TU WIEN Informatics

Supporting Laypeople in Learning Formal Medical Terminology

DISSERTATION

zur Erlangung des akademischen Grades

Doktorin der Technischen Wissenschaften

eingereicht von

Annisa Maulida Ningtyas, S.Kom., M.Eng.
Matrikelnummer 11936019

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Univ.Prof. Dr. Allan Hanbury
Zweitbetreuung: Dr. Florina Piroi

Diese Dissertation haben begutachtet:

Lorraine Goeuriot

Hussein Suleman

Wien, 14. August 2024

Annisa Maulida Ningtyas

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.at

TU WIEN Informatics

Supporting Laypeople in Learning Formal Medical Terminology

DISSERTATION

submitted in partial fulfillment of the requirements for the degree of

Doktorin der Technischen Wissenschaften

by

Annisa Maulida Ningtyas, S.Kom., M.Eng.
Registration Number 11936019

to the Faculty of Informatics

at the TU Wien

Advisor: Univ.Prof. Dr. Allan Hanbury
Second advisor: Dr. Florina Piroi

The dissertation has been reviewed by:

Lorraine Goeuriot	Hussein Suleman

Vienna, 14th August, 2024

Annisa Maulida Ningtyas

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.at

Erklärung zur Verfassung der Arbeit

Annisa Maulida Ningtyas, S.Kom., M.Eng.

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 14. August 2024

Annisa Maulida Ningtyas

Acknowledgements

First and foremost, I want to express my gratitude to God, who has consistently blessed me throughout my Ph.D. journey. Pursuing a Ph.D. is a significant milestone in my academic career, and it has been a journey filled with numerous ups and downs, moments of laughter and tears. I wish to sincerely thank everyone who has contributed to my successful completion of this remarkable journey.

I am deeply grateful for the valuable support and guidance I received from my supervisors, Prof. Allan Hanbury and Dr. Florina Piroi, throughout my Ph.D. journey. Their willingness to listen and offer excellent advice whenever I faced challenges was invaluable. I also appreciate the constructive feedback they provided allowing me to improve, learn, and grow from their experience.

I also express the deepest appreciation to all my colleague that helped and supported me. I thank Alaa El-Elbshihy, Fajar Juang, Tomasz Miksa, Andreas Rauber, Ryza Yasha, M.Syairaji, Angga E. Pramono, Dian Herawati for the collaboration works. Furthermore, I want to thank all the members of the Data Science group for the inspiring discussions we had during lunch or *jour fixe*.

I would also like to thank my Indonesian friends for their endless support and the wonderful travel experiences we shared: Rizki Mafthukah, Lisa P. Estianti, Diah F. Utami, Fernando Siahaan, Maria Tambunan, Agus Hidayat, Khairul Anam, Ayu Amalia, and Dini Khairunisa.

I want to express my gratitude to my family for their constant love and support throughout my life. My parents, Dede S. Hidayat and Rina Iriani, my parents-in-law, Bambang H. Budianto and Nuraeni Ekowati, my sisters, Rainy Rachmawati, Elisa N. Ismiah, Ranny F. Herwati, and Alethea Maharani, and my brothers, Justin A. W. Reza and Hardi B. Pamungkas, have all been sources of strength and encouragement for me.

I thank to my husband, Guntur Budi Herwanto for his support, understanding, especially during the stressful times we faced together on our Ph.D. journey. Going through this journey simultaneously was not easy, but we made it through together.

Last but not least, I would like to acknowledge AskAPatient.com for providing for providing access to the data essential to my research. This opportunity was important in completing my thesis.

Lastly, I would also like to acknowledge the Indonesian Ministry of Education and the OeAD for providing me with the opportunity to pursue my Ph.D. through the Indonesian-Austrian Scholarship Program (IASP), with reference number ICM-2019-13880.

Kurzfassung

Gesundheitskompetenz ist für Menschen ohne medizinischen Hintergrund unerlässlich, um fundierte Entscheidungen über ihre Gesundheitsversorgung zu treffen und ihre Lebensqualität zu verbessern. Eine wichtige Komponente dieser Kompetenz ist die funktionale Gesundheitskompetenz (Functional Health Literacy, FHL), die die grundlegenden Lese- und Schreibfähigkeiten umfasst, die erforderlich sind, um gesundheitsbezogene Informationen zu verstehen, z. B. Medikamentenanweisungen. Vielen Menschen fällt es jedoch schwer, Gesundheitsinformationen zu verstehen, was sich auf ihre Fähigkeit auswirkt, ihren Gesundheitszustand zu verstehen und wichtige Entscheidungen für die Gesundheitsversorgung zu treffen. Für Laien ist es wichtig, die medizinische Terminologie zu beherrschen, um Gesundheitsinformationen besser verstehen zu können. Mit der zunehmenden Zugänglichkeit von Gesundheitsinformationen im Internet stoßen Laien häufig auf medizinische Begriffe, verlassen sich aber in den sozialen Medien oft auf eine nicht standardisierte medizinische Sprache, was zu Verwirrung bei der Kommunikation mit medizinischen Fachkräften und anderen Personen führen kann.

Um dieses Problem zu lösen, wird in dieser Arbeit das Informal Medical Entity Linking (EL)-Modell vorgestellt, das Laien beim Erlernen medizinischer Terminologie durch Beiträge in sozialen Medien helfen soll. Dieses Modell identifiziert automatisch popularisierte medizinische Phrasen in Quellen wie Social-Media-Posts und normalisiert sie in standardisierte medizinische Fachterminologie in einer medizinischen Wissensbasis (KB), wie z. B. Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT) Konzepte, angereichert mit zusätzlichen Informationen aus relevanten Wikipedia-Artikeln. Der Abgleich von popularisierten Phrasen mit spezialisierten medizinischen Konzepten ist jedoch eine Herausforderung, da sich die medizinische Wissenschaft ständig weiterentwickelt und neben der vielfältigen popularisierten medizinischen Sprache auch neue medizinische Konzepte eingeführt werden. Die derzeitige Forschung im Bereich des Medical Entity Linking stützt sich auf überwachte Lernmethoden, die derzeit nur eine begrenzte Abdeckung medizinischer Konzepte aufweisen.

In dieser Arbeit werden Methoden zur Bewältigung von Datenbeschränkungen bei der Entwicklung des informellen medizinischen EL-Modells vorgestellt, insbesondere bei der Aufgabe der medizinischen Konzeptnormalisierung (MCN). Die MCN-Aufgabe zielt darauf ab, popularisierte medizinische Phrasen in spezialisierte medizinische Terminologien zu standardisieren. Wir schlagen einen Ansatz zur Augmentation von Textdaten vor, der

das Schreibverhalten von Laien nachahmt, um die Anzahl popularisierter medizinischer Phrasen für bestimmte medizinische Konzepte in öffentlich verfügbaren MCN-Datensätzen zu erhöhen, da viele der medizinischen Konzepte nur wenige Beispiele für popularisierte medizinische Phrasen aufweisen. Unsere Ergebnisse zeigen, dass die Augmentation Ansatz ist wirksam bei der Erhöhung der Anzahl der popularisierten medizinischen Phrasen für bestimmte medizinische Konzepte und Verbesserung der Modellleistung auf MCN-Modelle trainiert mit Daten Augmentation im Vergleich zu MCN-Modelle mit Original-Daten trainiert.

Darüber hinaus verwenden wir eine Fernüberwachungsmethode, um die Abdeckung medizinischer Konzepte in MCN-Datensätzen zu erweitern, indem wir Wikipedia und Wikidata nutzen, um automatisch beschriftete Daten zu generieren. Diese Strategie erweitert effektiv die Abdeckung medizinischer Konzepte und verbessert die Leistung von MCN-Modellen, wenn die automatisch beschrifteten Daten mit dem ursprünglichen Trainingsdatensatz für jeden öffentlichen MCN-Datensatz kombiniert werden, verglichen mit der Leistung von MCN-Modellen, die mit den ursprünglichen Trainingsdaten trainiert wurden.

Aufbauend auf der zuvor behandelten erweiterten Abdeckung wurde das informelle medizinische Entitätsmodell entwickelt. Dieses Modell besteht aus drei Phasen: (1) Die Phase der Erkennung benannter Entitäten (Named Entity Recognition, NER), in der popularisierte medizinische Phrasen im Text identifiziert werden. (2) Die Medical Concept Normalization (MCN)-Phase, in der jeder popularisierte medizinische Ausdruck auf die entsprechende medizinische Fachterminologie in SNOMED-CT normalisiert wird. (3) Die Phase der Entitäts-Disambiguierung (ED), in der der am besten geeignete Wikipedia-Artikel als Erklärungsquelle für die medizinische Fachterminologie gefunden wird.

Wir haben die Wirksamkeit des informellen Modells zur Verknüpfung medizinischer Entitäten beim Erlernen medizinischer Terminologie durch Benutzerexperimente evaluiert und die Teilnehmer in eine *Interventions*-Gruppe, die Unterstützung durch das Modell erhielt, und eine *Nicht-Interventions*-Gruppe, die keine Unterstützung erhielt, unterteilt. Ziel der Studie war es, festzustellen, ob die *Interventions*-Gruppe im Vergleich zur *Nicht-Interventions*-Gruppe signifikante Verbesserungen beim Erlernen medizinischer Terminologie zeigte. Die Ergebnisse deuten darauf hin, dass das Informal Medical Entity Linking-Modell ein potenzielles Instrument zur Unterstützung von Laien beim Erlernen medizinischer Terminologie in sozialen Medien sein kann.

Abstract

Health literacy is essential for individuals without a medical background to make informed choices about their healthcare and enhance their quality of life. A significant component of this literacy is Functional Health Literacy (FHL), which encompasses the basic reading and writing abilities needed to grasp health-related information, like understanding medication instructions. However, many people find it challenging to understand health information, which affects their capacity to understand their health conditions and make crucial healthcare decisions. Being knowledgeable in medical terminology is important for laypeople to grasp health information more effectively. With the growing accessibility of health information online, laypeople frequently encounter medical terms but often rely on non-standard medical language on social media, leading to possible confusion when communicating with healthcare professionals and others.

To address this issue, this thesis introduces the Informal Medical Entity Linking (EL) Model, designed to help laypeople learn medical terminology through social media posts. This model automatically identifies popularized medical phrases in sources like social media posts and normalizes them into standardized specialized medical terminology in a medical knowledge base (KB), such as Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT) concepts, enriched with additional information from relevant Wikipedia articles. However, aligning popularized phrases with specialized medical concepts is challenging due to the evolving nature of medical science and the introduction of new medical concepts, alongside the diverse popularized medical language. Current research in medical entity linking relies on supervised learning methods, which currently have limited coverage of medical concepts.

This thesis presents methodologies for addressing data limitations in the development of the Informal Medical EL model, specifically in the Medical Concept Normalization (MCN) task. The MCN task aims to standardize popularized medical phrases into specialized medical terminologies. We propose a textual data augmentation approach that mimics the writing behavior of laypeople to increase the number of popularized medical phrases for specific medical concepts in publicly available MCN datasets, as many of the medical concepts have few examples of popularized medical phrases. Our results show that the augmentation approach is effective in increasing the number of popularized medical phrases for specific medical concepts and improving model performance on MCN models trained with data augmentation compared to MCN models trained with original data.

Moreover, we utilize a distant supervision method to expand medical concept coverage within MCN datasets, leveraging Wikipedia and Wikidata to generate automatically labeled data. This strategy effectively broadens medical concept coverage and improves the performance of MCN models when combining the automatically labeled data with the original training dataset for each public MCN dataset, compared to the performance of MCN models trained on the original training data.

Building on the expanded coverage previously addressed, the Informal Medical Entity model was developed. This model consists of three phases: (1) The Named Entity Recognition (NER) phase, which identifies popularized medical phrases in the text. (2) The Medical Concept Normalization (MCN) phase, which normalizes each popularized medical phrase to its corresponding specialized medical terminology found in SNOMED-CT. Finally, (3) The Entity Disambiguation (ED) phase, which retrieves the most suitable Wikipedia article to serve as the source of explanation for the specialized medical terminology.

We evaluated the informal medical entity linking model's effectiveness in helping laypeople learn medical terminology through user experiments, dividing participants into an *intervention* group, which received assistance from the model, and a *non-intervention* group, which did not. The study aimed to determine if the *intervention* group showed significant improvement in learning medical terminology compared to the *non-intervention* group. The results indicate that the Informal Medical Entity Linking model can be a potential tool for assisting laypeople in learning medical terminology within social media settings.

Contents

CHAPTER 1

Introduction

As healthcare systems become more complex and public health challenges like COVID-19 arise, health literacy becomes essential to empower societies in managing their well-being and building resilience. Health literacy, defined by Sørensen, et al. [SVF+12], is *to effectively understand and use health information. It encompasses knowledge, motivation, and skills to access, comprehend, evaluate, and apply this information in everyday life.* Nutbeam et al. [Nut00] identifies three types of health literacy: *functional literacy*, *communication literacy*, and *critical literacy*. *Functional literacy* requires basic reading and writing skills for understanding everyday health information, such as reading medication instructions. *Communication literacy* is ability to extract information and derive meaning from different forms of communication and apply new information to changing situation. This including discussions with medical professionals about health and treatment options. *Critical literacy* is critically analyse information and to achieve policy and organisational changes.

Health literacy enables individuals to make informed decisions about healthcare, disease prevention, and health promotion, improving their quality of life throughout their lifespan. This issue has drawn significant attention from EU policymakers, health professionals, and researchers [SVF+12, SPR+15]. The World Health Organization (WHO) has also acknowledged health literacy as a fundamental pillar in successfully implementing the 2030 Agenda for Sustainable Development on a global scale [Wor17]. According to the findings of the European Health Literacy Survey (HLS-EU) [SPR+15], approximately half of the respondents from eight European countries - Austria, Bulgaria, Germany, Greece, Ireland, the Netherlands, Poland, and Spain - had limited health literacy. This limitation indicates that these individuals might struggle to understand health-related information, affecting their ability to understand their own health conditions and making it challenging for them to make important health decisions [HMCRT+18]. Furthermore, it can lead to delays in receiving appropriate medical treatment, resulting in poorer health outcomes [AO20].

Clear communication is crucial for effective engagement with medical professionals and comprehension of health information, but understanding health information depends on semantic skills that may vary based on an individual's health literacy level. Healthcare professionals are advised to communicate with patients or laypeople by avoiding medical terms or jargon and instead translating medical terms into lay language [GPH+22]. Proficiency in medical terms can significantly improve functional health literacy [FBNJ16a, OHLD+13], but it can be challenging for laypeople to understand [OHLD+13, LCW+19], leading to anxiety and dissatisfaction, especially for those with limited health literacy [LCW+19, AO20].

As technology and the internet have advanced, there is now an abundance of health information accessible to the public. Patients can access their electronic health record (EHR) notes through online portals, empowering them to actively manage their healthcare. However, individuals with limited literacy often rely on everyday medical knowledge, using non-medical terms, which makes it difficult for them to understand medical texts and may lead to miscommunication with medical professionals. To empower laypeople and improve functional literacy, one potential solution is enhancing their medical terms knowledge. Current research focuses on empowering laypeople by providing specialized medical terms with simpler definitions and clearer texts, thereby enhancing comprehension of EHR notes, radiology reports, and other clinical documents [ALL+20, CDPR+18, PRHB+13, KYJ+22].

Social media, including online patient forums, plays a significant role in healthcare, serving as a platform for community engagement, health promotion, patient education, and outreach [SHE+17]. This trend has led to a growing number of laypeople seeking medical knowledge, including specialized medical terms, through online resources [FBNJ16b]. As the use of online health resources continues to grow, familiarity with medical terms is increasing [FBNJ16b].

Moreover, with the advancement of social media, there is an abundance of health-related information available in the form of free text, often posted informally by laypeople. Identifying specialized medical concepts from this free text is valuable for medical companies, such as drug manufacturers, in summarizing the side effects of their products. This task is known as Medical Concept Normalization (MCN), where popularized medical phrases are linked to specialized medical concepts in a knowledge base (KB), such as the Unified Medical Language System (UMLS) [MT19]. While MCN has proven useful for medical organizations, such as for identifying adverse drug effects [TMNM18, MT19], its potential benefits for laypeople have not yet been explored.

Fage-Butler et al. [FBNJ16b] emphasize the importance of maintaining a lay language or patient-centered terminological level to keep forums engaging as learning environments. Furthermore, they point out that incorporating medical terms in online patient-patient communication could help expand patient navigation skills in the medical community [FBJ13]. Additionally, many laypeople appreciate the consistent and appropriate use of medical terms by professionals [HMMW06].

Many people find specialized medical terms difficult to understand, especially those with limited health literacy skills. Considering the educational potential of social media, this thesis introduces an informal medical entity linking (EL) model. This model links popularized medical phrases from social media to specialized medical terms and provides clear explanations to help laypeople learn medical terms.

In this thesis, we leverage SNOMED-CT (Systematized Nomenclature of Medicine - Clinical Terms) as the primary resource for mapping popularized medical terms to their specialized counterparts. Additionally, we use Wikipedia as a source to provide explanations for specific specialized medical terms. The integration of these resources is aimed at supporting laypeople in learning specialized medical terminology, potentially leading to enhanced functional health literacy of laypeople.

1.1 Motivation and Challenges in Informal Medical Entity Linking

The process of mapping words or phrases (mentions) in a text to concepts (entities) in a knowledge base is known as Entity Linking (EL) [KGH18]. Medical EL, as explained by [PAP+20], focuses on entity mentions associated with entity types such as drugs, diseases, symptoms, and so on. In contrast, Medical Concept Normalization (MCN) deals with phrases that might not be identified as medical entities by a standard medical named entity recognizer [PAP+20].

For instance, even though the phrase *cannot shut up for the whole day* is not recognized as a medical named entity, the MCN model will map it to Hyperactive Behavior [PAP+20]. In the context of our research, we define the Medical EL task as the process of identifying popularized medical phrases (referred to as mentions) found in user text and associating them with specific entities. These phrases are then standardized into specialized medical terms within a medical knowledge base, as shown in Figure 1.1. The overall process of Medical EL includes two main steps:

1. **Named Entity Recognition (NER)** or **mention detection**: identifies the specific words or phrases that may represent popularized medical expressions used by laypeople to describe medical concepts, referred to as *popularized medical phrases*.
2. **Medical Concept Normalization (MCN)**, which normalizes these popularized medical phrases into standardized medical concepts in a medical knowledge base.

The source texts used in medical Entity Linking (EL) typically consist of laypersons' language, which is more informal and descriptive compared to medical text. As an example, in Figure 1.1, the input text is "I feel a *bit drowsy* & have a *little blurred vision*, so far ...". In this text, *bit drowsy* is identified as an popularized medical phrases with the disease entity through Named Entity Recognition (NER). To define entity types, we followed the approach presented in a previous study [STT17], where they combined

categories like adverse drug reaction, disease, symptom, and finding into a single category called disease. The phrase *bit drowsy* refers to *Drowsy* in medical knowledge base (KB) through MCN [KMJKW15]. Bridging the gap between specialized medical terms and popularized medical phrases can be challenging, and simple dictionary matching doesn't work well for detecting medical terms.

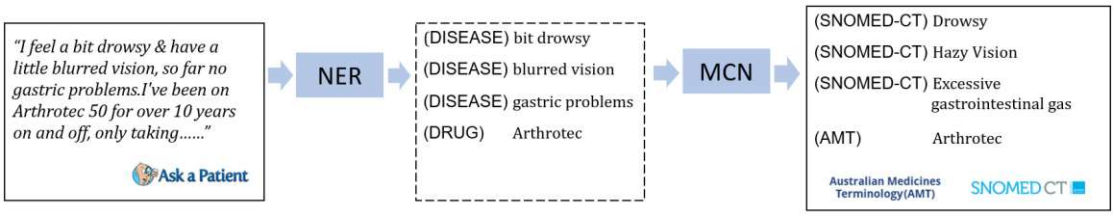


Figure 1.1: Generic Medical Entity Linking (MEL) Workflow. The MEL task involves analyzing text from social media to identify specific words or phrases that refer to medical entities, which are often expressed in informal language. The identified popularized medical phrases and their corresponding entities are then converted into standardized medical concepts. This process, known as Medical Concept Normalization (MCN), involves mapping these popularized phrases to specialized medical terms found in a designated medical knowledge base (KB).

The recent approaches treats MCN as a classification problem [LC15, LC16, TMNM18, MT19]. This approach relies on annotated labeled data sets, such as Psychiatric Treatment Adverse Reaction (PsyTAR) [ZFP+19], which covers 6,556 popularized medical phrases mapped to 755 medical concepts and CSIRO Adverse Drug Event Corpus (CADEC), which contains 9,111 popularized medical phrases mapped to 1,029 medical concepts [KMJKW15]. The most comprehensive data set is COMETA, which covers 20,015 popularized medical phrases mapped to 3,645 concepts of 350,000 concepts of Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT)[Don06]. Together these datasets (CADEC, PsyTAR, COMETA) cover about 10% of SNOMED-CT concepts, and many medical concepts have limited examples of associated popularized phrases (see Table 2.2). We refer to the popularized medical phrases associated with specific medical concepts as *support phrases*. Nevertheless, the problem of unseen medical concepts still occurs [BLSC20]. Increasing the coverage of the data sets by human annotation requires a lot of time and cost. Additionally, as medical science continues to evolve, the number of new medical concepts also grows. This constant growth, coupled with the diverse nature of informal language, makes it challenging for algorithms to achieve reliable results in the medical domain.

To address the scarcity of data sets problem, various approaches have been developed to explore alternative forms of labeled data as substitutes for manually labeled data [HLA+20]. These approaches have been categorized into two main categories: distant supervision and data augmentation [HLA+20]. Distant supervision automatically generates labeled data with existing knowledge bases or dictionaries. However, the existing distant

supervision approach [PAP+20] suffers from a language gap between popularized medical phrases used by laypeople to describe their specialized medical terms. For example the phrase *spinning sensation* refers to *vertigo*. Due to the similarity method used in [PAP+20], the difference cannot be detected. Data augmentation [WZ19] attempts to artificially add more data to boost the model performance and add various representations of popularized phrases. However, the impact of augmented data on Medical Entity Linking (MEL) model performance has yet to be investigated.

At the beginning of this chapter, we highlighted the limited impact of current medical EL in social media for laypeople. To address these challenges and leverage the potential of medical EL for laypeople in social media settings, we select SNOMED-CT and Wikipedia as our primary knowledge sources, to extend the current medical EL model. SNOMED-CT is a comprehensive, multilingual clinical healthcare terminology system that provides a standardized way to represent and organize medical concepts [sno]. It has a broad scope of coverage, including diseases, findings, and other clinical concepts [VVP23, BRB+07, sno]. SNOMED-CT also serves as a standardized terminology for electronic health records (EHRs) and other healthcare applications [VVP23]. Laypeople or patients may encounter these terms when accessing their health records or interacting with healthcare providers, and by utilizing SNOMED-CT as a source of specialized terms and by introducing its specialized medical terms to laypeople can help bridge the gap between professional language and consumer understanding.

Furthermore, we hypothesize that providing explanations from Wikipedia articles will enhance the benefits of medical entity linking for laypeople, empowering them to learn specialized medical terms in a social media setting. Wikipedia is the most frequently accessed resource for health information, with English medical content being accessed more often than any other health information source [HW15]. Despite its popularity, skepticism has been raised regarding potential inaccuracies [AAAA15]. However, research conducted by Giles [Gil06] found that Wikipedia's error rates are similar to those of Encyclopedia Britannica, a trusted, expert-reviewed resource. According to Scaffidi et al. [SKW+17], Wikipedia has a group of users under WikiProject Medicine who act as an expert review board to define the manual writing styles of medical content on Wikipedia.

Beyond the pros and cons, research has found that Wikipedia has a potential role in medical education [SKW+17, Smi20]. However, the quality of Wikipedia's health content for consumers requires further investigation [Smi20]. Polepalli et al. [PRHB+13] found that Wikipedia articles can improve laypeople's understanding of specialized medical terms. Based on these findings, we have chosen Wikipedia as a source of explanations for our enhanced medical EL model.

Previous research [SMM+20, ALL+20, CDPR+18] has employed text simplification to associate specialized medical terms in medical documents, such as EHR or radiology reports, to a knowledge base or dictionary, aiming to offer lay explanations. However, to the best of our knowledge, there is limited research on using medical EL in social media settings to support laypeople in learning medical terms. Our approach aims to address this gap.

1.2 Research Questions

Our research focuses on exploring the usefulness of informal medical entity linking in improving the health literacy of laypeople. Considering the broad scope of health literacy, our specific focus is on functional literacy, particularly the familiarity with and comprehension of specialized medical terms in social media settings. Laypeople often use different everyday phrases for specific medical terms, and our proposed informal entity linking model aims to introduce them to the corresponding specialized medical terms along with explanations for each popularized medical phrase within social media platforms. We hypothesize that this exposure has the potential to empower laypeople and enhance their understanding of medical terms. Thus, we define our first research question as:

> RQ1: How effective is medical entity linking, which maps popularized medical phrases to specialized medical terms with explanations, in increasing digital health literacy among laypeople?

Considering the data scarcity problem discussed in Section 1.1, we raise the second research question:

> RQ2: How effective are data augmentation and distant supervision methods in overcoming the problem of data scarcity in MCN tasks?

The effectiveness with respect to RQ2 refers to the impact of the performance of the supervised models on the MCN task and the increase in concept coverage of the current MCN datasets [ZFP+19, KMJKW15, BLSC20].

Data augmentation is used to increase the diversity of lay medical terms. In our research, we employed a textual data augmentation method, such as *synonym replacement* and *paraphrasing*. One of the challenges in data augmentation is that a substitution of words in popularized medical terms or phrases may lead to a different medical concept. For example, *weight gain* as a result of *obesity* can be transformed into *burden gain*, which may be associated with *struggle*. In contrast to data augmentation, distant supervision will be used to increase the coverage of the specialized medical terms from the available MCN datasets [KMJKW15, ZFP+19, BLSC20]. The challenge in distant supervision lies from the significant language gap between popularized medical expressions and specialized medical terms. For example, the popularized phrase *need to sleep constantly* used by laypeople to describe their health conditions on social media, which corresponds to *Somnolence* in specialized medical term, should ideally be synonymous.

1.3 Contributions

In this section, we present the contributions relevant to our research questions as follows:

- For Research Question 1 (RQ1), we created an end-to-end informal medical entity linking model to assist laypeople in learning medical terms. We conducted user experiments to evaluate how well this model helps users understand and become familiar with medical terms.

- In addressing Research Question 2 (RQ2), which involves the challenge of data scarcity, we introduced data augmentation and distant supervision methods. The goal of data augmentation is to increase the number of informal medical phrases that correspond to formal medical concepts, especially those with very few examples of informal medical phrase. Additionally, we employed distant supervision to expand the coverage of medical concepts within existing datasets ([KMJKW15, ZFP+19, BLSC20]) by automatically generating labeled data for the Medical Concept Normalization (MCN) task.

Figure 1.2 illustrates our primary contribution in this research, the ***end-to-end informal medical entity linking (EL) model***. The pipeline consists of three modules: (i) *identifying medical phrases*; (ii) *medical concept normalization*, and (iii) *entity disambiguation*. First, the identifying medical phrases module extracts informal medical phrases from text. This step is treated as a Named Entity Recognition (NER) task. Second, the output of the first module is parsed by the Medical Concept Normalization (MCN) module, which transforms popularized medical phrases into specialized medical terms, addressing normalization challenges as a multi-class classification problem. To enhance the performance of this module, data augmentation and distant supervision are employed to generate additional data, as indicated in the red boxes. This additional data is then combined with the existing datasets (as seen in the "Unify Datasets" step). Finally, the output of the MCN modules is processed by the last module, entity disambiguation, which extracts the relevant Wikipedia articles for use as sources of explanation. To address the challenge of data scarcity in MCN tasks (RQ2), our research focused on two primary contributions:

Data Augmentation We increased the number of informal medical phrases for concepts that had limited example of informal medical phrases by simulating the writing style of laypeople. This ensured the context of the informal phrases was maintained without any shift in its meaning. Our augmentation techniques included: 1) Character augmentation (e.g. keyboard errors); 2) Word augmentation (e.g. synonym replacement); and 3) Paraphrasing. Based on our experiments on the CADEC [KMJKW15] and PsyTar [ZFP+19] data sets, the augmentation increased the variation of informal medical phrases, and improved the model performance on the MCN module compared to the original data sets as discussed in this thesis and as reported in [NHPA21].

Distant Supervision To broaden the coverage of medical concepts within existing datasets [ZFP+19, KMJKW15, BLSC20], we extracted pairs of informal and formal concepts. The informal terms were extracted from Wikipedia articles, utilizing redirect

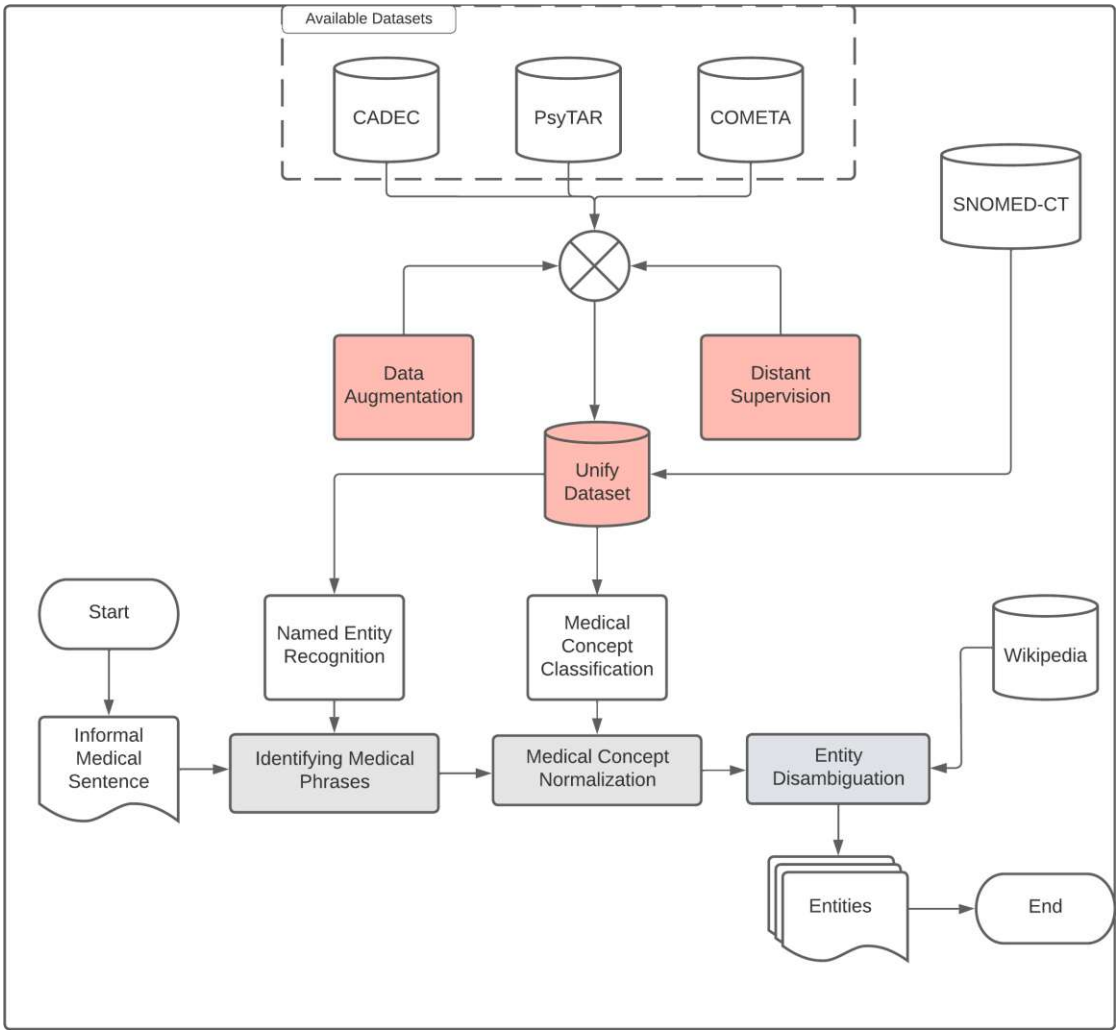


Figure 1.2: The end-to-end informal medical entity linking starts with a user-generated text and goes through the three main modules to extract medical entities. The grey boxes represent the main modules, while the red boxes highlight the contributions of our study.

names, wikilink, and the article summary [NEEH+22]. The formal concepts, on the other hand, were sourced from SNOMED-CT. We then matched these informal terms with corresponding formal concepts from SNOMED-CT. This distant supervision approach was able to increase the concept coverage in existing MCN data sets. Our findings demonstrated that our distant supervision method significantly enhanced model performance across three publicly available MCN datasets [ZFP+19, KMJKW15, BLSC20].

The next contribution is to improve the familiarity and understanding of laypersons with specialized medical terms with the support from our informal medical entity linking

model (RQ1). As discussed in Section1.1, the existing medical linking models mainly focus on (i) identifying medical phrases and on (ii) medical concept normalization (as shown in Figure 1.1). However, this usefulness for the laypeople is limited. To address this limitation, we introduced an **Entity Disambiguation (ED) Module**, which extracts relevant explanations from Wikipedia articles. The MCN module serves as a bridge to connect the informal medical phrases to the corresponding Wikipedia articles as source of explanation of the medical terms.

User Experiment: Assessing Informal Medical EL Model for Layperson Learning of Medical Terms To assess the effectiveness of our informal medical entity linking model in improving the medical terms knowledge among laypeople, we conducted user experiments involving two distinct groups: the *intervention* group and the *non-intervention* group. The *intervention* group received support from our informal medical entity linking model to complete the user experiment tasks, while the *non-intervention* group did not receive any model assistance. Our findings demonstrated that the informal medical entity linking model significantly improved both *surface-level* familiarity (the ability to recognize the word-form of formal medical terms from informal medical phrases found in social media) and *concept-level* familiarity (the ability to understand and identify the meaning of formal terms) among participants in the *intervention* group compared to the *non-intervention* group.

1.4 Published Research

This thesis heavily relies on research papers presented at conferences related to Information Retrieval (IR) and Natural Language Processing (NLP). The following papers form the essential foundation of this thesis:

- Ningtyas, A. M., El-Ebshihy, A., Piroi, F., Hanbury, A., & Andersson, L. (2020). TUW-IFS at TREC NEWS 2020: Wikification Task. In *TREC*. [NEEP$^+$20]
- *Ningtyas, A. M., Hanbury, A., Piroi, F., & Andersson, L.* (2021). Data augmentation for layperson's medical entity linking task. **In Proceedings of the 13th Annual Meeting of the Forum for Information Retrieval Evaluation (pp. 99-106)** [NHPA21]
- *Ningtyas, A. M.* (2022). Medical Entity Linking in Laypersons' Language. **In European Conference on Information Retrieval (pp. 513-519)**. Cham: Springer International Publishing [Nin22]
- *Ningtyas, A. M., El-Ebshihy, A., Herwanto, G. B., Piroi, F., & Hanbury, A.* (2022). Leveraging Wikipedia Knowledge for Distant Supervision in Medical Concept Normalization. **In International Conference of the Cross-Language Evaluation Forum for European Languages (pp. 33-47)**. Cham: Springer International Publishing [NEEH$^+$22]

The following papers are not directly relevant to this thesis:

- *El-Ebshihy, A., Ningtyas, A. M., Andersson, L., Piroi, F., & Rauber, A.* (2020). ARTU/TU Wien and artificial researcher@ LongSumm 20. **In Proceedings of the First Workshop on Scholarly Document Processing** (pp. 310-317) [EENA+20]
- *El-Ebshihy, A., Ningtyas, A. M., Andersson, L., Piroi, F., & Rauber, A.* (2022). A platform for argumentative zoning annotation and scientific summarization. **In Proceedings of the 31st ACM International Conference on Information & Knowledge Management (pp. 4843-4847)** [EENA+22]
- *El-Ebshihy, A., Ningtyas, A. M., Piroi, F., Rauber, A., Romadhony, A., Faraby, S. A., & Sabariah, M. K.* (2023). Using Semi-automatic Annotation Platform to Create Corpus for Argumentative Zoning. **In *International Conference on Theory and Practice of Digital Libraries* (pp. 132-145).** Cham: Springer Nature Switzerland [EENP+23]
- *Miksa, T., Suchánek, M., Slifka, J., Knaisl, V., Ekaputra, F.J., Kovacevic, F., Ningtyas, A.M., El-Ebshihy, A. and Pergl, R.*, (2023). Towards a Toolbox for Automated Assessment of Machine-Actionable Data Management Plans. ***Data Science Journal*, *22*, 28** [MSS+23]

1.5 Thesis Structure

The thesis is structured as follows: Chapter 2 provides the background information for this research. This includes information about (i) definition of functional health literacy; (ii) assessing functional health literacy; (iii) definition of medical terms; (iv) tools to enhance laypeople's comprehension of medical terms; (v) medical entity linking and medical concept normalization; (vi) the datasets used and a review of related works on medical concept normalization; and (vii) natural language processing in low resource scenarios. Chapter 3 presents our proposed data augmentation methods aimed at increasing the number of popularized medical phrases available in the datasets, particularly in tasks like named entity recognition (NER) and medical concept normalization (MCN). Chapter 4 describes the proposed distant supervision approach that leverages Wikipedia to expand the number of medical concepts in the current MCN datasets. Chapter 5 outlines the popularized medical entity linking model and presents the evaluation results, including both model performance assessment and expert evaluation regarding the reliability of the proposed model, before conducting user experiments. Chapter 6 describes the survey design and user experiment results to evaluate the effectiveness of the informal entity linking model in enhancing laypeople's comprehension of specialized medical terms. Lastly, Chapter 7 concludes the thesis by summarizing findings and discussing future research opportunities to support laypeople in enhancing medical terms knowledge.

CHAPTER 2

Background and State-of-the-Art

This thesis aims to improve the functional health literacy, specifically in the comprehension of the medical terms by developing the informal medical entity linking model. In this chapter, we present the overview and the definitions of the key concepts discussed in this thesis as our background. We begin with the definition of Functional Health Literacy (FHL) (Section 2.1). Next, we outline the tool used for evaluating FHL (Section 2.2). Then, we defined the medical terms used in this thesis (Section 2.3). Furthermore, we present the related work on the core topic of this thesis. We review related work on tools to support laypeople in enhancing comprehension of specialized medical terms (Section 2.5). Then, we describe the literature about the medical entity linking and medical concept normalization tasks (Section 2.6). Additionally, we review the datasets used in the research and its limitation (2.7). Finally, we review related works on natural language processing in low-resource scenarios 2.8), since informal medical entity linking is a low-resource problem.

2.1 Functional Health Literacy

Functional Health Literacy (FHL) is defined as a degree to which individuals, e.g. patient or laypeople have the capacity to obtain, process and understand basic health information to make appropriate health decisions [oM04, Nut00]. Figure 2.1 presents a conceptual model of the relationship between the health-related printed and oral literacy, and health outcomes proposed by Barker et al. [Bak06].

This model, aligning with the definition of FHL, emphasizes individual abilities including the skills to read and understand medical texts and instructions (prose literacy), and the capability to interpret and use numerical information, such as calculating medication dosages (quantitative literacy), and ability to locate and use information in documents (document literacy) [Bak06].

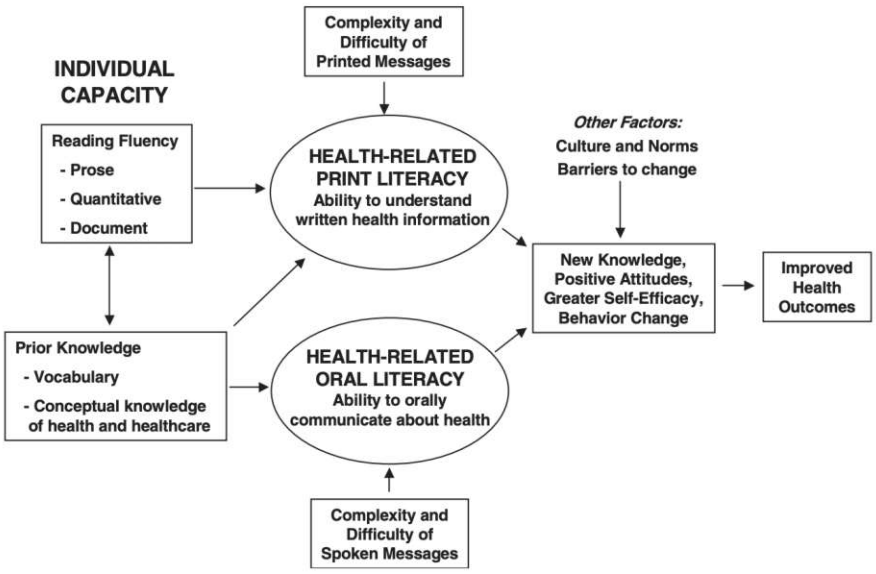


Figure 2.1: Conceptual model of the relationship between the health-related print literacy, health-related oral literacy, and health outcomes [Bak06]

The Institute of Medicine (IOM) [oM04] identifies health literacy into two main components: health-related print literacy and health-related oral literacy. The comprehension of health-related printed or oral literacy relies on an individual's health-related reading proficiency, health-related vocabularies, familiarity of health concepts found in the reading materials or discussions [Bak06]. Baker's model [Bak06] suggests that reading fluency is affected by the prior knowledge of each individual. This prior knowledge includes health-related vocabulary and conceptual knowledge of health and healthcare. FHL is important for empowering laypeople to take an active role in their healthcare. This affects laypeople's ability to manage and make appropriate health decisions for their health, and increase patient engagement. Low FHL can lead to misinterpretation of health information [HMCRT$^+$18], and result in poor health outcomes. This is often due to a lack of basic skills in reading and understanding instructions, informed consent documents, medication instructions, and more [GB08]. Therefore, improving FHL is considered as one important component within the broader scope of health literacy that contributes to better health outcomes.

2.2 Assessing Functional Health Literacy

The importance of understanding medical terms for good functional health literacy cannot be overstated. To address this, numerous surveys have been developed to evaluate health literacy, especially in terms of understanding medical terms and reading comprehension [PBWN95, BWP$^+$99, PMFN15, LWC$^+$18, GPH$^+$22]. These assessments predominantly utilize a multiple-choice format, though the framing of these questions can vary widely. A

popular methodology is the Cloze-type approach, which involves fill-in-the-blank questions [LWC+18, PBWN95, BWP+99]. One such instrument is the Rapid Estimate of Adult Literacy in Medicine (REALM). It specifically measures a patient's ability to pronounce medical and lay terms, with an emphasis on body parts and illnesses [DLJ+93]. Unlike many other tools that rely on multiple-choice or fill-in-the-blank formats, REALM is unique in its emphasis on verbal assessments. The test is concise, taking only 2 to 3 minutes to administer and score. Conversely, the Test of Functional Health Literacy in Adults (TOFHLA) assesses reading comprehension and numerical skills, drawing from materials such as healthcare documents, hospital forms, and medication labels [PBWN95]. TOFHLA has a shorter version, the Short Test of Functional Health Literacy in Adults (S-TOFHLA). The comprehensive TOFHLA includes a reading comprehension section with 50 questions and a 17-item numerical test, typically taking about 22 minutes to complete. In an effort to streamline the process, Baker et al. [BWP+99] reduced the TOFHLA's duration to 12 minutes by reducing the numerical and reading sections. New instruments are constantly developed for particular use scenarios in health domain. Puspitasari et al. [PMFN15] designed a medical terms questionnaire to investigate the influence of health topic familiarity on search behaviors. This questionnaire is divided into three sections: (1) Section 1 measures familiarity with medical terms, (2) Section 2 measures conceptual understanding of consumer-friendly terms, and (3) Section 3 measures conceptual understanding of advanced health terms. The questionnaires are related to specific health topics such as skin allergies, cardiovascular disease and cholesterol problems. Given the increasing importance of Electronic Health Records (EHRs), Lalor et al. [LWC+18] introduced ComprehENotes, an instrument tailored to assess the comprehension of EHR notes. This instrument, formulated using the Sentence Verification Technique (SVT), addresses diseases ranging from Cancer to Liver Failure. However, the authors realized the need for a more concise version and subsequently reduced the number of questions from 55 questions to 14 questions. Lastly, to emphasize the gap between medical jargon used by medical professionals and everyday language, Gotlieb et al. [GPH+22] developed a survey to evaluate how well patients understand medical terms. The survey showed that there can be confusion because some words mean different things in medical jargon than they do in everyday language. For example, the word *negative* in medical context usually means something good, which is the opposite of what it normally means in everyday language. In summary, the significance of functional health literacy, particularly the ability to grasp medical terms, lies in its role in empowering patients and enhancing their participation in effective healthcare. Several assessment tools have been developed for the purpose of evaluating this understanding. These tools range from assessing non-experts' pronunciation of medical terms [DLJ+93], finding synonyms and meanings for specific medical terms [PMFN15, GPH+22], to evaluating reading comprehension skills [PBWN95, BWP+99, LWC+18]. Furthermore, most of these instruments are designed to measure non-experts' knowledge of specialized medical terms and their ability to comprehend healthcare-related documents such as Electronic Health Records (EHR), medication prescriptions, and so on.

2.3 Medical Terminology

We use the term medical terminology to describe health-related vocabulary. Seiffe et al. [SMM+20] categorize medical terminology into two types: *medical technical terms* and *medical lay terms*. The authors define *medical technical terms* as terms used by medical professionals or terms that often include words of Latin or Greek origin. On the other hand, *medical lay terms* are health-related terms that are easily understood by patients, consisting of everyday language and words.

Fage-Buttler et al. [FBNJ16a] divided medical terminology used in online discussions between patients: *dictionary-defined medical terms*, *co-text-defined medical terms*, *medical initialisms*, *medication brand names*, and *colloquial technical terms*. *Dictionary-defined medical terms* are those found in medical dictionaries, like MedlinePlus, and often have Latin or Greek roots. *Co-text-defined medical terms* are words that have a normal, everyday meaning but get a special medical meaning when used in a medical situation (e.g. 'uptake'). *Medical initialisms* are abbreviations used in medical language. *Medication brand names* refer to both brand and generic medications. *Colloquial technical terms* are simplified technical terms used by laypeople.

In this thesis, we follow Seiffe et al.'s [SMM+20] categorization with some modifications. We use the term **specialized medical terminology** instead of medical technical terms, defining it as terms used by medical professionals, often derived from Latin or Greek, and found in medical knowledge bases like SNOMED-CT. We also use **popularized medical terms** instead of medical lay terms, defining them similarly as terms used by laypeople, composed of everyday language.

2.4 Medical Vocabulary

In this section, we present various medical vocabularies to motivate our selection of SNOMED-CT as the source of vocabulary for introducing specialized medical terminology to laypeople.

As we mentioned in the previous section, we defined medical terminology ("specialized medical terminology") as the terms used by healthcare professionals. Various medical vocabularies exist to categorize and standardize health-related terms. These range from vocabularies designed for consumers (or "laypeople") to those intended for healthcare professionals. Examples include the Consumer Health Vocabulary (CHV), Medical Subject Headings (MeSH), SNOMED-CT, and others.

CHV is a controlled vocabulary containing the mapping between medical vocabulary commonly used by laypeople to express a medical concept with Unified Medical Language System (UMLS) medical terms [MGL+21, ZTGW+06]. For example a consumer term *heart attack* can be "translated" to *myocardial infarction*. The purpose of CHV is to bridge the communication gap between laypeople and healthcare professionals. There are over 150,000 laypeople terms mapped to approximately 58,000 UMLS terms.

Medical Subject Headings (MeSH), on the other hand is a controlled vocabulary used for indexing, cataloging, and searching scientific articles or journals of biomedicine domain [Lip00]. MeSH stored the medical terms in a hierarchical tree structures, it organizes article or journals in 16 categories, such as category A (anatomic terms), category B (organisms), category C (diseases), etc. along with multiple subcategories [mes]. The descriptors has up to 13 levels of subcategories [MGL+21]. For example, articles concerning *Streptococcus pneumoniae* will be found under the descriptor *Streptococcus Pneumoniae* rather than the broader term *Streptococcus*, while an article referring to a new concept which not yet in the vocabulary, such as *streptococcal bacterium* will be listed under *Streptococcus* [mes].

In contrast, SNOMED-CT is a comprehensive, multilingual clinical healthcare terminology system that provides a standardized way to represent and organized medical concepts [sno]. Compared to CHV and MeSH, SNOMED-CT has a broad scope of coverage, including diseases, findings, body structure, procedures, and other clinical-related concepts [VVP23, BRB+07, sno]. Moreover, it offers granularity of clinical terminology across multiple clinical domains, with more than 350,000 concepts. The concepts in SNOMED-CT are organized in hierarchies, which can help laypeople understand relationships between medical terms. For instance, *heart attack* is synonymous with specialized medical term *myocardial infarction*, and this concept has its finding site in *myocardium structure (body structure)*[1]. Additionally, SNOMED-CT publishes Patient-Friendly Extension Releases for some member countries, such as the Netherlands, to support consumer health applications [vMdKNC18].

SNOMED-CT is designed as a standard for electronic health records (EHRs) and other healthcare applications for coding or classifying clinical information [VVP23]. While, EHRs primarily serve to provide patient care, and giving patients access to their EHRs may improve quality of care and patient engagement [NFL+20, vMDN+20]. However, laypeople may struggle with understanding the professional medical terminology used in their EHRs, leading to increased anxiety [TPK+21].

Furthermore, this thesis extends work in Medical Entity Linking, particularly the Medical Concept Normalization (MCN) task, where adverse drug effects found in social media are mapped to SNOMED-CT terms. This alignment with current research trends in the MCN task further motivates the use of SNOMED-CT.

Given its comprehensive coverage, granularity, and direct relevance to healthcare systems, SNOMED-CT emerges as an ideal resource for introducing laypeople to specialized medical terms used by healthcare professionals. Its broad scope, multilingual, comprehensive structure, and role as a standard terminology used by healthcare providers and organizations make it suitable for introducing specialized medical terms to laypeople.

[1] https://bit.ly/4cUEAO7

2.5 Tools to Enhance Laypeople's Comprehension of Medical Terminology

Given the significance of medical terms in functional literacy and the abundance of online health information, along with laypeople's interest in accessing health information through various systems like electronic health records (EHR), personal health records (PHR), and online health forums, researchers have proposed systems to assist laypeople in understanding medical terminology [ALL+20, CDPR+18, PRHB+13, ZTGK+07, KCZT10].

To enhance the readability of medical texts like EHR or PHR, Zeng-Treitler et al. [ZTGK+07] proposed simplifying medical terminology by substituting terms with synonyms and generating explanations. Their model involves (i) concept extraction, (ii) synonym identification, and (iii) explanation generation. For term extraction from EHRs, they utilized HITEx [ZTGW+06], a rule-based tool. The output was then fed into the synonym identification process, where each term's synonym from Open-Access Collaborative - Consumer Health Vocabulary (OAC-CHV) was matched. The choice to replace was guided by familiarity scores in OAC-CHV (0 for unfamiliar, 1 for familiar). Lastly, explanations were generated using non-hierarchical UMLS relations (e.g., *has site of* for *disease or syndrome*). System feasibility was assessed via user experiments with experts and non-experts, focusing on clinical document translation. Expert review indicated correctness and usefulness of the majority of text replacements and insertions.

Kandula et al. [KCZT10] introduced a simplification tool to tackle the semantic complexity of medical terms by replacing them with simpler synonyms or providing straightforward explanations using hierarchically and/or semantically related terms. They identified difficult medical terms based on their frequency of occurrence in lay reader-oriented biomedical sources, like Reuters News or MedlinePlus queries. Their reasoning was that terms found more frequently in these sources are likely to be easier for non-experts to understand. They associated noun phrases with UMLS and provided explanations by identifying key relationships, exemplified as *<original term semantic group (e.g., disease)> <connector (a condition affecting)> <explanation term semantic group (e.g., anatomical structure)>*. The tool also simplified lengthy sentences by breaking them into shorter grammatical ones. Validation involved comparing the readability of original and simplified texts in electronic health records and biomedical journal articles using readability metrics like Flesch Kincaid Grade Level and Simple Measure of Gobbledygook (SMOG), and user cloze tests. Results indicated significant enhancements in simplification, particularly in electronic medical records according to cloze tests.

With the advancement of deep learning, specifically utilizing supervised learning techniques, Chen et al. [CDPR+18, PRHB+13] focus on enhancing laypeople's grasp of medical terminology through a tool called NoteAid. Unlike earlier research [ZTGK+07, KCZT10] that relied on rule-based methods and dictionary matching for extracting medical terms, identifying synonyms, and generating explanations, Chen et al. [CDPR+18] employ supervised learning for text simplification. This novel approach involves two modules: (i) CodeMed, a lexical resource containing lay definitions for medical terms, and (ii)

MedLink, which establishes connections between medical terms and lay definitions. Their proposed method includes a distant supervision algorithm, training a supervised model to prioritize important medical terms for understanding electronic health records (EHRs). Feedback from cognitive walkthrough and post-session questionnaire for NoteAid was positive, highlighting its usability, visual display, speed, and adequacy of lay definitions.

Alfonso et al. [ALL+20] introduced the SIMPLE system, a web-based application designed to process medical text, such as electronic health records (EHRs), and provide annotated output with highlighted technical terms and corresponding info-buttons. The SIMPLE system consists of three modules: (i) **highlight** module, which identifies medical terminology using UMLS, MeSH (English and Italian versions), MedDRA, and MTHSMS vocabularies; (ii) **map** module, which translates the detected medical terms to consumer health vocabularies (CHV) to obtain simplified versions of the medical terms; and (iii) **define** module, which extracts the simple definitions of each medical term from WebMD and Italian dictionaries for health consumers. The proposed model was evaluated through user experiments with both non-experts and experts, as well as objective term familiarity scoring. The findings revealed that non-experts found the system helpful in comprehending medical texts, providing comparable information to medical experts, and presenting information in a more accessible manner.

The existing systems and methods particularly address the translation from specialized medical terms to plain medical phrases. Seiffe et al. [SMM+20] emphasized another potential resource for empowering non-experts, namely, social media. The study mentioned that the information shared online unveils directly or indirectly information about people's health situation and thus provides a valuable data resource, e.g adverse drug effects. The study aims to tackle the challenge of mapping popularized medical expressions with specialized terms and mapping specialized medical terms used in social media to lay terms, given that knowledgeable laypeople might employ specialized medical terms. The main focus in this research is to generate the datasets to be used to train a model to address the problem in the German language. To assess the datasets, two types of experiments were conducted: (1) mapping specialized medical terms to popularized equivalents through synonym replacement from UMLS, and (2) mapping popularized medical terms to medical terminology. This resource holds the potential to facilitate the mapping and translation between both language styles in the medical domain. Unlike previous studies [ZTGK+07, KCZT10, CDPR+18, ALL+20], this research does not assess the advantages of these experiments, which could be valuable for laypeople seeking to understand medical terminology.

In this thesis, we introduce a novel approach that uses social media as the primary data source to enhance laypeople's understanding of medical terminology. This strategy sets our work apart from existing systems such as NoteAid [CDPR+18] and SIMPLE [ALL+20], which primarily focus on simplifying medical terms within formal documents, such as electronic health records (EHRs). In contrast to NoteAid [CDPR+18] and SIMPLE [ALL+20], which use rule-based methods or text simplification techniques on EHR documents, our approach addresses the problem through domain entity linking

from social media. NoteAid and SIMPLE assist laypeople by adding definitions to medical terms found in EHRs, helping them understand the document's context. Our method reverses this process by leveraging social media to introduce specialized medical terms using popularized phrases commonly used by laypeople. We hypothesize that integrating our model with social media will support laypeople by exposing them to specialized medical terms derived from popularized medical phrases they already know. This approach may support the improvement of functional health literacy, particularly medical terminology knowledge. A detailed discussion of this novel approach is provided in Chapter 5, where we describe our proposed entity linking model. Furthermore, Chapter 6 demonstrates the effectiveness of this strategy, substantiating the utility of our approach.

2.6 Medical Entity Linking and Medical Concept Normalization

This section describes Entity Linking (EL) and Medical Concept Normalization (MCN), distinguishing between the two tasks. It then focuses on MCN task, providing an overview of the current research approaches, from early dictionary-based approaches to recent advanced machine learning methods.

2.6.1 Definition

Entity Linking (EL) is the process of identifying specific phrases (referred to as *mentions of entities*) in text, which may have multiple meanings, and linking them to their corresponding entries in a knowledge base (such as Wikipedia) [KGH18, TCL+16]. This task is also known as Wikification [MC07]. The main objective in EL and Wikification is to identify the important terms, usually represented as named entities, and associate them with the appropriate Wikipedia articles [MC07].

In their research, Trani et al. [TCL+16] provided an example to illustrate Entity Linking (EL) using the sentence: "Maradona played his first World Cup tournament in 1982, when Argentina played Belgium in the opening game of the 1982 Cup in Barcelona." They enhanced this sentence for better understanding by linking specific terms to their detailed entities: Maradona is linked to *Diego Maradona*, the World Cup tournament to *FIFA World Cup*, and so on for other terms like Argentina, Belgium, and Barcelona, each linked to their respective detailed entities.

Entity Linking (EL) involves two primary tasks: (i) Mention Detection or Named Entity Recognition (NER), which is the process of identifying specific word groups in a text that likely refer to entities (e.g *Maradona*) [KGH18, MC07]; and (ii) Entity Disambiguation – connects these identified phrases to relevant entities in a knowledge base, such as linking *Maradona* to *Diego Maradona* in Wikipedia pages [KGH18, TCL+16]. In our TREC report on Wikification, we further detailed the Entity Disambiguation process. It starts by detecting potential knowledge base entries (called *candidate entities*) that correspond to identified words and phrases. For instance, the mention *Argentina* might refer to

different entities like the country or the national football team. Then, it involves word sense disambiguation to determine the most relevant entities to link from each mention [NEEP+20].

In the medical field, EL is known as biomedical entity linking or medical entity linking, with a similar concept to general EL. This involves mapping text spans in biomedical texts to unique identifiers (medical concepts) in medical knowledge bases [FM23]. Research in normalizing popularized medical phrases to standard medical terms in a medical knowledge base (like SNOMED-CT) is known as medical concept normalization (MCN) [PAP+20]. In our thesis, we use both terms, where medical EL refers to the entire process and MCN is a part of this process, focusing on mapping popularized medical phrases to specialized medical terms in a medical knowledge base like SNOMED-CT. Our research primarily concentrates on MCN tasks, particularly developing medical EL for use in social media settings.

2.6.2 Research on Medical Concept Normalization

MCN has been implemented in several applications for improving patient care. For example, early detection of patients who require immediate treatment and medical support, such as depression [BMH17], digital disease surveillance [LHF+17], automated International Classification of Diseases (ICD) systems [WJW+20] and adverse drug reactions [ESNG18]. Aronson [Aro01] conducted early work on the dictionary-based method to build a tool called MetaMap. MetaMap is a linguistics-based tool, which aims to map biomedical text to the Unified Medical Language System (UMLS) Metathesaurus. This tool became the baseline for the next research on medical concept normalization. Dictionary-based methods obtain high precision but low recall. Some researchers extended the approach with additional features such as synonyms [LTMB17] and semantic features, such as semantic relatedness [ESNG18].

Due to the advances in machine learning and increasing the available source of annotated data, recent models rely on machine learning techniques. For the supervised learning approach, Limsopatham et al. [LC15] adapt phrase-based machine translation. The model maps the translation phrase to one of the concepts in SNOMED-CT based on the ranked similarity between vector representation of the translation output with concepts in the knowledge base. However, the model fails to deal with out of vocabulary (OOV) words.

Furthermore, Subramanyam and Sangeetha [SS20] treated the MCN task as a text classification task. They proposed a deep learning model, a BiLSTM model combined with deep contextualized and traditional word vectors, specifically highlighting the use of Embeddings from Language Models (ELMo) input features that can better learn phrase representations and predict concepts correctly compared to BiLSTM with only traditional embeddings.

Niu et al. [NYZ+19] proposed multi-task character-level attention network methods to encode the sentence with the character representations. They reported that character-

level encoding could reduce the OOV problems in the MCN task. The following research by Limsopatham et al. [LC16] outperforms the previous model by addressing the MCN task at the semantic level using deep learning and neural network to bridge the language gap between the popularized phrase and specialized medical term.

Lee et al. [LHF+17] improved the model from Limsopatham et al. [LC16] by removing the duplicate data from the original dataset and added various health-related datasets to enhance the embedding representation. This addition makes the vector have better semantic representation than the general pre-trained model. They also found that one phrase can have multiple specialized medical terms. Thus, they suggest MCN should be treated as a multi-class multi-label classification problem. Recent work employs deep learning and semantic representation for solving MCN. Tutubalina et al. [TMNM18] use a bidirectional RNN and GRU with attention on top of the embedding layer to transform the input phrases into a semantic vector representation. This RNN feature representation is appended with features extracted from the cosine similarity between the input phrase and medical concept from the UMLS knowledge base. They found that appending a set of semantic features to the model can improve the model performance. Miftahudinov et al. [MT19] use several models to transform the input phrases into a semantic vector representation. They use RNNs with pre-trained word embeddings: ELMo and BERT. To enhance the semantic representation, they also build a joint model to combine the deep learning vector representation with semantic similarity features between the vector representation of input phrases and the concept in the knowledge base.

More recently, Cao [CFZ22] introduced a multi-task approach that enriches the pre-trained BioBERT model to enhance MCN model performance. This method involves transforming not only the input phrase but also its contextual information (i.e., the full sentence containing the phrase) into BioBERT. Cao et al. [CFZ22] emphasized that this surrounding context provides supplementary information that can enhance the semantic meaning of the input phrase. Remy et al. [RSP23] proposed an approach by relying on a pre-trained model, BioLORD. BioLORD is a pre-trained model trained on UMLS. It a grounds medical concept representations using definitions, as well as short descriptions derived from a multi-relational knowledge graph of biomedical ontologies [RDD22]. The authors tried several pre-trained model such as PubMedBERT [GTC+20], and STAMB2[2]. The authors hypothesize that pre-training on general texts will help in understanding popularized medical phrases. Following this, they fine-tuned the pre-trained outputs using BioLORD.

The mentioned models [MT19, TMNM18, LC16, NYZ+19, CFZ22] treat the MCN task as a supervised classification task. Supervised classification has several shortcomings, particularly the labor of creating the training dataset and manually mapping it to medical concepts. Moreover, it failed to deal with the OOV problems. To overcome these shortcomings, Pattisapu et al. [PAP+20] develop a distant supervision model to generate training data automatically. The dataset is a pair of popularized medical phrases and

[2]https://huggingface.co/sentence-transformers/all-mpnet-base-v2

medical concepts in SNOMED CT. The popularized medical phrases are obtained from a patient discussion forum, and the medical concepts are obtained from MetaMap. This pair is determined by a similarity function, such as cosine similarity. The dataset is used to train the medical concept normalization by encoding the target knowledge [PPPV20]. Finally, during the inference, the model [PPPV20] finds the highest similarity between medical phrase and medical concept using a common embedding space.

The challenge of an end-to-end model for medical entity linking in social media data is the candidate generation. The candidate is a span of phrase that represents medical terms. In contrast with scientific publications [Kat15, GK14, LDl13, ZHZ+15], social media such as patient forums and question-answer websites have a vastly varied representation of phrases in terms of medical terms [PAP+20]. To solve this problem, datasets for the supervised approach have been created [KMJKW15, ZFP+19, LC16]. However, more data are needed in order for systems to perform better [PAP+20, TMNM18]. The latest medical entity linking approach [BLSC20] still suffers from a lack of sufficient regularity in the data due to the various ways of expressing popularized medical phrases in social media. Additionally, Cao et al. [CFZ22] highlighted another problem: the uneven distribution of classes. Certain concepts, notably in PsyTAR, are underrepresented with very few examples of popularized phrases.

In this thesis, we approach the MCN task as a supervised multi-class classification problem. Our approach aligns closely with the work of Tutubalina et al. [TMNM18]. As previously noted, Tutubalina et al. evaluated various deep neural networks, with a particular focus on RNN and GRU with attention mechanisms. Their experiments revealed that the GRU architecture, combined with HealthVec [MTT17] with TF-IDF for data representation, yielded best performance. In this research, we propose an MCN model architecture using GRU to process short, popularized medical phrases. Our approach incorporates both state-of-the-art embeddings, such as BERT-based models, and traditional embeddings like GloVe [PSM14]. A comprehensive comparison between our MCN model and current state-of-the-art methods is presented in Chapter 3 and Chapter 4.

2.7 Datasets of Medical Concept Normalization

This section provides an overview of the benchmark datasets for tasks related to Medical Concept Normalization (MCN). As a reminder, the MCN task involves the mapping of word(s) or phrase(s) that are identified as popularized medical phrases to corresponding concepts (entities) within a knowledge base, such as SNOMED-CT (as illustrated in Figure 1.1). There are various data sets available for handling the MCN task, namely CSIRO Adverse Drug Event Corpus (CADEC) [KMJKW15], Psychiatric Treatment Adverse Reaction (PsyTAR) [ZFP+19], and COMETA [BLSC20]. Each of these datasets provides examples of popularized medical phrases that correspond to specific medical concepts. We refer to these examples as *support phrases*.

CADEC The CSIRO Adverse Drug Event Corpus (CADEC) is a collection of medical forum posts where laypeople share their experience with Adverse Drug Events (ADE)[3] [KMJKW15]. These posts are often written in informal language and ignore the standard English grammar. CADEC focuses on ADEs related to anti-inflammatory drugs, specifically divided into two categories: Diclofenac and Lipitor. The corpus includes popularized medical phrases linked to medical concepts from widely-used medical vocabularies, particularly the SNOMED-CT and Australian Medical Terminology (AMT) [KMJKW15]. The annotators have identified various potential popularized medical phrases, such as *blurred vision*, and subsequently linked them to corresponding medical concepts, such as *hazy vision*, in SNOMED-CT. This corpus consists of 1,250 user posts and is labeled with predefined disease-related categories: ADR, Disease, Symptoms, and Clinical Finding.

PsyTAR Another publicly accessible dataset utilized for MCN tasks is the Psychiatric Treatment Adverse Reaction dataset (PsyTAR) [ZFP+19]. Similar to CADEC [KMJKW15], PsyTAR encompasses 887 drug review posts that discuss patient-reported Adverse Drug Events (ADE). In PsyTAR, each user post has been annotated for Adverse Drug Reactions (ADR), withdrawal symptoms, drug indications, as well as signs/ symptoms/illnesses. This dataset focuses on ADEs related to psychiatric drugs, including medications like Zoloft and Lexapro from the Selective Serotonin Reuptake Inhibitor (SSRI) class, as well as Cymbalta and Effexor XR from the Serotonin Norepinephrine Reuptake Inhibitor (SNRI) class.

COMETA This dataset is one of the largest public social media corpora and comprises English biomedical entity mentions sourced from Reddit. These mentions were annotated by experts and linked to SNOMED-CT [BLSC20]. The corpus's construction involved leveraging over 800,000 user posts from 68 diverse subreddits. In contrast to CADEC and PsyTAR, which focus around particular drugs or substances, COMETA collects data from diverse subreddits, extending beyond the scope of Adverse Drug Events (ADEs). The annotated entities cover a wide range of concepts including symptoms, disease, anatomical expressions, chemicals, genes, devices and procedures across a range of conditions. The experts annotated each mention by both general and specific level of SNOMED-CT concepts. The *General* level is concerned with the literal meaning of the terms, while the *Specific* level takes into account the context in which the entity appears. The dataset is accessible to the public[4]. All of these datasets are explaining *disease*, *symptoms*, and *drug-related issues* taken from social media, AskaPatient.com or Reddit. Most of the entities in the datasets are linked to the UMLS Metathesaurus using SNOMED-CT for *disease*, *symptoms*, and *findings*. Table 2.1 shows the statistics distributions of each datasets. Among the datasets, COMETA covers broader concepts and larger a number of entities than both CADEC and PsyTAR. The combined datasets (CADEC, PsyTAR, and COMETA) encompass less than 10% of the concepts from SNOMED-CT, with around 4,500 concepts out of 350,000 concepts. Furthermore, in these datasets, 60% of

[3] https://www.askapatient.com/

[4] https://github.com/cambridgeltl/cometa/tree/master/data

Table 2.1: Dataset Statistics for CADEC[KMJKW15], PsyTAR[ZFP+19], and COMETA[BLSC20].

	CADEC	PsyTAR	COMETA
Posts	1,250	887	800,000
Sentences	7,632	6,009	-
Communities	2 threads	2 threads	68 threads
Entities	9,111	6,556	20,015
Unique Concepts	1,029	755	General: 3,645 Specific: 4,003

[*] COMETA did not provide full sentences from Reddit posts collected for each entity, which means we are unable to calculate the total number of sentences.

the specialized medical terms have less than 4 popularized medical phrases mapped to them, as shown in Figure 2.2. This indicates that the issue of data scarcity remains a challenge. Table 2.2 provides statistics on the number of concepts with a limited number of supporting popularized phrases, specifically ranging from one to four support phrases and Table 2.3 shows examples of medical concepts with corresponding supporting phrases.

Table 2.2: Distribution of medical concepts based on the number of supporting phrases, ranging from one to four supporting phrases

Number of Support Phrases	Count of Medical Concepts
1	456
2	182
3	1,989
4	186
Total Concepts	**2,813**

2.8 Natural Language Processing in Low-Resource Scenarios

We define the term low-resource as the lack of annotated datasets that cover various forms of text in laypersons' language, disease category, and concepts from a specialized medical meta-thesaurus for medical entity linking. In other words, we have a data scarcity situation. Data augmentation [KPT+20] and distant supervision [PAP+20] have the similar purpose of tackling the data scarcity problems in supervised learning for low-resource scenarios.

Textual data augmentation addresses the data scarcity problem in supervised learning

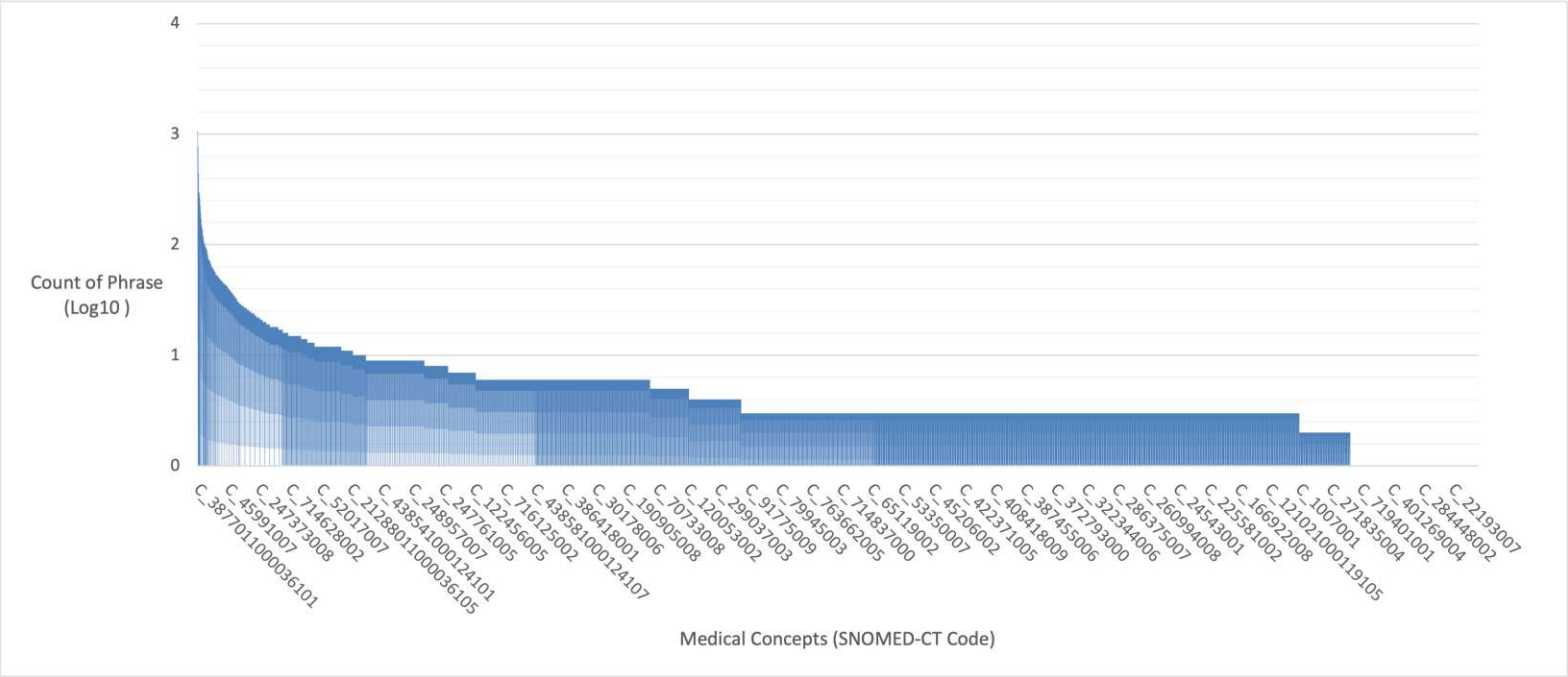


Figure 2.2: Distributions of phrases per medical concept. We applied a log transformation to normalize the phrase counts for each medical concept. The diagram shows a left-skewed distribution, which illustrates that some medical concepts have a large number of supporting phrases, and decreases towards the right, where many concepts have fewer supporting phrases.

Table 2.3: Examples of medical concepts with limited supporting phrases and corresponding support phrases

Number of Support Phrases	SNOMED-CT Code	SNOMED-CT Description	Example of Support Phrases
4	85898001	Cardiomyopathy	*cardiomyopathy, significant heart muscle weakness*
	85799002	Amnemonic aphasia	*loss of vocabulary, can't think of the right words to use, at a loss for words quite often, unable to recall names*
3	95847005	Injury of muscle	*permanent, irreversible muscle damage, muscle damage*
	95421005	Intercostal myalgia	*muscles pain on the left side of my chest, pain in side muscles on one side of chest, muscles on the sides of my back were constantly painful*
2	8009008	Nocturnal enuresis	*bed wetting, can't even wake up to use the bathroom*
	79471008	Sudden hearing loss	*totally deaf in left ear, sudden hearing loss experience*
1	84769001	Cervical vertigo	*turn my head my brain is a bit behind*
	82470000	Muscle fasciculation	*fasciculations (muscle twitches)*

tasks by adding automatically labelled data to the existing datasets [KPT+20]. Data augmentation can be applied by various techniques. The first technique is back-translation [SHB15, FBM17, CG20], which translates a sentence into another language and back to the original language to represent the sentence in a different form with the same meaning. The next technique is by replacing a word with a synonym [WZ19], or the same entity type [DA20]. In terms of replacement, a language model (LM) can be used to provide a better suggestion by the masking technique [GR20], or contextual word probability [Kob18]. Another method is swapping the words in the sentence [WZ19]. Wei and Zou [WZ19] randomly choose two words in the original sentence and swap their positions to generate new data. Moreover, data augmentation also can be conducted based on the syntactic level of the sentence, such as the dependency tree [VKSL19]. Vania et al.

[VKSL19] proposed two methods to generate the new sentence: the first is to rotate parts of the dependency tree and the second is removing parts of the dependency tree from the original sentence.

Knowledge bases [RZNC20] and language models [DLB+20] are also beneficial for data augmentation. For instance, Tikhomirov et al. [TLSD20] proposed a data augmentation technique using word replacement based on a cybersecurity knowledge base. Kang et al. [KPT+20] extended the Easy Data Augmentation (EDA) from Wei and Zou [WZ19] by incorporating Unified Medical Language System (UMLS) knowledge, and called the method UMLS-EDA. One of the UMLS-EDA methods is the Synonym Replacement with UMLS that tries to find UMLS concepts in a medical literature sentence and replace them with randomly selected synonym from UMLS. In a recent study, Scaboro et al. [SPC+22] proposed a different data augmentation approach by manually altering selected original tweets to create negations and speculations in an adverse drug event dataset. In contrast, Falis et al. [FDBA22] adapted Kang et al.'s [KPT+20] synonym replacement technique, expanding the synonym dictionary source from UMLS to include ICD-9, ICD-10, and SNOMED-CT. They replaced medical concepts detected by the NER+Linking model with synonyms randomly selected from these sources. These approaches [TLSD20, KPT+20, FDBA22, SPC+22] demonstrate the effectiveness of using knowledge bases and language models from different domains, such as cybersecurity and medical, to address the challenge of data scarcity.

In contrast to data augmentation, distant supervision aims to generate a labeled data set from an unlabeled data set. The corresponding labels are collected from external sources through a semi-automatic process. Distant supervision is a popular approach for information extraction tasks, such as Named Entity Recognition (NER) or Relation Extraction (RE) where the labels can be obtained from the knowledge base, gazetteers, and another external source is used to obtain the external information [MBSJ09, LBHT20, CHC+19, LAS19, HK18, DWK19].

Distant supervision has also been used in various tasks to perform an automatic annotation process, such as sentiment analysis [BGdLG18, AMU17] using an emoticon dictionary. Vashishth et al. [VJN+20] create a training data set for medical entity linking from PubMed abstracts to extract medical entities automatically. The entities are retained when the abstract's spans exactly match with the entity in Medical Subject Heading (MeSH).

Moreover, distant supervision has also been applied in MCN. Pattisapu et al. [PAP+20] performed distant supervision by obtaining medical phrases from medical forums using text classification. To obtain the corresponding medical concept, they calculate the semantic similarity between medical phrases and medical concepts. This approach still produces a mismatch between medical phrases and medical concepts due to the ambiguity of popularized medical phrases [PAP+20].

2.9 Summary

This chapter provides an overview of the background of this research and the relevant literature related to the thesis topic. The chapter begins with a discussion of the background to this work and includes the definition of functional health literacy (FHL). Next, the chapter describes the tools used to assess the Functional Health Literacy (FHL) of laypeople. Following this, we define the medical terminology utilized in this thesis. Related work is then discussed, focusing on existing tools developed to improve laypeople's medical terminology knowledge, the definition of Medical Entity Linking (MEL) and Medical Concept Normalization (MCN), and research challenges in MCN tasks, available datasets for MCN tasks, and natural language processing in low-resource scenarios.

The literature review begins by addressing tools designed to enhance laypeople's understanding of medical terminology. Many of these studies concentrated on empowering patients or non-experts to better understand medical documents, such as Electronic Health Records (EHRs), Personal Health Records (PHRs), or radiology reports. This was achieved by translating complex medical terminology into simpler synonyms and providing concise explanations that link the medical concepts with their associations to other types of medical concepts, such as *Pulmonary atresia (a type of birth defect) and oropharyngeal (e.g. mouth)* [KCZT10]. In contrast to these research, our approach introduces a model that supports laypeople in enhancing their comprehension of medical terminology by extending the objectives of the Medical Concept Normalization (MCN) task to be more beneficial for non-experts, as elaborated in the motivation presented in Chapter 1.

We also discuss related work on the development of MCN models and explore available datasets for MCN tasks. Many researchers have approached MCN as a classification task, utilizing deep learning methods. We emphasize the research challenges posed by data scarcity issues, particularly given the limited coverage of medical concepts in existing datasets, especially in SNOMED-CT. Additionally, some medical concepts lack popularized medical phrases. To address these challenges, our research has aimed to tackle the problem of limited medical datasets. To do so, we reviewed research on the topic of natural language processing in low-resource settings, specifically focusing on semi-automated methods that can be beneficial for generating additional datasets.

CHAPTER 3

Data Augmentation for Popularized Medical Phrase Extraction

Automatic detection of medical phrases in social media postings is challenging not least because of the lexical and grammatical diversity of terminology used by individuals that have varied backgrounds and expertise levels [MYS+19]. Additional difficulties include the widespread use of informal language, typographic and abbreviation errors, and non-standard syntax [PAP+20, MYS+19]. New colloquial medical terms are continually emerging, which is also a consequence of the complexity of the language, in general, and of the language in online communication in particular [KRS16]. As a result, the laypersons' health vocabularies, that is the terminology and discourse used by laypersons, has little overlap with the vocabulary used by medical experts and medical documents.

Current publicly available data sets for Medical Concept Normalization (MCN) [KMJKW15, BLSC20, ZFP+19] contain mappings between layperson phrases and medical expert phrases. However, less than 10% of the concepts listed in the Systematized Nomenclature of Medicine - Clinical Terms (SNOMED-CT) occur in these data sets. Furthermore, in these data sets, 60% of the expert medical phrases have less than 4 popularized medical phrases (i.e *support phrases*) mapped to them (see Table 2.2). Together these data sets contain a little over 4500 SNOMED-CT concepts. This is supported by findings by Cao et al. [CFZ22] that certain medical concepts in PsyTAR are supported with very few examples of popularized medical phrases. As deep learning models are data hungry, their use in the biomedical natural language processing is limited as these publicly available data sets are small. Extending such data sets to be able to trained deep learning models is expensive as annotations must be done with expert help.

In this chapter we present a set of data augmentation techniques for textual data that

we apply to laypersons' texts in the informal medical entity linking. These methods are designed to generate more data by simulating the writing behavior of laypersons.

The first set of techniques is based on typographic and grammatical errors in their writing (e.g. keyboard errors, misspelling, misplaced words). A second set of augmentation techniques is based on semantic information used for word replacement. We replace words with synonyms and with semantically similar words that are found by means of distributional word representations (i.e. word embeddings). A third technique for textual data augmentation uses a paraphrasing engine that is based on transformer architecture to generate new data. We apply the augmentation methods to the data used for the Named Entity Recognition (NER) and the MCN tasks which are part of our MEL model (Figure 1.1), and investigate their impact on them.

The contributions of this chapter are as follows:

- We proposed a data augmentation techniques to increase the number of medical phrases used by laypeople in the MCN task, which translates to more data available for training.
- We evaluated the impact of data augmentation on both NER and MCN tasks by training the models and comparing their performance against a baseline.

The remaining content of this chapter is organized as follows: Section 3.1 outlines our methodology. Next, Section 3.2 details the experiment setup. Our results and discussions regarding the experiments can be found in Section 3.3. We conclude the chapter with a summary in Section 3.4.

3.1 Methodology

The datasets we use in this work are CADEC [KMJKW15], PsyTAR [ZFP+19] (explained in Section 2.7), and additional dataset, MedRed [SMLQB20], a dataset used for the extraction of medical entities in social media (NER task). Each dataset contains posts written by laypeople on medical health forums. We explain in this section the augmentation techniques that we use to augment the data from the datasets [KMJKW15, ZFP+19, SMLQB20]. We train two separate models for the two tasks (NER and MCN) using training data that has been augmented with these methods. We analyze the models' effectiveness on the test data provided by the original datasets, which has been kept hidden from the model training, by looking at F1 measure scores.

3.1.1 Data

We demonstrate our data augmentation methods on the available datasets (presented in Section 2.7) with additional datasets (e.g., MedRed). We used different sets of data to train the two sub-tasks of informal medical entity linking: NER and MCN tasks.

There are two available datasets for identifying medical NER for laypersons' language: (1) CADEC is an annotated dataset on patient-reported Adverse Drug Events (ADE), which contains 1,250 annotated texts from the "Ask a Patient" forum[1] [KMJKW15]. The dataset has been previously used to identify adverse drug reactions posted on social media, see, for example, Tutubalina et al. [TN17]; (2) MedRed, a dataset for extracting medical entities in social media, which contains 1,977 Reddit posts [SMLQB20]. This dataset was used to extract disease, symptoms, and drug names.

For the MCN task, we use two datasets: (1) PsyTAR, a corpus related to ADE and psychiatric medication effectiveness, which contains 887 reviews from the "Ask a Patient" forum [ZFP+19]; and (2) CADEC, which we use as a baseline dataset not only for the NER task but also for the MCN task. CADEC contains medical terms associated with medical concepts from standard medical vocabularies, more specifically, from the SNOMED-CT and Australian Medical Terminology (AMT) [KMJKW15]. All of these datasets will be used as baseline datasets which refers to the original, unaugmented dataset that we use to train our initial model. This model then serves as a point of comparison for evaluating the performance of models trained on augmented datasets for NER and MCN tasks. Table 3.1 gives details on the datasets' content, describing the number of social posts per dataset, the number of sentences, the number of threads dedicated to specific medical topics of discussions, and the number of medical entities.

Table 3.1: CADEC, MedRed, and PsyTAR dataset statistics

Number of/Types of	CADEC	MedRed	PsyTAR
posts	1,250	1,977	887
sentences	8,575	9,190	6,009
discussion forums	2 forums	18 forums	2 forums
all entity types	7,906	3,768	-
DIS type	6,151	2,931	-
DRUG type	1,755	837	-
popularized terms/mentions	9,111	-	6,556
SNOMED-CT concepts	1,029	-	755
task	NER&MCN	NER	MCN

3.1.2 Data Augmentation for NER and MCN

We propose several data augmentation techniques for the NER and MCN tasks. These techniques aim to generate more data to increase the volume of training data for our models. The augmentation techniques we employ mimic laypeople's writing behavior. We divide the techniques into *character augmentation*, *word augmentation*, and *paraphrasing*.

Character-based augmentation is inspired by the typographical and orthographical errors that occur during written communication [MWM00]. Typographical spelling errors

[1] https://www.askapatient.com/

are a form of mistyping (typing a neighbouring letter incorrectly on the keyboard). Orthographical errors are forms of error in guessing or selecting the wrong word, such as *to* instead of *two* or *too*. We also add Optical Character Recognition (OCR) errors in our character augmentation. OCR is used to recognize text from an image source. When images are automatically converted into text, OCR algorithms can produce noise in recognizing characters, such as 'o' and '0'.

Word-based augmentation aims to augment the popularized medical phrase(s) at word level and we use several techniques: synonym, hypernym replacement, hyponym replacement, and swap words.

Paraphrase-based augmentation aims to generate new colloquial medical terms by paraphrasing the original popularized medical phrases. We used a paraphrasing engine based on a *Text-to-Text Transfer Transformer* fine-tuned on the Google PAWS data set [ZBH19, YZTB19]. The engine is used to paraphrase the popularized medical phrases, for example, *hard to get out of bed in the morning* augmented to be *getting from bed in the morning is hard*.

Semantic Mention Replacement augmentation aims to replace the named entity of another with the same type. This method was inspired by the *mention replacement* proposed by Dei and Adel [DA20]. We applied this method to the NER task only because it depends on the consistency of entity types. For example, replacing one disease entity with another disease entity. Therefore, the replacements are performed with entities of the same type, making it suitable specifically for the NER task.

Table 3.2 describes each of the augmentation approaches used in this work, by the NER and MCN tasks for MEL. Table 3.3 illustrates the augmentation approaches used in a NER task. These techniques are equally applicable in MCN tasks that focus solely on augmenting data at the phrase level. We notice from the examples in Table 3.3 that augmentation may lead to different lengths of the popularized medical phrases of interest: (*pain in my foot*) becomes shorter (*my foot aches*) or longer (*pain sense in my foot*). This affects the label sequence of the original sentence for the NER task.

We treat the NER task as a sequence labeling task, where the sequence (one or more words) corresponds to a specific entity type. We use the *BIO* tagged format with types, where *B* marks the beginning word of the sequence, *I* marks words that are not the beginning of the sequence (i.e. are inside the sequence), while *O* marks words not part of the sequence of interest (outside). Each word in the NER dataset that has either a *B* or *I* tag is additionally tagged with either *DIS* or *DRUG* types, where *DIS* marks a disease/symptom and *DRUG* marks a drug. For example, the phrase *bit drowsy* is labeled *B-DIS I-DIS*, and the other words that do not represent the medical terms are labeled as *O*.

3.1.3 Model Training

We train one model for each of the NER and MCN tasks using contextual embedding and Recurrent Neural Network (RNN) deep learning.

Table 3.2: Detailed descriptions of the augmentation methods employed in this work, per MCN and NER task. The first row for each technique describes the concrete augmentation operation for the MCN task, while the second row describes additional rules that must be considered in order to maintain the BIO tag necessary for the NER task. We refer to *mentions* as named-entities, and *label sequence* as BIO tag.

Technique (Category)	Detail
Keyboard Errors (Character)	MCN: Replace one character at random, simulating a common typing error caused by the character's position on the keyboard, which is typically very close to the one replacing the original character.
	NER: Similar replacement as above, only targets mentions without changing the label sequence.
OCR Errors (Character)	MCN: Replace a character at random, simulating an OCR error. The replacement is done using an existing list of common OCR mistakes [Ma19].
	NER: Replacement similar to the MCN task, only targeting mentions without changing the label sequence.
Misspelling Errors (Character)	MCN: Replace a character at random, simulating a misspelling error, according to a pre-defined list of errors in the misspelling corpus [Nie20].
	NER: Replacement similar to the MCN task, only targeting mentions without changing the label sequence.
WordNet Synonym Replacement (SR) (Word)	MCN: Randomly replace word(s) in with their synonym extracted from WordNet
	NER: Replacement similar to the MCN task, only targeting mentions. We reconstruct the label sequence (Sub Section 3.1.3) using the length of the new phrase.
CHV-Drug SR (Word)	MCN: Randomly replace word(s) in the phrase with synonyms from the Consumer Health Vocabulary (CHV) [ZT06, VMHZ14] and Drug Bank [WFG+18]. If there are multiple words in the input sentence that are a match for synonyms, we replace the longer phrase. The matching is done by n-gram matching.
	NER: Replacement similar to the MCN task, only targeting mentions. We reconstruct the label sequence (Section 3.1.3) using the length of the new phrase.
Hypernym Replacement (Word)	MCN: Randomly replace phrase word(s) with hypernyms fetched from WordNet.
	NER: Replacement similar to the MCN task, only targeting mentions. We reconstruct the label sequence (Section 3.1.3) using the length of the new phrase.
Hypnonym Replacement (Word)	MCN: Randomly replace phrase word(s) with hyponyms fetched from WordNet.
	NER: Replacement similar to the MCN task, only targeting mentions. We reconstruct the label sequence (Section 3.1.3) using the length of the new phrase.
Swap Word (Word)	MCN: Choose two words at random from the input phrase and swap their positions.
	NER: Replacement is similar to the MCN task, only targeting mentions.
Semantic Mention Replacement (Semantic MR)	MCN: not applicable.
	NER: Replace the mention with another of the same type. The replacement is determined by calculating the mention's highest similarity score to similar entities in the corpus. To perform semantic textual similarity, we use the Sentence-BERT for sentence representation.
Paraphrase	MCN: Generate a new phrase based on the input one using a paraphrasing engine based on a T5 model trained on the Google PAWS Dataset.
	NER: Replacement similar to the MCN task, only targeting mentions. We reconstruct the label sequence (Section 3.1.3) using the length of the new phrase.

Table 3.3: Augmentation example. The bold face words mark the input sequence, the italic, blue-colored words indicate the augmentation based changes to the input sentence.

Original Sentence	I find that the **pain in my foot**[**DIS**] is subsiding and i can walk a lot better.
Original word sequence	[“I”, “find”, “that”, “the”, **“pain”**, **“in”**, **“my”**, **“foot”**, “is”, “subsiding”, “and”, “i”, “can”, “walk”, “a”, “lot”, “better”, “.”]
Original tag sequence	[“O”, “O”, “O”, “O”, “B-DIS”, “I-DIS”, “I-DIS”, “I-DIS”, “O”, “O”, “O”, “O”, “O”, “O”, “O”, “O”, “O”, “O”]
Keyboard Errors	Words: [“I”, “find”, “that”, “the”, **“pain”**, ***“im”***, **“my foot”** , “is”, “subsiding”, “and”, “i”, “can”, “walk”, “a”, “lot”, “better”, “.”]
	Tags: [“O”, “O”, “O”, “O”, “B-DIS”, “I-DIS”, “I-DIS”, “I-DIS”, “O”, “O”, “O”, “O”, “O”, “O”, “O”, “O”, “O”, “O”]
OCR Errors	Words: [“I”, “find”, “that”, “the”, **“pain”**, **“in”**, **“my”**, ***“fo0t”***, “is”, “subsiding”, “and”, “i”, “can”, “walk”, “a”, “lot”, “better”, “.”]
	Tags: [“O”, “O”, “O”, “O”, “B-DIS”, “I-DIS”, “I-DIS”, “I-DIS”, “O”, “O”, “O”, “O”, “O”, “O”, “O”, “O”, “O”, “O”]
Misspelling Errors	Words: [“I”, “find”, “that”, “the”, ***“psin”***, **“in”**, **“my”**, **“foot”**, “is”, “subsiding”, “and”, “i”, “can”, “walk”, “a”, “lot”, “better”, “.”]
	Tags: [“O”, “O”, “O”, “O”, “B-DIS”, “I-DIS”, “I-DIS”, “I-DIS”, “O”, “O”, “O”, “O”, “O”, “O”, “O”, “O”, “O”, “O”]
Semantic MR	Words: [“I”, “find”, “that”, “the”, ***“severe”, “foot”, “pain”*** , “is”, “subsiding”, “and”, “i”, “can”, “walk”, “a”, “lot”, “better”, “.”]
	Tags: [“O”, “O”, “O”, “O”, “B-DIS”, “I-DIS”, “I-DIS”, “O”, “O”, “O”, “O”, “O”, “O”, “O”, “O”, “O”, “O”, “O”]
Paraphrase	Words: [“I”, “find”, “that”, “the”, ***“my”, “foot”, “aches”*** , “is”, “subsiding”, “and”, “i”, “can”, “walk”, “a”, “lot”, “better”, “.”]
	Tags: [“O”, “O”, “O”, **“B-DIS”, “I-DIS”, “I-DIS”**, “O”, “O”, “O”, “O”, “O”, “O”, “O”, “O”, “O”, “O”, “O”]
WordNet SR	Words: [“I”, “find”, “that”, “the”, ***“hurting”, “in”, “my”, “foot”*** , “is”, “subsiding”, “and”, “i”, “can”, “walk”, “a”, “lot”, “better”, “.”]
	Tags: [“O”, “O”, “O”, “O”, “B-DIS”, “I-DIS”, “I-DIS”, “I-DIS”, “O”, “O”, “O”, “O”, “O”, “O”, “O”, “O”, “O”, “O”]
CHV-Drug SR	Words: [“I”, “find”, “that”, “the”, ***“pain”, “sense”, “in”, “my”, “foot”*** i, “is”, “subsiding”, “and”, “i”, “can”, “walk”, “a”, “lot”, “better”, “.”]
	Tags: [“O”, “O”, “O”, “O”, “B-DIS”, **“I-DIS”**, “I-DIS”, “I-DIS”, “I-DIS”, “O”, “O”, “O”, “O”, “O”, “O”, “O”, “O”, “O”, “O”]
Hypernym Replacement	Words: [“I”, “find”, “that”, “the”, “pain”, “in”, “my”, ***“walk”*** , “is”, “subsiding”, “and”, “i”, “can”, “walk”, “a”, “lot”, “better”, “.”]
	Tags: [“O”, “O”, “O”, “O”, “B-DIS”, “I-DIS”, “I-DIS”, “I-DIS”, “O”, “O”, “O”, “O”, “O”, “O”, “O”, “O”, “O”, “O”]
Hyponym Replacement	Words: [“I”, “find”, “that”, “the”, “pain”, “in”, “my”, ***“flatfoot”*** , “is”, “subsiding”, “and”, “i”, “can”, “walk”, “a”, “lot”, “better”, “.”]
	Tags: [“O”, “O”, “O”, “O”, “B-DIS”, “I-DIS”, “I-DIS”, “I-DIS”, “O”, “O”, “O”, “O”, “O”, “O”, “O”, “O”, “O”, “O”]
Swap Word	Words: [“I”, “find”, “that”, “the”, ***“in”, “pain”***, “my”, “foot”, “is”, “subsiding”, “and”, “i”, “can”, “walk”, “a”, “lot”, “better”, “.”]
	Tags: [“O”, “O”, “O”, “O”, “B-DIS”, “I-DIS”, “I-DIS”, “I-DIS”, “O”, “O”, “O”, “O”, “O”, “O”, “O”, “O”, “O”, “O”]

NER Task. Figure 3.1 shows the outline of our training architecture. The model takes user-posted sentences as input, represented in the (Conference on Natural Language Learning) CoNLL format[2] [TKSDM03]). Initially, we tokenize these sentences and label them using the *BIO* tag scheme. These tokens are then sent through an embedding layer to create contextual representations as the feature representations. We concatenate GloVe [PSM14] and RoBERTa [LOG+19] embeddings, which allows us to capture richer information [ABV18]. GloVe helps in capturing word-level relationships, while RoBERTa captures contextual information from the entire input text. These embeddings serve as the input for the Bi-LSTM layer, followed by the CRF layer, which generates the final tags. This architecture, originally proposed by Huang et al. [HXY15], has been shown to be effective in extracting entities for NER tasks [ABV18, SMLQB20]. In the end, our Named Entity Recognition (NER) model returns the input text annotated with *BIO* tags.

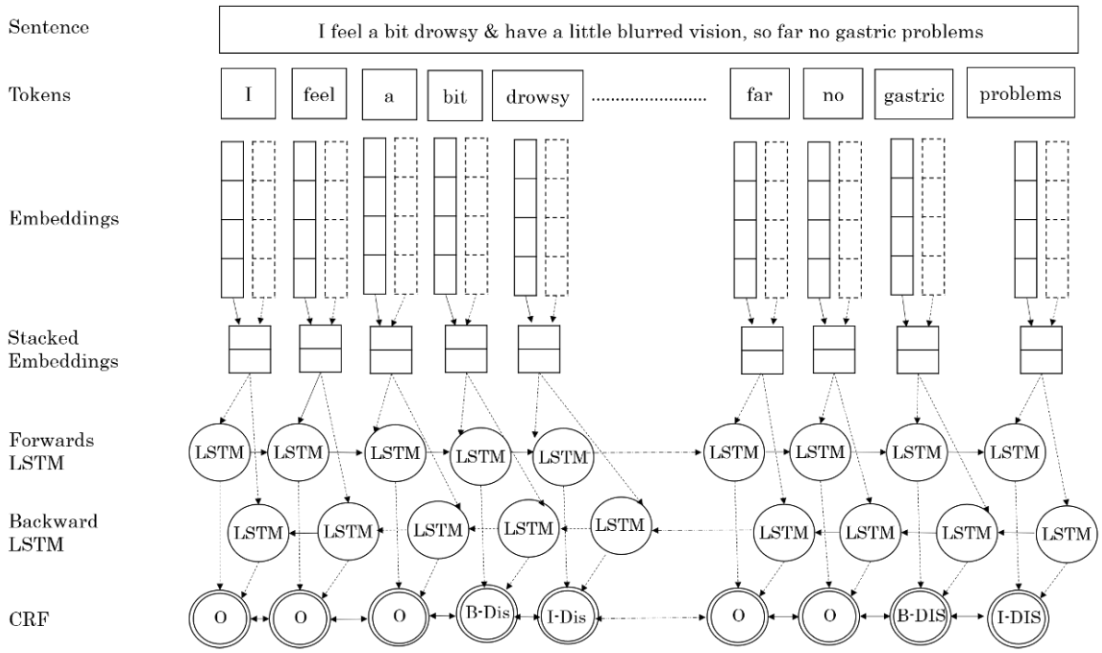


Figure 3.1: The architecture used for training the Named Entity Recognition (NER) model.

MCN Task. We handle this task as a multi-class classification task. The idea is to classify popularized medical phrases into specialized medical concepts, represented by SNOMED-CT codes as class labels. Similar to the NER task, we use the concatenated GloVe and RoBERTa word embeddings. We use, then, Gated Recurrent Units (GRU) [CvMG+14] to learn the SNOMED-CT class labels for medical phrases. We chose GRU

[2]This format typically includes one word per line with annotations. Sentences are separated by a blank line (detailed in `https://universaldependencies.org/docs/format.html`.

over LSTM to better cater to the short input of popularized medical entities. Figure 3.2 presents the model architecture used to train the MCN model. Once the embedding for each word in the sequence is computed, these embeddings are passed to the GRU, which includes a reset gate and an update gate. This enables the model to capture context both before and after each word, resulting in a more comprehensive representation of the sequence. Finally, a softmax layer is applied to perform multi-class classification, determining the SNOMED-CT code class to which the input phrase (i.e popularized medical phrase) belongs.

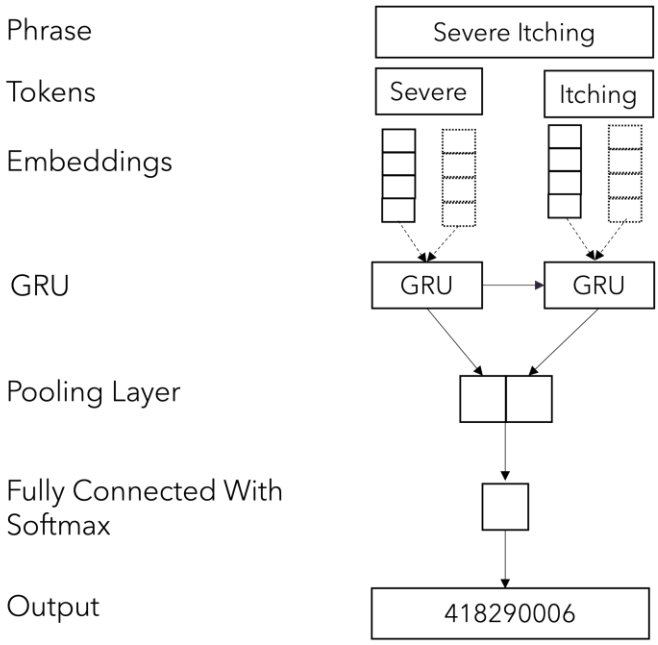


Figure 3.2: The architecture used for training the Medical Concept Normalization (MCN) model.

When the MCN is integrated into a workflow of a complete pipeline of informal medical entity linking, as shown in Figure,1.1, its input can be set as the output from the NER, detailed in Chapter 5. To demonstrate the data augmentation, we utilize labeled data from the CADEC and PsyTAR datasets to train the MCN. The resulting output from this trained model is a classification of popularized medical phrases according to the SNOMED-CT codes.

3.2 Experiment Setup

We designed several experiments to evaluate the effect of the proposed data augmentation methods on the informal medical entity linking by comparing the baseline models, trained with the non-augmented data, with models that use the augmented data sets in the training phase, particularly for medical NER and MCN. We took the original fold of

training, validation, and testing sets provided by the authors of CADEC [KMJKW15], MedRed [SMLQB20], and PsyTAR [ZFP+19]. We only augment the training sets, leaving out the validation sets for evaluating the model training and the testing sets for demonstrating our model's performance. We used CADEC and MedRed datasets to perform medical NER tasks and the CADEC and PsyTAR datasets for the MCN tasks.

Experiment Setup for NER Task We analyze the impact data augmentation has on the NER task by augmenting different amounts of the baseline training data. This approach is influenced by the study conducted by Dai and Adel [DA20], where they explored the impact of training data size in low-resource settings by dividing their training data into small, medium, and large subsets. In our experiment, we randomly sampled 20% , 50%, and 100% of training data as the source of augmentation, applying all the augmentation techniques described in Section 3.1.2 (see also Table 3.2).

For each applied technique in the character and word augmentation categories, we augment a popularized medical phrase only once. For semantic mention and paraphrase augmentation approaches, we augment a term three times to have a balance of the various annotation categories in the training data (e.g. three character-based annotation techniques vs. one Semantic Mention Replacement technique). The combination of original and augmented training data is used to train the NER model. We repeat the training process five times with different random seeds, to detect patterns in our model's performance.

We also explored augmenting the context around entity mentions, where, *context* refers to the token (e.g., word) labeled with the "O" tag. For this experiment, we applied the augmentation techniques listed in Table 3.2, focusing specifically on *character-based* and *word-based augmentation*. Table 3.4 provides an example of how we implemented these data augmentation techniques.

Additionally, we combined the augmented context with augmented entity mentions. This approach created a more diverse set of training sentences, allowing the NER model to learn from a wider variety of examples and potentially improve the model's performance. We conducted two sets of experiments:

1. GloVe + RoBERTa + biLSTM-CRF + Data Augmentation: our proposed model enhanced by different data augmentation methods.

2. BERT-biLSTM-CRF + Data Augmentation: we compared our proposed model with the model architecture from [KPT+20], where BlueBERT-base, pretrained on PubMed literature and MIMIC-III (Medical Information Mart for Intensive Care-III)[PYL19] used as to represent the data.

Table 3.4: Text augmentation examples at different levels and techniques. Augmentation details: *Blue italic* indicates augmented changes. At context-level and combination level, CHV-Drug SR example omitted due to lack of suitable terms.

Augmentation Level	**Augmentation Method**	**Augmented Text**
Original	Original Sentence	I find that the **pain in my foot**[**DIS**] is subsiding and I can walk a lot better.
Original	Original Word Sequence	[‘I’, ‘find’, ‘that’, ‘the’, **‘pain’**, **‘in’**, **‘my’**, **‘foot’**, ‘is’, ‘subsiding’, ‘and’, ‘I’, ‘can’, ‘walk’, ‘a’, ‘lot’, ‘better’, ‘.’]
Original	Original Tag Sequence	[‘O’, ‘O’, ‘O’, ‘O’, ‘B-DIS’, ‘I-DIS’, ‘I-DIS’, ‘I-DIS’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’]
Context-Level	Keyboard Errors	Words: [‘I’, ***fund***, ‘that’, ‘the’, *‘pain’*, *‘in’*, *‘my’*, *‘foot’*, ‘is’, ‘subsiding’, ‘and’, ‘I’, ‘can’, ‘walk’, ‘a’, ‘lot’, ‘better’, ‘.’]
		Tags: [‘O’, ‘O’, ‘O’, ‘O’, ‘B-DIS’, ‘I-DIS’, ‘I-DIS’, ‘I-DIS’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’]
	OCR Errors	Words: [‘I’, ‘find’, ‘that’, ‘the’, *‘pain’*, *‘in’*, *‘my’*, *‘foot’*, ‘is’, ‘subsiding’, ‘and’, ‘I’, ‘can’, ‘walk’, ‘a’, ‘lot’, ***b3tter***, ‘.’]
		Tags: [‘O’, ‘O’, ‘O’, ‘O’, ‘B-DIS’, ‘I-DIS’, ‘I-DIS’, ‘I-DIS’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’]
	Misspelling Errors	Words: [‘I’, ‘find’, ‘that’, ‘the’, *‘pain’*, *‘in’*, *‘my’*, *‘foot’*, ‘is’, ‘subsiding’, ‘and’, ‘I’, ‘can’, ***wakl***, ‘a’, ‘lot’, ‘better’, ‘.’]
		Tags: [‘O’, ‘O’, ‘O’, ‘O’, ‘B-DIS’, ‘I-DIS’, ‘I-DIS’, ‘I-DIS’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’]
	WordNet SR	Words: [‘I’, ‘find’, ‘that’, ‘the’, *‘pain’*, *‘in’*, *‘my’*, *‘foot’*, ‘is’, ‘subsiding’, ‘and’, ‘I’, ‘can’, ‘walk’, ‘a’, ‘lot’, ***improve***, ‘.’]
		Tags: [‘O’, ‘O’, ‘O’, ‘O’, ‘B-DIS’, ‘I-DIS’, ‘I-DIS’, ‘I-DIS’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’]
	Hypernym Replacement	Words: [‘I’, ‘find’, ‘that’, ‘the’, *‘pain’*, *‘in’*, *‘my’*, *‘foot’*, ‘is’, ‘subsiding’, ‘and’, ‘I’, ‘can’, ***travel***, ‘a’, ‘lot’, ‘better’, ‘.’]
		Tags: [‘O’, ‘O’, ‘O’, ‘O’, ‘B-DIS’, ‘I-DIS’, ‘I-DIS’, ‘I-DIS’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’]
	Hyponym Replacement	Words: [‘I’, ‘find’, ‘that’, ‘the’, *‘pain’*, *‘in’*, *‘my’*, *‘foot’*, ‘is’, ‘subsiding’, ‘and’, ‘I’, ‘can’, ‘walk’, ‘a’, ‘lot’, ***perfect***, ‘.’]

Continued on next page

Table 3.4 – continued from previous page

Augmentation Level	Augmentation Method	Augmented Text
	Swap Words	Tags: [‘O’, ‘O’, ‘O’, ‘O’, ‘B-DIS’, ‘I-DIS’, ‘I-DIS’, ‘I-DIS’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’] Words: [‘I’, ‘find’, ‘that’, ‘the’, *‘pain’*, *‘in’*, *‘my’*, *‘foot’*, ‘is’, ‘subsiding’, ‘and’, ‘I’, ***walk***, ***can***, ‘a’, ‘lot’, ‘better’, ‘.’]
		Tags: [‘O’, ‘O’, ‘O’, ‘O’, ‘B-DIS’, ‘I-DIS’, ‘I-DIS’, ‘I-DIS’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’]
Entity-Level	Keyboard Errors	Words: [‘I’, ‘find’, ‘that’, ‘the’, **‘pain’**, ***im***, **‘my’**, **‘foot’**, ‘is’, ‘subsiding’, ‘and’, ‘I’, ‘can’, ‘walk’, ‘a’, ‘lot’, ‘better’, ‘.’]
		Tags: [‘O’, ‘O’, ‘O’, ‘O’, ‘B-DIS’, ‘I-DIS’, ‘I-DIS’, ‘I-DIS’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’]
	OCR Errors	Words: [‘I’, ‘find’, ‘that’, ‘the’, **‘pain’**, **‘in’**, ***myt***, **‘foot’**, ‘is’, ‘subsiding’, ‘and’, ‘I’, ‘can’, ‘walk’, ‘a’, ‘lot’, ‘better’, ‘.’]
		Tags: [‘O’, ‘O’, ‘O’, ‘O’, ‘B-DIS’, ‘I-DIS’, ‘I-DIS’, ‘I-DIS’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’]
	Misspelling Errors	Words: [‘I’, ‘find’, ‘that’, ‘the’, **‘pain’**, **‘in’**, **‘my’**, ***fooot***, ‘is’, ‘subsiding’, ‘and’, ‘I’, ‘can’, ‘walk’, ‘a’, ‘lot’, ‘better’, ‘.’]
		Tags: [‘O’, ‘O’, ‘O’, ‘O’, ‘B-DIS’, ‘I-DIS’, ‘I-DIS’, ‘I-DIS’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’]
	Synonym Replacement	Words: [‘I’, ‘find’, ‘that’, ‘the’, **‘pain’**, **‘in’**, **‘my’**, **‘foot’**, ‘is’, ‘subsiding’, ‘and’, ‘I’, ‘can’, ‘walk’, ‘a’, ‘lot’, ***suffering***, ‘.’]
		Tags: [‘O’, ‘O’, ‘O’, ‘O’, ‘B-DIS’, ‘I-DIS’, ‘I-DIS’, ‘I-DIS’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’]
	WordNet SR	Words: [‘I’, ‘find’, ‘that’, ‘the’, **‘pain’**, **‘in’**, **‘my’**, **‘foot’**, ‘is’, ‘subsiding’, ‘and’, ‘I’, ‘can’, ‘walk’, ‘a’, ‘lot’, ***ache***, ‘.’]
		Tags: [‘O’, ‘O’, ‘O’, ‘O’, ‘B-DIS’, ‘I-DIS’, ‘I-DIS’, ‘I-DIS’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’]
	Hypernym Replacement	Words: [‘I’, ‘find’, ‘that’, ‘the’, **‘pain’**, **‘in’**, **‘my’**, **‘foot’**, ‘is’, ‘subsiding’, ‘and’, ‘I’, ‘can’, ***distress***, ‘a’, ‘lot’, ‘better’, ‘.’]
		Tags: [‘O’, ‘O’, ‘O’, ‘O’, ‘B-DIS’, ‘I-DIS’, ‘I-DIS’, ‘I-DIS’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’]

Continued on next page

Table 3.4 – continued from previous page

Augmentation Level	Augmentation Method	Augmented Text
	Hyponym Replacement	Words: [‘I’, ‘find’, ‘that’, ‘the’, **‘pain’**, **‘in’**, **‘my’**, **‘foot’**, ‘is’, ‘subsiding’, ‘and’, ‘I’, ‘can’, ‘walk’, ‘a’, ‘lot’, ***cramp***, ‘.’]
	Swap Word	Tags: [‘O’, ‘O’, ‘O’, ‘O’, ‘B-DIS’, ‘I-DIS’, ‘I-DIS’, ‘I-DIS’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’] Words: [‘I’, ‘find’, ‘that’, ‘the’, ***‘in’, ‘pain’***, ‘my’, ‘foot’, ‘is’, ‘subsiding’, ‘and’, ‘i’, ‘can’, ‘walk’, ‘a’, ‘lot’, ‘better’, ‘.’]
	CHV-Drug SR	Tags: [‘O’, ‘O’, ‘O’, ‘O’, ‘B-DIS’, ‘I-DIS’, “I-DIS”, ‘I-DIS’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’] Words: [‘I’, ‘find’, ‘that’, ‘the’, ***‘pain’, ‘sense’, ‘in’, ‘my’, ‘foot’*** i, ‘is’, ‘subsiding’, ‘and’, ‘i’, “can”, ‘walk’, ‘a’, ‘lot’, ‘better’, ‘.’]
		Tags: [‘O’, ‘O’, ‘O’, ‘O’, ‘B-DIS’, **‘I-DIS’**, ‘I-DIS’, ‘I-DIS’, ‘I-DIS’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’]
Combination	Keyboard Errors	Words: [‘I’, ***fund***, ‘that’, ‘the’, **‘pain’**, ***fund***, **‘im’**, **‘foot’**, ‘is’, ‘subsiding’, ‘and’, ‘I’, ‘can’, ‘walk’, ‘a’, ‘lot’, ‘better’, ‘.’]
		Tags: [‘O’, ‘O’, ‘O’, ‘O’, ‘B-DIS’, ‘I-DIS’, ‘I-DIS’, ‘I-DIS’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’]
	OCR Errors	Words: [‘I’, ‘find’, ‘that’, ‘the’, **‘pain’**, **‘in’**, **‘my’**, ***fo0t***, ‘is’, ‘subsiding’, ‘and’, ‘I’, ‘can’, ‘walk’, ‘a’, ‘lot’, ***‘b3tter’***, ‘.’]
		Tags: [‘O’, ‘O’, ‘O’, ‘O’, ‘B-DIS’, ‘I-DIS’, ‘I-DIS’, ‘I-DIS’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’]
	Misspelling Errors	Words: [‘I’, ‘find’, ‘that’, ‘the’, ***psin***, **‘in’**, **‘my’**, **‘foot’**, ‘is’, ‘subsiding’, ‘and’, ‘I’, ‘can’, ***wakl***, ‘a’, ‘lot’, ‘better’, ‘.’]
		Tags: [‘O’, ‘O’, ‘O’, ‘O’, ‘B-DIS’, ‘I-DIS’, ‘I-DIS’, ‘I-DIS’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’]
	WordNet SR	Words: [‘I’, ‘find’, ‘that’, ‘the’, ***‘hurting’***, **‘in’**, **‘my’**, **‘foot’**, ‘is’, ‘subsiding’, ‘and’, ‘I’, ‘can’, ‘walk’, ‘a’, ‘lot’, ***‘improve’***, ‘.’]
		Tags: [‘O’, ‘O’, ‘O’, ‘O’, ‘B-DIS’, ‘I-DIS’, ‘I-DIS’, ‘I-DIS’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’]
	Hypernym Replacement	Words: [‘I’, ‘find’, ‘that’, ‘the’, **‘pain’**, **‘in’**, **‘my’**, ***‘walk’***, ‘is’, ‘subsiding’, ‘and’, ‘I’, ‘can’, ***‘travel’***, ‘a’, ‘lot’, ‘better’, ‘.’]

Continued on next page

Table 3.4 – continued from previous page

Augmentation Level	Augmentation Method	Augmented Text
		Tags: [‘O’, ‘O’, ‘O’, ‘O’, ‘B-DIS’, ‘I-DIS’, ‘I-DIS’, ‘I-DIS’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’]
	Hyponym Replacement	Words: [‘I’, ‘find’, ‘that’, ‘the’, **‘pain’**, **‘in’**, **‘my’**, ***‘flatfoot’***, ‘is’, ‘subsiding’, ‘and’, ‘I’, ‘can’, ‘walk’, ‘a’, ‘lot’, ***perfect***, ‘.’]
	Swap Word	Tags: [‘O’, ‘O’, ‘O’, ‘O’, ‘B-DIS’, ‘I-DIS’, ‘I-DIS’, ‘I-DIS’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’] Words: [‘I’, ‘find’, ‘that’, ‘the’, ***‘in’, ‘pain’***, ‘my’, ‘foot’, ‘is’, ‘subsiding’, ‘and’, ‘i’, ***‘walk’, ‘can’***, ‘a’, ‘lot’, ‘better’, ‘.’]

Lastly, we conducted an additional experiment by merging the CADEC and MedRed datasets for training the NER model. This involved combining the training data from both CADEC and MedRed into a single corpus, while still keeping the original test data separated for evaluating the model's performance. The objective was to train the model with a larger training set. This experiment aims to assess if the combination of the data leads to increased model performance.

Experiment Setup for MCN Task While training the models to handle the MCN task, we augmented 100% of the original training data (adding it to the original training data), because of the data scarcity problem as we mentioned in Section 2.7. We used similar augmentation techniques as in the NER case to augment the MCN training datasets, except the mention replacement. We repeat the training process five times with different random seeds.

For comparison, we replicated the state-of-the-art baseline from [TMNM18] based on recurrent neural networks, specifically using GRU with HealthVec [MTT17] and TF-IDF as the representation of data.

3.2.1 Evaluation Metrics

The micro F1 score is employed for evaluating NER and MCN models that used augmented training data due to its balanced measurement of precision and recall, especially useful in imbalanced or multi-class scenarios. We assess statistical significance through a *paired t-test* with $p < 0.05$.

Computing F1 for the NER Task To evaluate models trained for Named Entity Recognition (NER), we use two evaluation methods as in [SBMHZ13]:

1. Strict Matching: Requires exact match in both span and type.
2. Partial Matching: Considers matches in span, regardless of type.

Precision and Recall for strict and partial span matching are computed using the Message Understanding Conference (MUC) scores system, created for evaluating Information Extraction systems [CS93], adapted for the SemEVAL tasks. The MUC scoring categories are a comparison of a system's performance with the ground truth. We give below an explanation of each of the MUC categories:

- Correct (COR): The output of the system and the ground truth are in agreement;
- Incorrect (INC): the output of a system does not match the ground truth;
- Partial (PAR): the system and the ground truth are similar but not identical;
- Missing (MIS): A ground truth is not captured by a system.

- Spurious (SPU): the system generates an output that isn't present in the ground truth;

To compute Precision and Recall we need to define the True Positives (TP), False Negatives (FN), and False Positives (FP) quantities. We define them as sums of MUC categories as described in the Evaluation of SemEval-2013 Task 9.1 [3]:

$$TP_{strict} = COR \tag{3.1}$$

$$TP_{partial} = COR + 0.5 * PAR \tag{3.2}$$

$$ACT = COR + INC + PAR + SPU \tag{3.3}$$

$$POS = COR + INC + PAR + MIS \tag{3.4}$$

$TP_{partial}$: True Positives in a partial sense. As per Equation 3.2, $TP_{partial}$ includes all Correct (COR) instances plus half of the Partial (PAR) instances. This approach gives some credit to partially correct outputs.

ACT: This represents the total number of actual responses from the system, as shown in Equation 3.3. It's the sum of Correct (COR), Incorrect (INC), Partial (PAR), and Spurious (SPU) instances. Essentially, it's a measure of all outputs generated by the system, whether they are correct, partially correct, incorrect, or spurious.

POS: This term denotes the total number of positive instances in the ground truth, as indicated in Equation 3.4. It includes Correct (COR), Incorrect (INC), Partial (PAR), and Missing (MIS) instances. It represents all instances that should have been recognized by the system, including those it failed to identify.

Using these definitions we can compute the Precision and Recall for the strict and partial evaluations:

$$\begin{aligned} Precision_{strict} &= \frac{COR}{ACT} \\ Recall_{strict} &= \frac{COR}{POS} \end{aligned} \tag{3.5}$$

$$\begin{aligned} Precision_{partial} &= \frac{COR + 0.5 * PAR}{ACT} \\ Recall_{partial} &= \frac{COR + 0.5 * PAR}{POS} \end{aligned} \tag{3.6}$$

Finally, the F1 scores for the partial and strict evaluations are defined as the harmonic mean for the respective precision and recall quantities.

[3] https://www.cs.york.ac.uk/semeval-2013/task9/

Computing F1 for the MCN Task We evaluate the MCN model performance by using standard Precision, Recall and F1-Measure. Since MCN is a multi classification task, we only report the F1 scores. F1-scores are calculated by counting overall TP, FN, and FP from all of the classes.

3.3 Results and Discussion

Data Augmentation for NER task The experiments aim to evaluate the effect of data augmentation on the output of the NER model in our MEL workflow (Figure 1.1). Table 3.5 and Table 3.6 shows the evaluation results for the various NER models trained on the different datasets we created with the augmentation techniques described in Section 3.1.2.

Table 3.5: NER Performance for popularized medical entity recognition on the MedRed (Strict and Partial)

Model	Strict			Partial		
	Ori + 20%	Ori + 50%	Ori + 100%	Ori + 20%	Ori + 50%	Ori + 100%
Ori (baseline)	**69.6 ± 1.1**			**74.2 ± 0.9**		
+Char	69.5 ± 1.1	69.5 ± 1.2	68.9 ± 0.7	74.1 ± 1.0	73.5 ± 1.0	72.6 ± 0.8
+Word	*69.5 ± 0.6*	68.5 ± 0.2	68.8 ± 0.7	*74.2 ± 0.6*	72.9 ± 0.3	73.5 ± 0.7
+WordChar	68.8 ± 0.6	68.3 ± 0.6	67.8 ± 0.7	73.8 ± 0.6	72.8 ± 0.8	72.1 ± 0.8
+Paraphrase	67.3 ± 1.6	65.0 ± 1.5	62.3 ± 1.7	72.3 ± 1.4	69.9 ± 1.6	66.9 ± 1.7
+MR_Sem	68.5 ± 0.4	68.1 ± 0.7	69.0 ± 0.3	73.1 ± 0.7	73.1 ± 0.8	73.8 ± 0.2
+Combination	67.3 ± 1.2	67.3 ± 0.8	68.2 ± 0.8	73.1 ± 0.9	72.6 ± 0.7	73.0± 0.6

Table 3.6: NER Performance for popularized medical entity recognition on the CADEC

Model	Strict			Partial		
	Ori + 20%	Ori + 50%	Ori + 100%	Ori + 20%	Ori + 50%	Ori + 100%
Ori (baseline)	**81.2±0.1**			**85.9±0.1**		
+Char	78.7±0.8	78.5 ± 0.5	79.0 ± 0.7	83.4 ± 0.9	83.4 ± 0.8	83.5 ± 0.9
+Word	*79.6 ± 0.4*	78.4 ± 0.5	78.7 ± 0.9	*84.0 ± 0.5*	83.2 ± 0.6	83.3 ± 0.8
+WordChar	78.3 ± 0.5	77.6 ± 0.3	77.5 ± 0.3	83.4 ± 0.3	82.6 ± 0.4	82.6 ± 0.4
+Paraphrase	78.6 ± 0.6	78.1 ± 0.5	77.8 ± 0.6	83.5 ± 0.6	83.0 ± 0.4	82.9 ± 0.6
+MR_Sem	78.9 ± 0.2	79.4 ± 0.3	78.3 ± 1.1	83.9 ± 0.1	84.1 ± 0.2	83.2 ± 1.3
+Combination	78.9 ± 0.1	78.7 ± 0.5	78.9 ± 0.3	83.8 ± 0.2	83.7 ± 0.4	83.9 ± 0.3

From the experimented results, we conclude that, compared to the baseline, the models for all augmentation approaches under-performed on both the CADEC and MedRed datasets. Furthermore, the partial evaluation performs better than the strict evaluation. This is to be expected as the partial evaluation is more relaxed than the strict evaluation,

allowing for more TP cases. Among the proposed techniques, the model that performs best is the one that uses word augmentation to increase the amount of training data by 20%. According to our experiments, adding more data does not guarantee that the model's performance will improve. As discussed in [DA20], data augmentation may alter the ground truth label or create erroneous occurrences. This adverse effect may be particularly pronounced in large training sets. This negative impact was particularly evident when we augmented 100% of the original training data, leading to a decline in model performance across all augmentation techniques. We further do a failure analysis to assess if the set of false positive terms identified from all augmentation techniques could be useful as inputs to the MCN model.

Sensitivity to Data Augmentation: Furthermore, we assessed the sensitivity of various models to data augmentation. This evaluation involved training the models using different levels of data augmentation (see Table 3.4). We employed three distinct augmentation levels: (1) context-level augmentation, which involved modifying the surrounding contextual information; (2) entity-level augmentation, which focused on altering the entity mentions and served as our primary experimental focus; and (3) combined augmentation, which applied both context-level and entity-level simultaneously. This approach allowed us to evaluate whether a model is more sensitive to certain types of augmentation than others.

The proposed data augmentation techniques might unintentionally alter important features or information that could potentially impact the model performance. For instance, in entity-level augmentation, the methods applied to entity mentions may affect the model's ability to recognize important entities (e.g., medical entities). The augmentation at the context level may influence entity recognition by potentially changing the semantic meaning of the original sentence. We compared our proposed NER (refer Section 3.1.3) as our baseline with BERT-biLSTM-CRF [KPT+20]. In this experiment, we evaluated using 20% augmented training data to increase the amount of training data. As seen in the previous experiment (Table 3.5 and Table 3.6), the results with 20% augmented data showed slightly better performance compared to augmenting 50% or 100% of the training data.

Tables 3.7 and 3.8 present the results for the MedRed and CADEC datasets, showing F1-scores for strict and partial evaluation of each model under different augmentation levels and techniques. Our proposed model outperformed the state-of-the-art model on the original data for both datasets. For MedRed, it achieved an F1-score of 69.6% for strict evaluation and 74.2% for partial evaluation. For CADEC, it achieved 81.2% and 83.3%, respectively.

However, our model appears more sensitive to noise introduced by both augmentation techniques and levels, particularly for context-level augmentation on both datasets, especially when using the character-based augmentation technique, where performance decreases observed for both strict and partial F1-scores. For the MedRed dataset, the F1-score decreases to 61.3% for strict evaluation, and 65.3% for partial evaluation, while for

CADEC the F1-score for strict evaluation drops to 74.1% and 78.6% for partial evaluation. Additionally, our model also shows sensitivity to noise in the combine-level, where we augment the context and also the mention from the original text. The model performance decreases for both character and word-based augmentation techniques. Furthermore, the proposed model has minimal impact on the entity-level.

In contrast, the BERT-biLSTM-CRF model started with lower performance on the original data but showed slight improvements across different augmentation levels, especially in context-level augmentation for the CADEC dataset, for both strict and partial F1-scores. For strict evaluation, the F1-score improved from 77.5% to 78.3% using character-based augmentation and to 78.0% using word-based augmentation. For partial evaluation, the F1-score increased from 83.3% to 83.9% using character-based augmentation and to 83.5% using word-based augmentation.The model also showed slight improvements at other augmentation levels. Entity-level augmentation with word-based techniques yielded a slight improvement in performance. Similarly, combination-level augmentation, incorporating both character and word-based approaches. In the MedRed dataset, the BERT-biLSTM-CRF model demonstrated a slight improvement for entity-level augmentation using character-based techniques, resulting a strict F1-score of 68.4% and a partial F1-score of 74.4%.

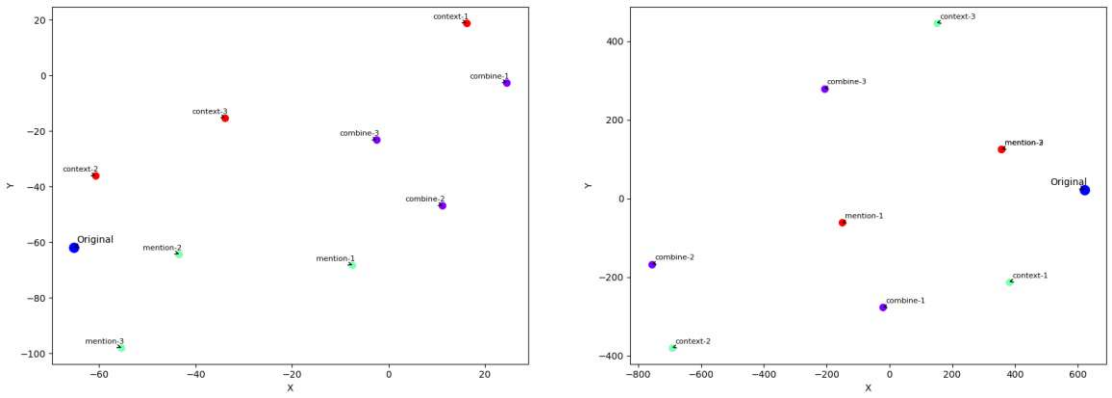
Table 3.7: NER performance comparison between the proposed model and BERT-biLSTM-CRF [KPT+20] on the MedRed dataset

Level	Method	**Strict**		**Partial**	
		Glove+RoBERTa -biLSTM-CRF	BERT -biLSTM-CRF	Glove+RoBERTa -biLSTM-CRF	BERT -biLSTM-CRF
Context	Ori	**69.6**	68.0	**74.2**	73.5
	+Char	61.3	67.5	65.3	72.7
	+Word	60.4	66.5	65.7	72.2
Entity	+Char	69.5	68.4	74.1	74.4
	+Word	69.5	67.7	74.2	74.4
Combine	+Char	60.6	67.5	65.4	73.1
	+Word	60.7	68.3	65.5	73.3

Furthermore, to understand these results, we conducted a t-SNE visualization analysis of the augmented texts in comparison to the original texts. We took two examples from CADEC dataset (see Appendix 7.3.3) to illustrate the effect of different augmentation techniques in increasing the variation of the sentence while preserving semantics of the original text. The t-SNE plots in Figures 3.3b and 3.4b show that the augmented text using word-based augmentation techniques has relatively tight clustering compared to the original texts at all augmentation levels. This were noticed for Text 1 (Figure 3.3b): *Extremely severe pain in right shoulder as if from extreme workout or injury (none which apply). Stopped taking lipitor seven days ago and still experiencing pain in shoulder and*

Table 3.8: NER performance comparison between the proposed model and BERT-biLSTM-CRF [KPT+20] on the CADEC dataset

Level	Method	**Strict**		**Partial**	
		Glove+RoBERTa -biLSTM-CRF	BERT -biLSTM-CRF	Glove+RoBERTa -biLSTM-CRF	BERT -biLSTM-CRF
Context	Ori	**81.2**	77.5	**85.9**	83.3
	+Char	74.1	78.3	79.6	83.9
	+Word	75.2	78.0	80.5	83.5
Entity	+Char	78.7	77.1	83.4	83.2
	+Word	79.6	78.0	84.0	83.6
Combine	+Char	75.3	77.8	80.8	83.5
	+Word	75.2	78.7	80.8	84.0



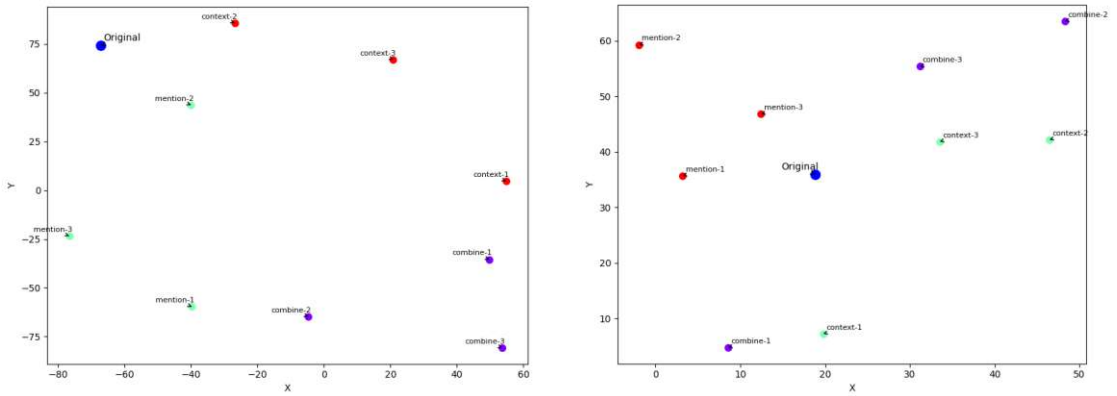
(a) Original text vs augmented text using character-based augmentation for Text 1

(b) Original text vs augmented text using word-based augmentation for Text 1

Figure 3.3: t-SNE visualization: original vs augmented Text 1

tingling and numbness down right arm radiating into fingers. No more medications. will attempt holistic approach. vitamins C and/or niacin. The same pattern is seen for Text 2 (Figure 3.4b): *Pain in hip, lower back, knees & elbow. Stiffness in lower back and legs when getting up in the morning . Excercise intolerance and general muscle weakness. I ask my doctor about these pains months ago and he said lipitor would not cause my problems. However, after reading the common side effects with others on this site, I will stop this medication immediately, it not worth feeling like an old man at age 46 !.*

Contrasted to word-based augmentation techniques, the text augmentation using character-based augmentation techniques show a wider spread of augmented texts, as can be seen in Figures 3.3a and 3.4a for both original texts. The tighter clustering of augmented texts with word-base augmentation techniques indicates the technique is able to increase the



(a) Original text vs augmented text using character-based augmentation for Text 2

(b) Original text vs augmented text using word-based augmentation for Text 2

Figure 3.4: t-SNE visualization: original vs augmented Text 2

variation of the text, while maintaining the original semantics of the texts. However, the character-based augmentation technique indicates higher variations of sentences than the word-based augmentation technique, while maintaining reasonable semantic similarity with the original texts.

For the augmentation level, context-level augmentations tended to cluster to the original texts, which may explain the slight improvements of the model performance for the state-of-the-art model for this augmentation level, specifically for CADEC dataset. Entity-level augmentation shows various degrees of similarity, but remained close to the original points. This indicates that our model is robust to these augmentation for both MedRed and CADEC datasets. The combined augmentations demonstrated a wider spread from the original points, especially in character-based augmentation techniques. This indicates that the combination of augmentation level, both context and mentions introduces more noise and may change the semantics from the original texts. This may suggest a decrease in model performance for both our proposed model and state-of-the-art model.

These findings highlight the complex nature of augmentation effects, which can vary model performance under different augmentation techniques for different datasets, especially for NER tasks. Our proposed model achieved higher performance when trained on the original datasets (both MedRed and CADEC), but it may be sensitive to the noise introduced by certain augmentation techniques and levels. Conversely, the state-of-the-art model appears more robust to the augmentation, as evidence by the slight improvement for the model performance under some augmentation techniques and levels. To achieve better generalization and a more comprehensive understanding of augmentation effects in NER tasks, further experiments need to be conducted. These could include exploring other proposed augmentation techniques and testing their impact on different models.

Data Augmentation for MCN tasks We present experimental results of the MCN model's performance on data augmented by different techniques. Our proposed model, using GRU as the neural architecture with stacked GloVe and RoBERTa embeddings, is compared to the state-of-the-art model from Tutubalina et al. [TMNM18], which used HealthVec and TF-IDF as the representation of the data.

Table 3.9 summarizes each trained model's performance on the baseline and augmented data in terms of micro F1-score, for the MCN task. Bold values in the table show the best model performance. Each data augmentation techniques was applied once to the entire training data. In comparison to NER tasks, augmentation techniques applied to MCN models outperformed the models trained on original data for both datasets.

Our proposed model, combining GRU with stacked GloVe and RoBERTa embedding, consistently outperformed the state-of-the-art model across all augmentation techniques and the original dataset for both CADEC and PysTAR datasets. For instance, in the original dataset (None), our model achieved F1-score 72.5 ± 4.9 for CADEC and 79.6 ± 0.2 for PsyTAR, compared to 63.0 ± 3.7 and 56.0 ± 6.7 respectively for the state-of-the-art model. We argue that the stacked embedding contains rich information by combining context-independent (GloVe) and context-dependent (RoBERTa) features, potentially capturing a broad range of semantic and syntactic information. Although HealthVec is a domain-specific embedding trained on health-related user comments and TF-IDF provides statistical features, our experimental results suggest that the combination of GloVe and RoBERTa is more beneficial for this task.

Furthermore, the model trained with augmented data using WordChar, Paraphrase, and Combination techniques consistently outperformed those using Word-based and Char-based techniques. This consistency was observed in both our proposed model and the state-of-the-art model.

The state-of-the-art model showed improved performance when trained with augmented data compared to the original training data. For CADEC, the Combination technique yielded the best results with an F1-score of 67.3 ± 2.7 ($t = 5.68$, p-value = 0.002). Similarly, for the PsyTAR dataset, the Combination technique outperformed other models, with an F1-score of 76.2 ± 0.4 ($t = 7.03$, p-value = 0.001). The WordChar technique demonstrated the second-best performance for PsyTAR, with an F1-score of 74.6 ± 2.5 ($t = 9.45$, p-value = 0.0003). According to our observation, the word replacement introduces new vocabulary to the corpus. This finding is aligned with Wei and Zou's work [WZ19].

Our proposed model demonstrated different results. On the PsyTAR dataset, the Paraphrasing technique outperformed other augmentation methods. The model trained with data augmented by this technique showed a significant improvement over the baseline model, 81.4 ± 0.5 of F1-score ($t = 11.08$, p-value = 0.00018858). This was followed by the Combination model, where the model was trained on the augmented data used by all combinations of the augmentation techniques, with 81.2 ± 0.4 of F1-score ($t = 8.65$, p-value = 0.00049).

Table 3.9: MCN performance on CADEC and PsyTAR datasets.

		CADEC	**PsyTAR**
GRU+HealthVec+TFIDF [TMNM18]	None	63.0 ± 3.7	56.0 ± 6.7
	Char	62.8 ± 4.6	66.6 ± 6.5
	Word	63.9 ± 4.4	70.8 ± 0.2
	Paraphrase	65.9 ± 1.9	70.5 ± 0.4
	WordChar	65.2 ± 3.1	**74.6 ± 2.5**
	Combination	**67.3 ± 2.7**	**76.2 ± 0.4**
GRU+GloVe+RoBERTa (Our)	None	72.5 ± 4.9	79.6 ± 0.2
	Char	74.5 ± 5.4	79.0 ± 0.4
	Word	76.8 ± 4.6	80.2 ± 0.5
	Paraphrase	77.1 ± 4.2	**81.4 ± 0.5**
	WordChar	76.9 ± 5.5	80.3 ± 0.1
	Combination	**78.8 ± 3.3**	**81.2 ± 0.4**

According to our findings, the paraphrasing engine preserved the semantics of popularized medical phrases while producing new semantically similar phrases to the original terms. Figure 3.5 shows some examples of cosine similarity between a reflection of input and augmented phrases. For instance, the phrase *hard to get out of bed in the morning* is paraphrased into similar phrases like *getting from bed in the morning is hard* and *harder to get out of bed in the morning*. This approach ensures that the original meaning of the phrase is preserved, avoiding the introduction of noise into the data.

Additionally, in the CADEC experiments, combining all augmentation techniques (the Combination model in Table 3.9) outperforms other techniques. The Combination model, which incorporates all augmentation methods, showed a significant improvement over the baseline model (t = 6.90 and a p-value = 0.001). This improvement can be attributed to the broader range of popularized medical phrases introduced in the training data.

Meanwhile, character augmentation performed poorly in comparison to other techniques. On PsyTAR data, character-based augmentation methods cannot be used to improve our proposed model trained on the original dataset, showing a slight decrease in performance (79.0 ± 0.4 vs baseline 79.6 ± 0.2). This poor result is also evident in the state-of-the-art model for CADEC, where character-based augmentation yielded the lowest score (62.8 ± 4.6). This under-performance of character-based augmentation techniques might introduce noise, potentially affecting the model learning. However, further investigations are needed as the character-based augmentation technique performed better for PsyTAR dataset trained on the state-of-the-art model, improving from a baseline of 56.0 ± 6.7 to 66.6 ± 6.5.

Failure Analysis in NER Data Augmentation This section discusses and analyzes the failure analysis on the NER augmentation model, with a focus on our proposed model

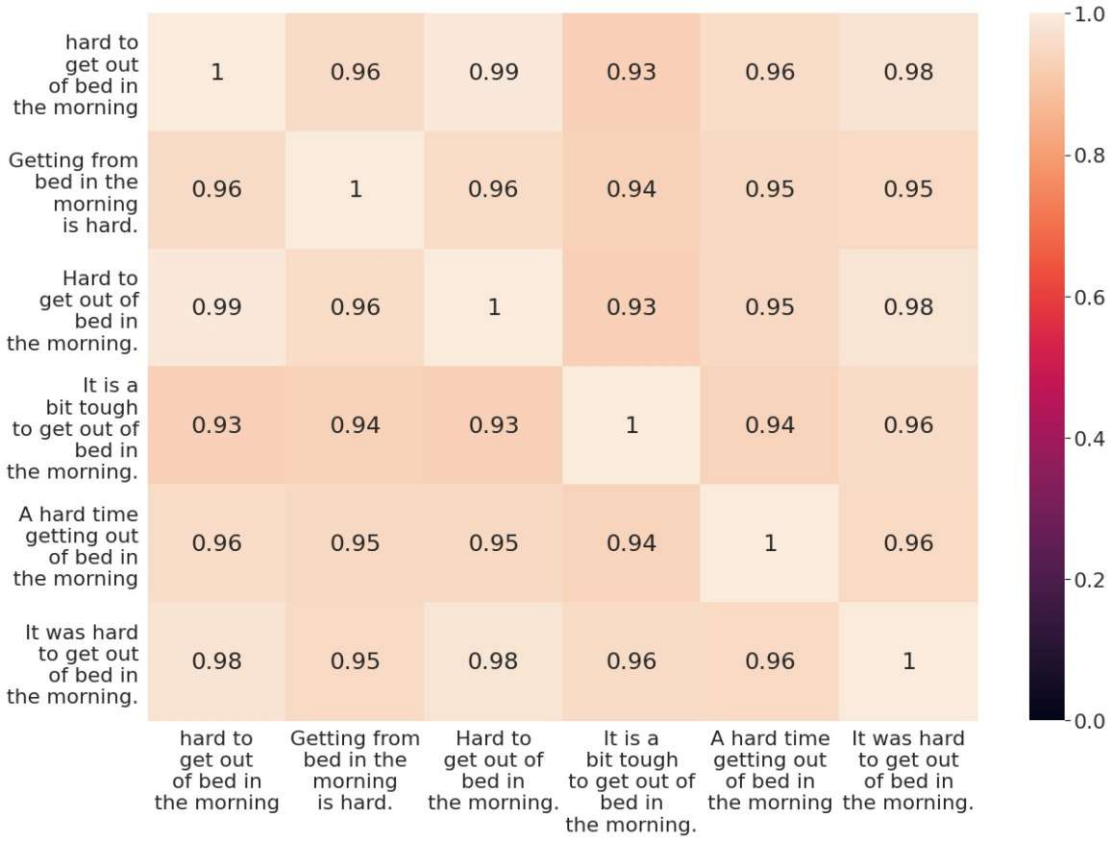


Figure 3.5: Pairwise cosine similarity between original (first column & row) and augmented phrases from the paraphrase method for an example phrase

architecture. The goal is to assess whether the false positive terms generated by the NER augmentation models could be used as an input to the MCN model in the informal medical entity linking model pipeline. This means the false positives can be indicated as popularized medical phrases in an informal medical entity pipeline.

Our analysis is driven by the hypothesis that these false positive terms, despite their initial misclassification, may nonetheless represent valid popularized medical phrases. This is particularly pertinent for predicted spans that were not present in the ground truth data. We conducted a manual assessment done by two medical experts focused on inspecting these false positive terms. The assessment aimed to identify the correctness of the entity types (*Disease* or *Drug*) and if it is correctly classified to the specialized medical terms extracted from the MCN model trained on CADEC.

The analysis specifically focused on the CADEC test dataset. We used the best-performing NER models (based on partial match evaluations) from each of augmentation techniques. We collected all the FP terms from each technique and we created a unified distinct FP terms list. We identified terms that were false positives. A total of 452 false positive

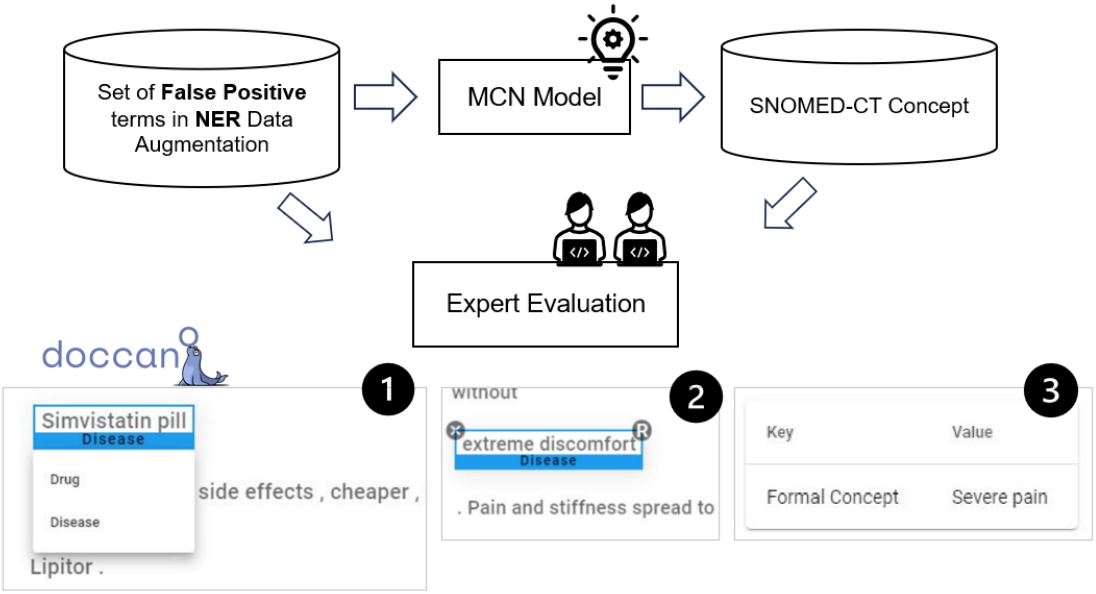


Figure 3.6: Analysis of False Positive in Named Entity Recognition. The expert identifies two main tasks (1) instances where popularized medical phrases are correctly linked to their specialized medical terms but are labeled with the wrong entity type, (2) cases where spans are not popularized medical phrases, leading to the removal of their entity type labels. Both of these tasks are based on the knowledge of (3) the identified specialized medical terms.

terms were identified. These terms were then processed through the highest performing MCN model trained on the CADEC dataset, which classified these popularized phrases into specialized medical terms as per SNOMED-CT. This dataset was then used as the basis for our evaluation. Figure 3.6 illustrates the failure analysis process.

We set two tasks for the evaluation: (1) evaluate whether these FP terms were assigned to correct entity types, either *Disease* or *Drug*, and (2) the expert also conducted a further analysis to verify the correctness of the specialized term predicted for each false positive (FP) term. We used `Doccano`, an annotation tool with which experts provided their feedback for this failure analysis[4].In the first task, as illustrated in Figure 3.7, the interface shows sentences containing the false positive (FP) term, allowing the experts to assess if the span is an popularized phrase and if its label is correct. If both the popularized phrase and its label are found to be correct, the term and its label are retained as they are.

In contrast, if the FP term span is an popularized phrase but its label is incorrect, the expert modifies the label accordingly. Finally, if either the span or its label is incorrect, expert will remove the term's label, and provide a note explaining the reason

[4]https://mel-fp.herokuapp.com/

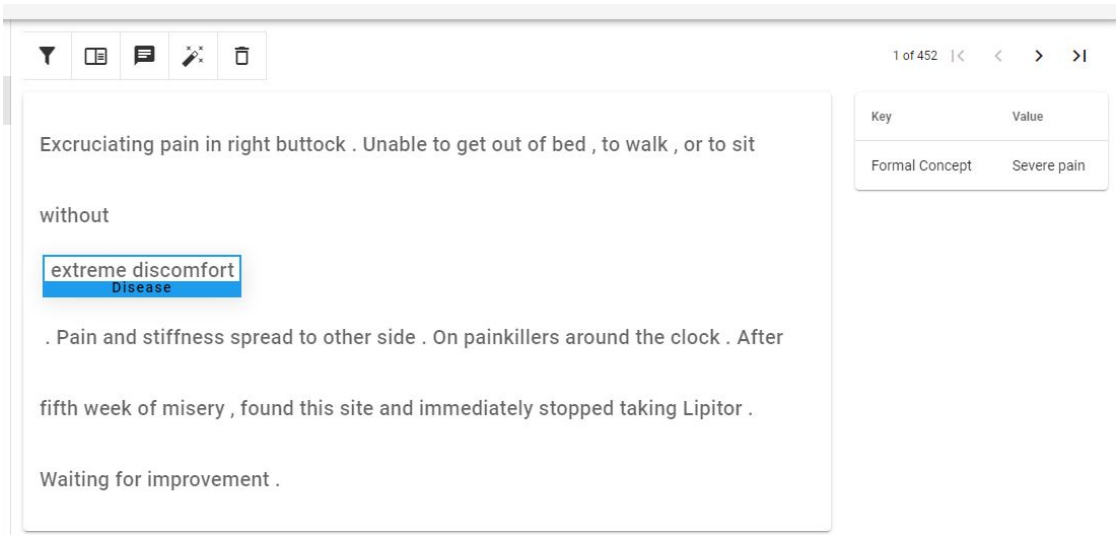


Figure 3.7: A User Interface for Evaluating Failure Analysis in Doccano Annotation Tools

for its removal. In the second task, experts are tasked with evaluating whether the FP terms identified as popularized phrases have been correctly classified into their corresponding specialized terms. However, due to time constraints, only one expert was able to participate in this task. We acknowledge that this limited participation of experts is a constraint in our evaluation analysis for the second task.

Based on our analysis, among these 452 terms, both experts agree that 407 are medical terms with correct entity type, and 10 are not medical terms. The Cohens' Kappa coefficient for this task is 0.3 (fair agreement). Based on the expert assessment, the terms that are not disease/symptom are actually related to medical procedures, body parts, or medical tests, whereas the terms that are not drugs are nutritional or dietary supplements. Furthermore, in order to validate the 407 terms identified as medical terms, we used the MCN model to cross-check them with the specialized medical term. Among the 407 terms, 343 of the specialized medical concepts are correctly identified. For instance, medical terms *muscle ached* and *unable to lose weight* can be normalized to SNOMED-CT medical concept of *myalgia* and *weight gain*. This observation demonstrates that the false positive for popularized medical phrases could lead to correct specialized medical terms. Based on this assessment, we argue that 343 of 452 (75%) terms detected by our model using augmented data and marked as false positives in the evaluation, can in fact be considered valid medical entities. This means that the NER model could actually be beneficial to the whole pipeline of informal medical entity linking model.

NER Model Trained on Combined CADEC and MedRed Datasets In the previous experiment, we attempted to enhance the baseline model's performance on the CADEC [KMJKW15] and MedRed [SMLQB20] datasets by introducing additional

data through augmentation techniques. Unfortunately, the results showed that these augmentation efforts did not lead to an improvement in model performance of our proposed model architecture.

As a result, in this new experiment, we made the decision to merge the CADEC and MedRed datasets for training our model. This is different from the augmentation techniques that we propose before. This combination aims to increase training data to train the NER model, where both datasets share a commonness used to extract the medical entities of disease or drugs from social media. This involved combining the training data from both CADEC and MedRed into a single corpus, while still keeping the original test data separate for evaluating the model's performance. The objective was to train the model with a larger training dataset.

In this experiment, we examined whether increasing the size of the training dataset by merging existing data would improve the performance of the NER model. We trained the model with the model architecture explained in Section 3.1. The results of this experiment can be found in Table 3.10.

Table 3.10: NER performance trained on combination of CADEC and MEDRED train data. The baseline is a model trained with the original training data, as shown in Table 3.6 and Table 3.5.

Training Data Description	Test Data	F1-Score	
		Strict	Partial
CADEC (Baseline)	CADEC	81.2	85.9
MedRed (Baseline)	MedRed	69.6	74.2
$MedRed \cup CADEC$	CADEC	75.72	82.38
	MedRed	69.75	74.49

The results shows that in the strict evaluation, we achieved an F1-score of 75.72% for CADEC dataset, while in the partial evaluation, the score was notably higher at 82.38%. For MedRed, the strict evaluation achieved a score of 69.75%, whereas the partial evaluation resulted in a score of 74.49%. We assume this difference in performance due to the fact that CADEC has a larger number of entities for both entity types compared to MedRed, which likely allowed the model to learn more effectively from the CADEC data. Nonetheless, the NER model trained on this merged dataset did not perform as well as when trained on the original training sets, particularly with the CADEC dataset (see Table 3.6). However, for the MedRed dataset, the performance of the NER model trained on the combined dataset was at least comparable to its performance when trained on MedRed's original training set (see Table 3.5).

We examined several features to measure the characteristics of CADEC and MedRed. The dataset features presents in Table 3.11. The comparison between CADEC and MedRed datasets shows differences in their features. MedRed contains more sentences and has longer sentences across all splits than CADEC. The average sentence lengths of MedRed ranging from 15.28 to 15.94 words, compared to CADEC's 13.85 to 14.55 words. In contrast to MedRed, CADEC has a significantly higher number of entities (7,906 vs. 3,768) and consequently a greater entity density (0.10-0.12 vs. 0.04).

Table 3.11: Comparing the features of CADEC and MedRed

Dataset		**# Sents**	**Avg. Sent. Len.**	**# Entities**	**Entity Density**
CADEC	Train	4,716	13.85	4,323	0.11
	Dev	1,190	14.55	1,136	0.10
	Test	2,669	13.91	2,447	0.12
MedRed	Train	4,514	15.94	1,825	0.04
	Dev	2,353	15.28	939	0.04
	Test	2,325	15.45	1,004	0.04

This higher entity density indicates that CADEC has more entities per sentence, which could indeed be beneficial to the model during the learning process. The lower entity density in MedRed could potentially explain its apparently negative effect on model performance. According to [LWJ+23], higher density could increase the task difficulty of the NER model, which brings brings a challenge to the performance of the NER model. Our findings diverge from this expectation. In our case, the higher entity density proved beneficial for NER model learning, particularly when compared to the sparse entity distribution found in MedRed. This observation underscores the complex relationship between dataset characteristics and model performance in NER tasks. This suggests that further investigation is necessary to fully understand the impact of dataset features on model performance. To this end, Fu et al. [FLN20] have proposed an evaluation framework for interpretable assessment of NER tasks, which could provide valuable insights in future analyses.

Based on our earlier findings, it appears that the partial evaluation consistently outperforms the strict evaluation, which is expected because the partial evaluation is more relaxed than the strict evaluation, where the partial matching spans are allowed. Our observations revealed that the NER model occasionally predicts spans that are longer than the ground truth. For instance, in the CADEC, there is a sentence: "Excruciating pain in right buttock. Unable to get out of bed...". Here, the ground truth entity is *Excruciating pain*, while the model predicted *Excruciating pain in right buttock*. On the other hand, we also observed instances where the model predicted only a partial span of the ground truth. For example in MedRed, a sentence "Right now , the only things the doctors have to go on is a positive Rocky Mountain Spotted Fever test...", the ground truth is *Rocky Mountain Spotted Fever*, but the model predicted *Spotted Fever*.

We recognize that it can be beneficial for the model to capture longer spans, as this allows

it to capture more information. For example, the model can detect the specific body part where the user experiences intense discomfort. However, when the model predicts only a partial portion of the ground truth, there is a risk of losing important information. In the example provided, we missed the key term *Rocky Mountain*, which refers to the type of bacteria causing the fever. Future analysis should investigate whether longer spans introduce unnecessary noise and if shorter spans cause topic shifts. We leave this as future works.

3.4 Summary

This chapter describes the proposed data augmentation techniques to address the limitation of the number of support popularized medical for the specialized medical terms from the publicly available datasets in Medical Entity Linking, which consists of NER and MCN tasks. We proposed several data augmentation techniques, such as character-based, word-based, paraphrase-based, and semantic replacement augmentation techniques. We evaluated the impact of data augmentation on NER and MCN tasks by training the model and compared their performance against the model trained on the original training data.

The experimental results show that for NER tasks, specifically for our proposed model training architecture, the augmentation approach could not outperform the baseline model on CADEC and MedRed datasets. The word-based augmentation technique outperforms compared to the other techniques. Based on the failure analysis, we discovered that 343 of the 452 false-positive terms caused by augmentation could be classified as a medical entity and correctly normalized to the SNOMED-CT medical concept.

We further explored augmenting the context around entity mentions and both context and entity mentions to increase sentence variation. We compared the impact of different levels of augmentation on our proposed model and a state-of-the-art model. Our model showed better performance on the original data, but was sensitive to the noise introduced by certain augmentation techniques. The state-of-the-art model showed slight improvements across different levels of augmentation, but had lower overall performance compared to our model when trained on the original data. Further experiments are needed to understand the impact of augmentation levels on the NER task.

We also attempted to train the NER model on the combined CADEC and MedRed data to evaluate whether this could lead to an increase in model performance. However, the evaluation results showed that the NER model did not show improved performance when tested on the CADEC dataset compared to the model trained on the original CADEC dataset. On the other hand, its performance was comparable to the model trained on the original MedRed dataset.

Meanwhile, the augmentation techniques in the MCN task significantly improve our proposed model performance on both the CADEC and PsyTAR datasets compared to the state-of-the-art model. Models trained on data augmented by paraphrasing or a

combination of augmentation techniques outperformed the baseline model. However, character-based augmentation methods performed poorly compared to other techniques in the MCN task.

Based on our findings, the data augmentation method is a valuable approach to enhance Medical Entity Linking models, specifically MCN tasks, when the number of support phrases of medical concepts in training data is limited. However, improving NER task performance through data augmentation remains challenging and requires further investigation. The sensitivity of models to certain augmentation techniques and levels indicates the importance of carefully selecting the augmentation approach based on the characteristics of the dataset and model architecture. Future work will focus on the challenges of applying data augmentation to NER tasks.

CHAPTER 4

Leveraging Wikipedia Knowledge for Distant Supervision

In Chapter 3, we mentioned challenges in MCN tasks. One issue is that there is a language mismatch problem between the laypeople and medical professionals when referring to medical terminology. Laypeople usually use lay or slang terms (e.g. *can't sleep*), rather than specialized medical terms (e.g *insomnia*), when describing their symptoms or their medical experiences, such as medical treatments in social media. This lay medical expression can vary widely among laypeople. Another problem is the data scarcity problems, as the current research approaches the MCN as a supervised text classification task [LC15, LC16, TMNM18, MT19, PPPV20]. Examples of MCN task outputs are shown in Table 4.1.

One approach to addressing the data scarcity is to automatically generate labelled data with distant supervision methods using existing knowledge bases or dictionaries. Distant supervision in the MCN task is one of the approaches to overcome the low concept coverage of SNOMED-CT [PAP+20]. However, the current approach [PAP+20] is limited when the language gap between colloquial and specialized medical terms is wide. For example, the popularized phrase *need to sleep constantly* and *Somnolence* should be synonymous in medical terminology. However, the proposed distant supervision approach [PAP+20] does not map between the two phrases due to their low linguistic similarity.

Wikipedia has a large number of articles related to the medical domain. According to Shafee et al. [SMK+17], the English Wikipedia contains approximately 30,000 medical articles, and Ngo et al. [NTK+19] estimate that around 80% of the SNOMED-CT concepts are covered by Wikipedia articles. Wikipedia's Manual of Style for medical-related articles[1] recommends that authors write in plain English and as simply as possible. For instance, when introducing new technical terms, authors must provide an explanation

[1] https://bit.ly/Wikipedia_Manual_of_Style_Medicine-related_articles

Table 4.1: Example mappings between popularized medical phrases and medical terminology in SNOMED-CT

Popularized Medical Phrase	Normalized Medical Concept
Muscle pain	Myalgia (SNOMED-CT ID: 68962001)
Mellows me out	Feeling content (SNOMED-CT ID: 271599002)
Extreme pain	Severe pain (SNOMED-CT ID: 76948002)

in plain English, followed by the technical terms in parentheses. It is advisable to use hyperlinks to direct readers to other pages for further information. Additionally, *redirect* pages are created to aid in the search process by providing alternate names. For example, *heart attack* redirects to *Myocardial Infarction*. Considering all these, we hypothesise that medical related Wikipedia articles incorporate colloquial medical terminology.

In this chapter, we explore the suitability of the medical related Wikipedia articles as a source of distant supervision dataset for the automatic collection of labelled data, to supplement the existing MCN datasets (PsyTAR [ZFP+19], CADEC [KMJKW15], and COMETA [BLSC20]) with additional popularized phrases per concept and expand their concept coverage.

The contribution of this chapter is as follows:

- We explore the potential of using medical Wikipedia articles for distant supervision. We extract popularized medical phrases from various Wikipedia elements like abstracts, redirect pages, and wikilinks, and then link them to SNOMED-CT concepts.
- We evaluate the effectiveness of the distant supervision data for MCN tasks.

The remaining content of this chapter is organized as follows: Section 4.1 details our proposed distant supervision method. Section 4.2 describes the experimental setup for evaluating the impact of distant supervision on MCN tasks. We discuss the results in Section 4.3 and conclude the chapter in Section 4.4.

4.1 Distant Supervision Approach

This section describes our approach to the Medical Concept Normalization (MCN) task. Figure 4.1 shows the pipeline of our approach. Initially, we extracted medical articles from Wikipedia using Wikidata, leveraging three biomedical external identifier properties: SNOMED-CT, Unified Medical Language System (UMLS), and International Classification Disease, Tenth Revision (ICD-10). These biomedical external identifier properties not only helped us identify medical articles on Wikipedia but also served as sources for mapping selected articles to SNOMED-CT codes. The goal of Medical Concept Normalization is to normalize popularized medical phrases into specialized

medical terminology in the Knowledge Base (KB), specifically SNOMED-CT. Once we extracted medical articles from Wikipedia and mapped them to SNOMED-CT, the following step was identifying popularized medical phrases within Wikipedia structures, such as the first sentence of article summaries, Wikipedia redirect pages, and Wikilinks. A detailed explanation of this process is provided in the following sections.

4.1.1 Data Description

Our approach uses two main data sources, Wikidata and Wikipedia, where we use the January 2020 dump. This particular dump was used during our participation in the TREC Wikification 2020 [NEEP+20][2]. The Wikipedia dump contains several features that we will be using to produce our distant supervision data[3]:

(1) The **article summary** is the first section of a Wikipedia article. We specifically consider the first sentence in this summary. A Wikipedia article's summary section contains several phrases that explain the article's concept. The majority of a concept's basic explanation and central idea are typically contained in the first sentence. The second sentence and its remainder are used to explain details that may be irrelevant to our MCN task. As a result, we limit our extraction of medical terms to the first sentence.

(2) The Wikipedia ***redirect*** **pages** are alternative names, spellings, abbreviations, and common misspellings. Page redirects are frequently used to obtain the synonym for the titles of Wikipedia articles.

(3) The **wikilinks** are internal links between Wikipedia articles. Similar to redirects, they may point to synonyms and common phrases that refer to a concept.

In the typical MCN task, every popularized medical phrase will be classified to one of the SNOMED-CT codes. Wikipedia articles do not directly store SNOMED-CT codes, therefore, we incorporate information from Wikidata[4]. Wikidata is a free, collaborative, multilingual, collecting structured data as a central storage repository that can be used by Wikimedia projects, including Wikipedia [Wik].

Wikidata consists of items, where each item includes a label, description, and several aliases [Wik]. Wikidata items represent entities from human knowledge, such as people, countries, books, and diseases. For instance, *Myocardial Infarction* is an item representing a disease.

Wikidata utilizes the RDF (Resource Description Framework) format to store information, structured in "Subject-Predicate-Object" triples. "Subject" refers to an item being described, the "Predicate" is attribute of the "Subject", and the "Object" is the value of the "Property". The data structure of a Wikidata item as illustrated in Figure 4.2.

[2] some experiments were conducted as part of our participation in this shared task
[3] *Distant supervision data* is defined in Section 4.2
[4] https://www.wikidata.org

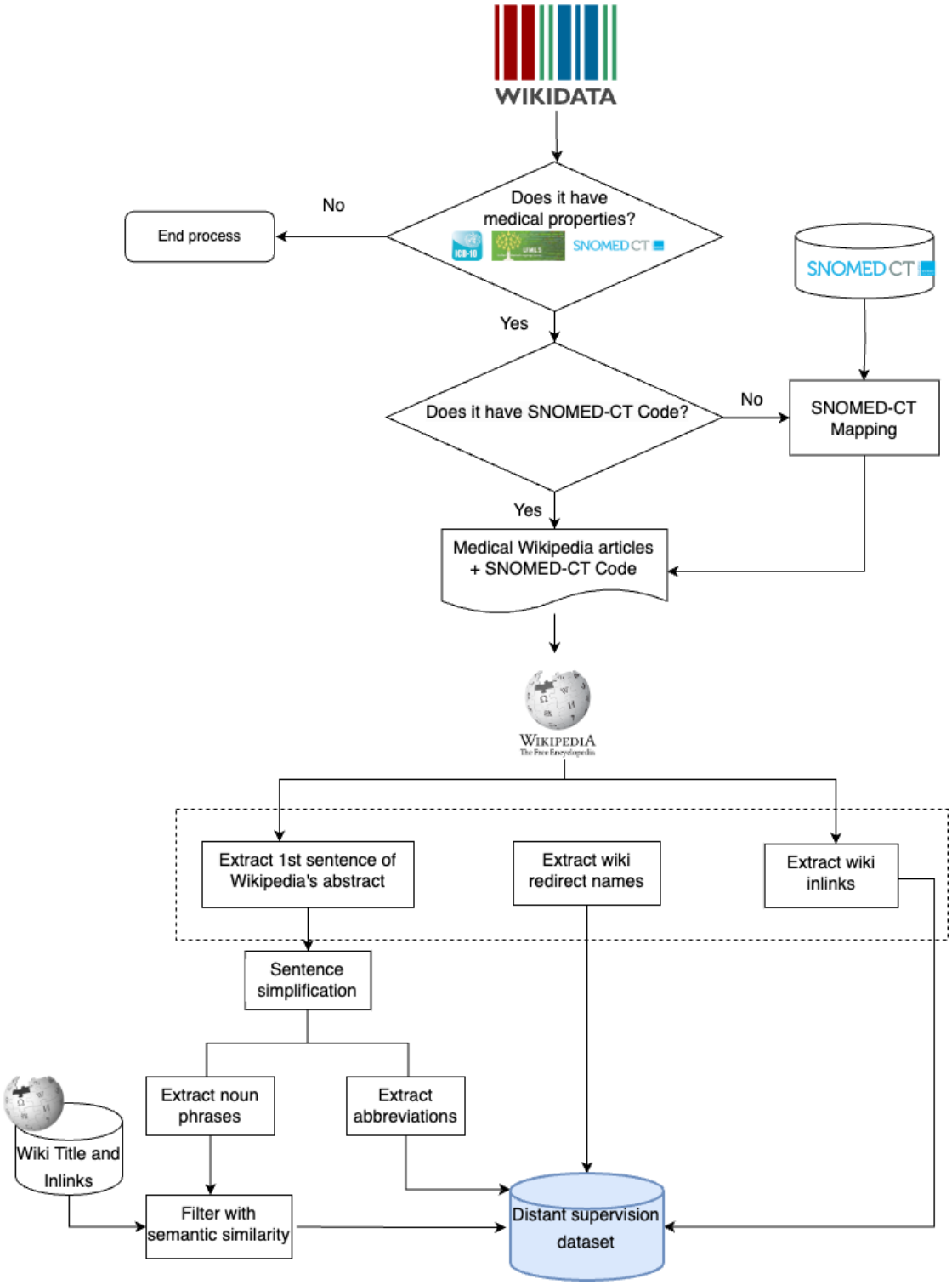


Figure 4.1: The distant supervision approach

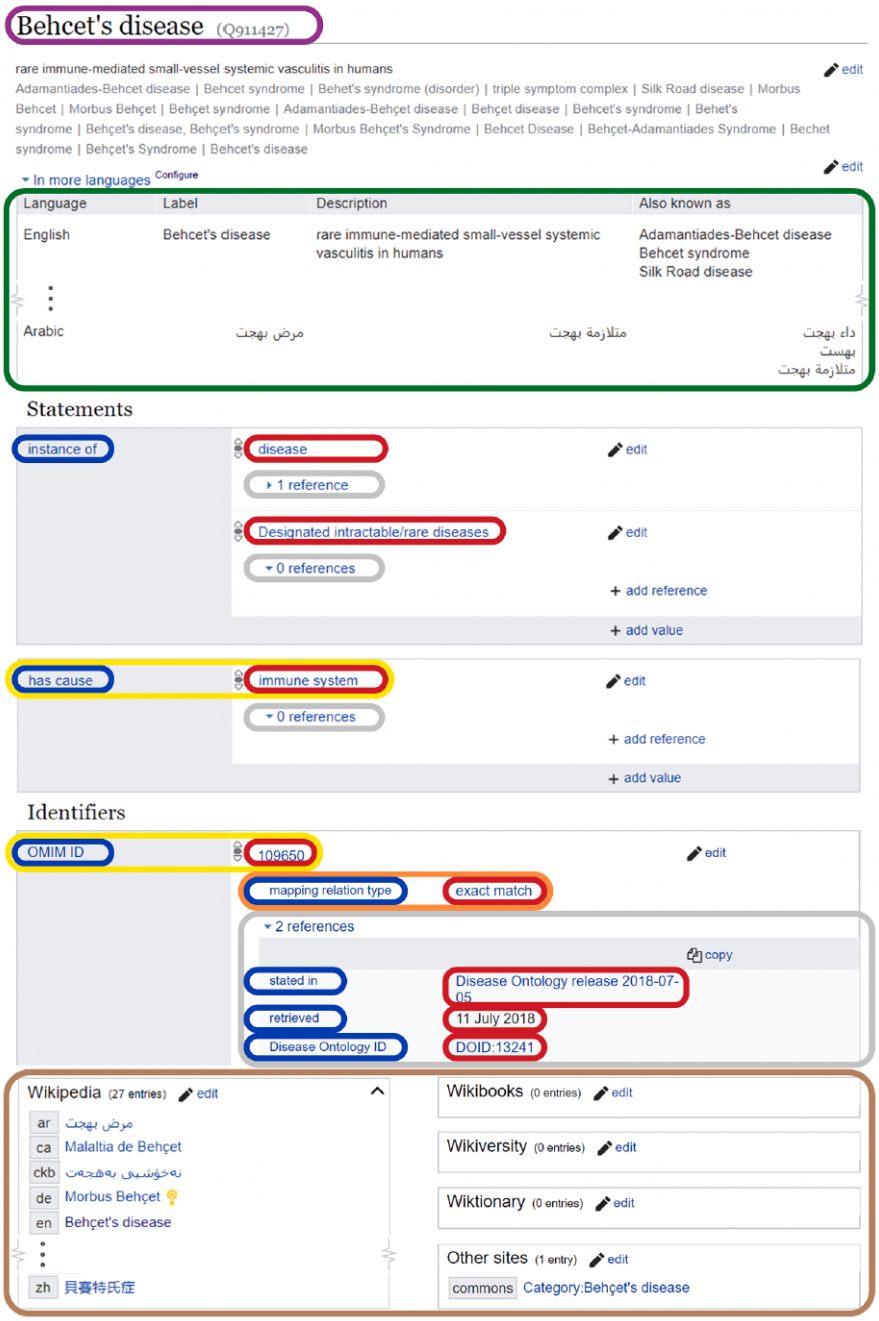


Figure 4.2: Data structure of Wikidata [TSH+19]. The components of a Wikidata item include: identifier (purple), multilingual labels, descriptions, and aliases (green), sitelinks to Wikimedia pages (brown), and statements comprising claims (yellow) and qualifiers (orange). Statements form triples where predicates are Wikidata properties (blue) and objects (red) can be various data types [TSH+19, Wik].

Wikidata is a large-scale ontological database covering various disciplines and includes numerous medical entries (items). These items cover a broad range from human genes and proteins, to diseases, drugs, and anatomical entities, with the most significant ones linked to corresponding articles in the four largest language editions of Wikipedia [TSH+19]. All of these items are interconnected, creating an extended biomedical taxonomy using taxonomic Wikidata properties such as *instance of (P31)*, *subclass of (P279)*, *part of (P361)*, and *has part (P527)* [TSH+19]. Additionally, medical entries in Wikidata can be linked to other items through medical relations, such as *symptoms (P789)* or *side effect (P1909)* [TSH+19]. Wikidata is also connected to external vocabularies, terminologies, or classifications used in the biomedical field, including Disease Ontology ID (P699), MeSH ID (P486), and SNOMED-CT identifier (P5806). For example, the Wikidata item for *Behcet's disease* (see Figure 4.2) is linked to the external database *Online Mendelian Inheritance in Man (OMIM ID) (P492)*, corresponding to OMIM ID *109650* (*Behcet syndrome*).

To effectively map Wikipedia articles to SNOMED-CT codes, we utilize Wikidata properties that correspond to SNOMED-CT identifiers. According to [NTK+19] around 46% of SNOMED-CT concepts are directly found in Wikipedia. Additionally, we integrate properties from the Unified Medical Language System (UMLS) and the International Classification of Diseases, Tenth Revision (ICD-10), to broaden the range of medical-related Wikipedia articles we use in our research. The UMLS encompasses various biomedical vocabularies and standards covering drugs, disorders, genes, anatomy, and medical devices[5]. Additionally, we include the ICD-10 property to extract Wikipedia articles, leveraging the existing mappings between SNOMED-CT and ICD-10. Both SNOMED-CT and ICD-10 are also utilized in Electronic Health Records (EHR).

Despite some Wikidata items lacking a direct SNOMED-CT link, many are associated with UMLS or ICD-10 codes. Furthermore, although UMLS encompasses various biomedical vocabularies, including SNOMED-CT and ICD-10, we have identified Wikidata items linked to Wikipedia articles associated only with SNOMED-CT or ICD-10. Table 4.2 shows the statistics of the number of Wikidata items associated with each property. Among the three biomedical external identifier properties selected, the number of Wikidata

Table 4.2: Number of Wikidata items associated with UMLS, SNOMED-CT, and ICD-10

#Wikidata w/ UMLS	**#Wikidata w/ SCT**	**#Wikidata w/ ICD-10**
731,428	1,101	4,649

items linked only to UMLS codes is higher than those linked to SNOMED-CT and ICD-10, with 731,428 items linked to UMLS. This is regardless of their association with Wikipedia entities. In comparison, 4,649 Wikidata items are associated with ICD-10 codes. Among these, 1,437 out of the 4,649 items (30.9%) are linked to ICD-10 codes without being linked to any UMLS concepts. Additionally, 1,101 Wikidata items are

[5] *https://www.nlm.nih.gov/bsd/disted/video/clininfo/umls.html*

mapped to SNOMED-CT concepts, with 245 Wikidata items (22.25%) linked only to SNOMED-CT concepts and not to UMLS concepts.

Table 4.3 presents the distribution of medical-related Wikidata items per category, linked to English Wikipedia entities. We sampled the categories based on the most common categories searched by laypeople [MBS24]. Among Wikidata items that have corresponding Wikipedia entities, more are linked to UMLS and ICD-10 than to SNOMED-CT.

For the *disease* category, out of 4,700 Wikidata items, 3,070 are linked to UMLS and 1,630 to ICD-10. In contrast, only 290 Wikidata items are linked to SNOMED-CT. In the *Drugs* category, there are 972 Wikidata items, with 112 of these linked to UMLS. Based on these observations, we incorporated these three properties to capture a broader range of Wikipedia articles and medical concepts. However, the inconsistent coverage of SNOMED-CT in Wikidata items requires an additional mapping step from ICD-10 and UMLS to SNOMED-CT. In our work, we use the National Library of Medicine (NLM), including SNOMED-CT US edition[6] and its UMLS[7], and SNOMED International[8] knowledge base.

Table 4.3: Summary of Data for Different Categories

Category	**Total Items**	**Items w/ UMLS**	**Item w/o UMLS**	**Items w/ ICD-10**	**Items w/o ICD-10**	**Items w/ SCT**	**Items w/o SCT**
Disease	4,700	3,070	1,630	1,099	3,601	290	4,410
Drugs	972	112	860	0	972	0	972
Symptom	623	449	174	161	462	46	577
Procedures	1,300	270	1,030	4	1,296	4	1,296
Anatomical Structure	2,967	1,603	1,364	0	2,967	52	2,915

4.1.2 Identifying Medical Articles from Wikipedia

Wikipedia contains a wide range of articles, including those related to medical information. To identify the medical-related Wikipedia articles relevant to our distant supervision approach, we rely on Wikidata's biomedical external identifier properties rather than Wikipedia's category labels. We focus on three specific biomedical external identifier properties from Wikidata: (1) SNOMED-CT, (2) ICD-10, and (3) UMLS codes. If an article is associated with a concept from any of these three properties, we consider it to contain medical information. The identification process can be illustrated in Figure 4.3. Let W represent the set of all Wikipedia articles. Define three sets representing the biomedical external identifier properties from Wikidata:

[6]https://www.nlm.nih.gov/healthit/snomedct/us_edition.html
[7]https://www.nlm.nih.gov/research/umls/index.html.
[8]https://www.nlm.nih.gov/healthit/snomedct/international.html

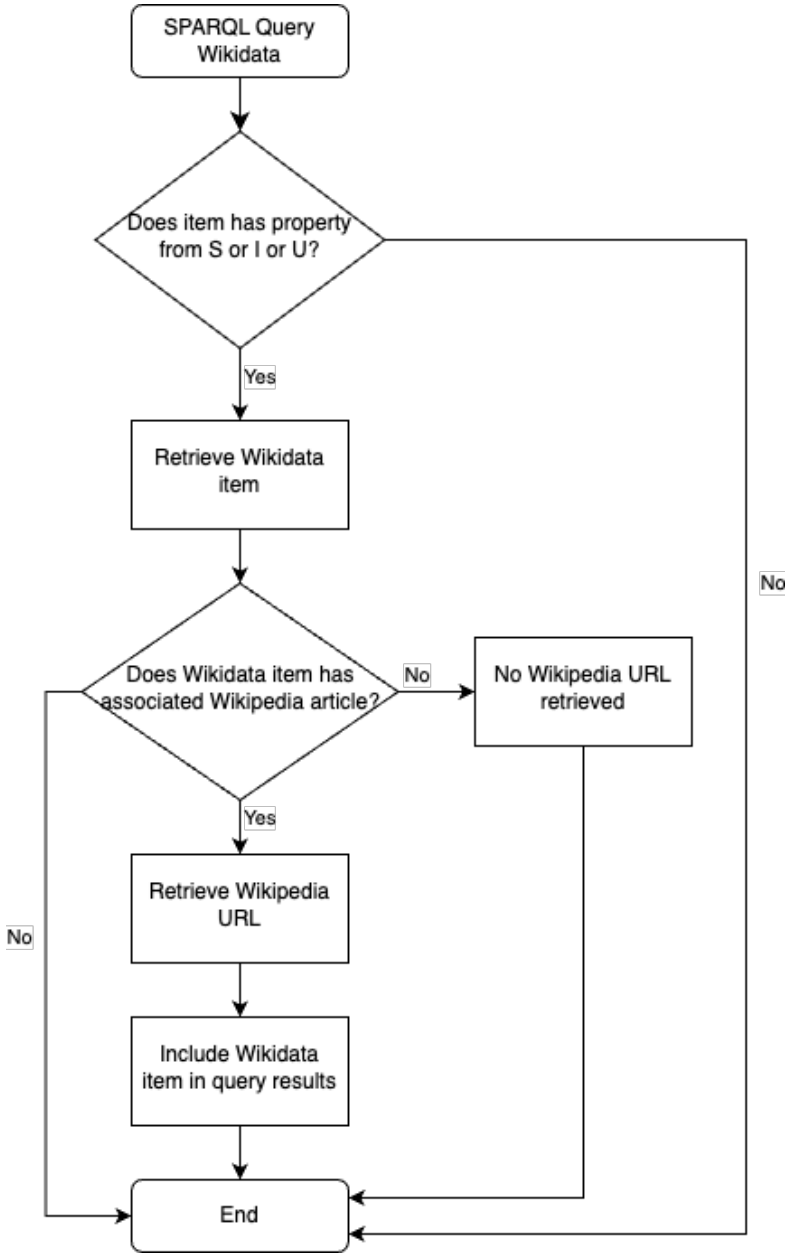


Figure 4.3: Medical-related Wikipedia Articles Extraction Workflow

- S: the set of SNOMED-CT codes
- I: the set of ICD-10 codes
- U: the set of UMLS codes

For each Wikidata item $d(w)$, check if it has at least one property from the sets S, I, or

U. If $d(w)$ has a property from S, I, or U, and is associated with a Wikipedia article w, we retrieve the item's label and the URL of the associated Wikipedia article. This process can be implemented by issuing a SPARQL query to the Wikidata SPARQL query endpoint. The query is as follows:

```
SELECT DISTINCT ?item ?itemLabel ?sct ?cui ?icd ?wiki_url WHERE {
  {
    SELECT ?item ?sct WHERE {
      ?item wdt:P5806 ?sct.
    }
  } UNION {
    SELECT ?item ?cui WHERE {
      ?item wdt:P2892 ?cui.
    }
  } UNION {
    SELECT ?item ?icd WHERE {
      ?item wdt:P494 ?icd.
    }
  }
  OPTIONAL {
    ?wiki_url schema:about ?item;
              schema:inLanguage "en".
    FILTER (STRSTARTS(STR(?wiki_url), "https://en.wikipedia.org/"))
  }
  SERVICE wikibase:label {
    bd:serviceParam wikibase:language "en".
    ?item rdfs:label ?itemLabel.
  }
}
```

The query, denoted as Q, retrieves the relevant information for each Wikidata item, including its unique ID (*item*), most common name (*itemLabel*), SNOMED-CT code (*sct*), UMLS code (*cui*), ICD-10 code (*icd*), and the URL for the corresponding Wikipedia article (*wiki_url*). Figure 4.4 illustrates Wikidata item's structure with its related Wikipedia article. The result of the

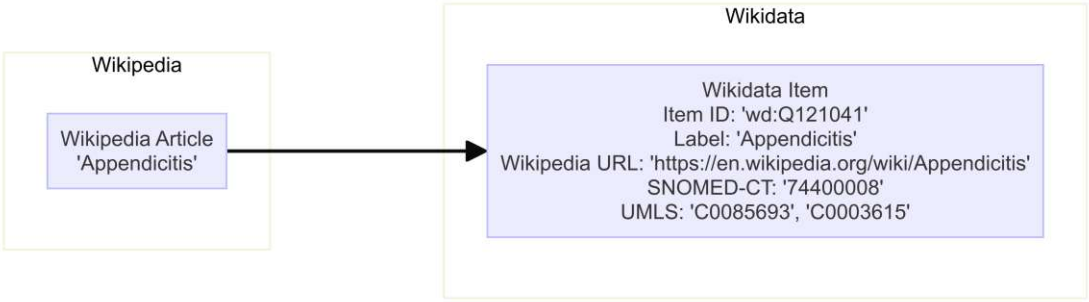


Figure 4.4: Wikidata Item's Structure with Biomedical External Identifier Properties and Related Wikipedia Article

SPARQL query Q is a set of Wikidata items, where each item corresponds to a Wikipedia article w corresponds to a Wikidata item, denoted as $d(w)$, which is the result of executing Q. The properties of a Wikidata item $d(w)$ are represented as key-value pairs in the following format:

$$\begin{aligned} d(w) = \{&\text{Wikidata ID} : \{\text{key} : \text{value}\} \mid \\ &\text{key} \in \{\text{itemLabel}, \text{Wikipedia URL}, \text{SNOMED-CT}, \text{ICD-10}, \text{UMLS}\}, \\ &\text{value} : \text{string (constant)}\} \end{aligned}$$

In this context, *key* refers to a specific property or attribute of Wikidata item $d(w)$, such as its label, Wikipedia URL, or medical codes. Meanwhile, the *value* is the data associated with each *key*, typically a string of text. For example, in the context of biomedical external identifier properties like SNOMED-CT, the *key* is the property name (e.g 'SNOMED-CT') and the *value* is the corresponding code associated with $d(w)$. For each Wikidata item $d(w)$, three subsets are derived to extract and categorize the biomedical external identifier properties:

- S(d(w)): Set of SNOMED-CT codes associated with the Wikidata item $d(w)$.
- U(d(w)): Set of UMLS codes associated with the Wikidata item $d(w)$.
- I(d(w)): Set of ICD-10 codes associated with the Wikidata item $d(w)$.

These subsets are populated based on the corresponding medical code properties in the Wikidata. For illustration, the subsets $S(d(w))$, $U(d(w))$, and $I(d(w))$ can be exemplified using the following Wikidata items:

- For the Wikidata item wd:Q121041 (Appendicitis):

$$\begin{aligned} S(d(\text{Q121041})) &= \{74400008\} \\ U(d(\text{Q121041})) &= \{C0085693, C0003615\} \\ I(d(\text{Q121041})) &= \emptyset \text{ (no ICD-10 code)} \end{aligned}$$

- For the Wikidata item wd:Q68833 (Bone Fracture):

$$\begin{aligned} S(d(\text{Q68833})) &= \emptyset \text{ (no SNOMED-CT code)} \\ U(d(\text{Q68833})) &= \{C0016658\} \\ I(d(\text{Q68833})) &= \{T14.2\} \end{aligned}$$

- For the Wikidata item wd:Q147362 (Ovarian Cyst):

$$\begin{aligned} S(d(\text{Q147362})) &= \emptyset \text{ (no SNOMED-CT code)} \\ U(d(\text{Q147362})) &= \{C0029513\} \\ I(d(\text{Q147362})) &= \emptyset \text{ (no ICD-10 code)} \end{aligned}$$

4.1.3 Mapping Wikipedia Articles to SNOMED-CT

For each Wikipedia article w identified as a medical article (Figure 4.4), we aim to map it to a SNOMED-CT concept. The mapping process utilizes the properties retrieved from the associated Wikidata item ($d(w)$) and the subsets ($S(d(w))$), ($I(d(w))$), and ($U(d(w))$) defined previously.

We use a lookup table from the UMLS Metathesaurus and SNOMED-CT to map Wikipedia articles to SNOMED-CT concepts. This table allows us to retrieve the appropriate medical concept by matching the properties of a Wikidata item to its corresponding SNOMED-CT concept. The lookup process can be performed using SQL queries or a simple dictionary look-up method. The mapping process is described in Algorithm 4.1.

Algorithm 4.1: Mapping Wikipedia Articles to SNOMED-CT via Wikidata

Input : Set of Wikipedia articles W identified as medical articles
Output : Mapping of Wikipedia articles to SNOMED-CT concepts

foreach $w \in Wiki$ **do**
  $d_w \leftarrow$ Wikidata item associated with w
  if *"SNOMED-CT" key exists in* d_w **then**
    $s_w \leftarrow S(d_w)$ `// Direct mapping to SNOMED-CT`
  end
  else
    $candidates \leftarrow \emptyset$ `// Initialize candidate list`
    if *"ICD-10" key exists in* d_w **then**
      $icd10\_code \leftarrow I(d_w)$
      $snomed\_from\_icd10 \leftarrow$ mapICD10toSNOMED($icd10\_code$)
      $candidates \leftarrow candidates \cup \{snomed\_from\_icd10\}$
    end
    if *"UMLS" key exists in* d_w **then**
      $umls\_code \leftarrow U(d_w)$
      $snomed\_from\_umls \leftarrow$ mapUMLStoSNOMED($umls\_code$)
      $candidates \leftarrow candidates \cup \{snomed\_from\_umls\}$
    end
    if $candidates \neq \emptyset$ **then**
      $best\_match \leftarrow$ Levenshtein_dist($candidates$.SCT_description, d_w.label)
      $s_w \leftarrow best\_match$
    end
    else
      $s_w \leftarrow$ "No mapping found"
    end
  end
  Output: $w \rightarrow s_w$
end

The output of these functions, $snomed\_from\_icd10$ and $snomed\_from\_umls$, is a pair consisting of the SNOMED Concept Unique Identifier (SCUI) and the associated SNOMED description (refer to SCT_description). Below is the illustration of the mapping Wikipedia articles to SNOMED-CT

- For the Wikidata item `wd:Q121041` (Appendicitis):

$$\begin{aligned} S(d(\text{Q121041})) &= \{74400008\} \\ U(d(\text{Q121041})) &= \{C0085693, C0003615\} \\ I(d(\text{Q121041})) &= \emptyset \text{ (no ICD-10 code)} \\ \text{Mapping Output: } s_w &= \{74400008\} \text{ (Direct SNOMED-CT mapping)} \end{aligned}$$

- For the Wikidata item `wd:Q68833` (Bone Fracture):

$$\begin{aligned} S(d(\text{Q68833})) &= \emptyset \text{ (no SNOMED-CT code)} \\ U(d(\text{Q68833})) &= \{C0016658\} \\ I(d(\text{Q68833})) &= \{T14.2\} \\ s_{u_w} &= \text{mapUMLStoSNOMED}((U(d(\text{Q68833})))) \\ &= \{(125605004, \text{fracture of bone})\} \\ s_{i_w} &= \text{mapICD10toSNOMED}(I(d(\text{Q68833}))) \\ &= \{(1003502008, \text{traumatic fracture of bone})\} \\ \text{Best Match: } &= \{(125605004, \text{fracture of bone})\} \\ &\quad \text{(Selected based on shortest Levenshtein distance)} \end{aligned}$$

- For the Wikidata item `wd:Q147362` (Ovarian Cyst):

$$\begin{aligned} S(d(\text{Q147362})) &= \emptyset \text{ (no SNOMED-CT code)} \\ U(d(\text{Q147362})) &= \{C0029513\} \\ I(d(\text{Q147362})) &= \emptyset \text{ (no ICD-10 code)} \\ s_{u_w} &= \text{mapUMLStoSNOMED}(U(d(\text{Q147362}))) \\ &= \{(79883001, \text{cyst of ovary})\} \\ \text{Mapping output: } &= \{(79883001, \text{cyst of ovary})\} \end{aligned}$$

For Wikidata item `wd:Q121041` (appendicitis), since there is a direct SNOMED-CT code available $S(d(\text{Q121041})) = \{74400008\}$, we use this code directly. There is no need to use the UMLS or ICD-10 codes as we already have a direct mapping. Meanwhile, for Wikidata item `wd:Q68833`, there is no direct mapping to SNOMED-CT code ($S(d(\text{Q68833})) = \{\emptyset\}$); as a result, we look at the other codes.

We would utilize both UMLS and ICD-10 mapping functions to identify the corresponding SNOMED-CT concepts. Specifically, let $s_{u_w} = \text{mapUMLStoSNOMED}(U(d(\text{Q68833})))$ and $s_{i_w} = \text{mapICD10toSNOMED}(I(d(\text{Q68833})))$. The results from s_{u_w} and s_{i_w} provide SNOMED-CT code (SCUI) and description (SCT description) pairs. The most suitable SNOMED-CT match for the Wikidata item is then selected based on the minimal Levenshtein distance to the Wikidata item label. For example, selecting (125605004, 'fracture of bone') for the label "bone fracture".

Furthermore, for Wikidata item `wd:Q147362` (ovarian cyst), where only a UMLS code is available, the mapping function $s_{u_w} = \text{mapUMLStoSNOMED}(U(d(\text{Q147362})))$ is applied to find a matching SNOMED-CT concept. The selected SNOMED-CT match is based on the shortest Levenshtein distance to the Wikidata item label, with s_{u_w} in this case yielding $(79883001, \text{cyst of ovary})$. Table 4.4 shows the final output of the mapping process of the Wikipedia article to SNOMED-CT.

Table 4.4: Example of the results from aligning Wikipedia articles with SNOMED-CT codes via Wikidata. This table displays SNOMED-CT codes that are obtained using the previously outlined method.

Item	ItemLabel	Wikipedia URL	SNOMED-CT
wd:Q121041	appendicitis	Appendicitis	74400008
wd:Q68833	bone fracture	Bone Fracture	125605004
wd:Q147362	ovarian cyst	Ovarian Cyst	79883001

4.1.4 Lay Medical Term Extraction

Once the extracted medical articles from Wikipedia via Wikidata have been mapped to SNOMED-CT concepts, we extract the medical phrases from the *article summaries*, *redirect pages*, and *wikilinks* as detailed in Section 4.1.1. We denote the Wikipedia dump as $Wiki$. Each article $wiki$ within $Wiki$ is represented as $d(wiki)$, where $wiki$ refers to a Wikipedia article extracted from the previous step. The properties of $d(wiki)$ are defined as follows:

$$\begin{aligned} d(wiki) = \{&\text{Wikipedia ID} : \{\text{key} : \text{value}\} \mid \\ &\text{key} \in \{\text{title}, \text{abstract summary}, \text{redirect pages}, \text{wikilinks}\}, \\ &\text{value} : \text{string}\} \end{aligned}$$

Extracting Medical Phrases from the Article Summary

According to our observations, the *first sentence* of a Wikipedia medical article may be too technical (i.e. already has the specialized medical terms) to represent the MCN task. We first experimented with Simple English Wikipedia. Simple English Wikipedia is an online encyclopedia intended to provide a simple English version of Wikipedia for people whose first language is not English [Sim, PSG08]. However, these simplified articles are not as complete in their information as the main Wikipedia articles. Therefore, we simplified the text of the main Wikipedia article using the Multilingual Unsupervised Sentence Simplification (MUSS) sentence simplification model [MFdlC+20]. The goal of the MUSS is to paraphrase the sentence to create simpler versions of sentences while preserving their original meaning. We decided to use MUSS sentence simplification as this model is trained, among others, also on Simple English Wikipedia.

By using MUSS we transform the *first sentence* into a simpler version. Based on this simplified sentence, the layman's definition of a medical term (that is, the popularized medical phrase) is extracted. This simplified sentence then served as the basis for extracting noun phrases and abbreviations, which are essential for identifying popularized medical phrases.

We recognize that simplifying sentences may cause some important details to be lost or the information to be changed slightly. However, using MUSS still has benefits, as it aims to create simpler versions of sentences while preserving their original meaning, making complex information more accessible and easier to understand. Additionally, we plan to conduct further evaluations to assess the impact of simplification on dataset quality and model performance. The process is describe in Algorithm 4.2.

Noun Phrase Detection and Extension Medical terms are initially extracted through Noun Phrase Detection, which identifies groups of words functioning as nouns in a sentence. However, this method can yield incomplete phrases. To refine this, we examine the sentence's dependency tree. This grammatical structure shows how words in a sentence are related, and by identifying the leftmost and rightmost syntactic descendants of a token within this tree, we can extend the noun phrases for a more comprehensive extraction. For this task, we first simplify the

Algorithm 4.2: Extract Lay Medical Terms

Input: WikipediaArticle
Output: MedicalTerms
1 **Function** *ExtractMedicalTermsWikipediaArticle*:
2 $simplified\_summary \leftarrow$ `MUSS`$(WikipediaArticle.abstract\_summary)$;
3 $noun\_indices \leftarrow$ `NounPhraseDetection`$(simplified\_summary)$;
4 $dependency\_tree \leftarrow$ `DependencyTree`$(simplified\_summary)$;
5 $extended\_noun\_phrases \leftarrow []$;
6 **for** $noun\_index$ ***in*** $noun\_indices$ **do**
7 $left\_edge \leftarrow$ `FindLeftmostDescendant`$(noun\_index, dependency\_tree)$;
8 $right\_edge \leftarrow$ `FindRightmostDescendant`$(noun\_index, dependency\_tree)$;
9 $extended\_noun\_phrase \leftarrow simplified\_summary[left\_edge : right\_edge + 1]$;
10 $extended\_noun\_phrases.append(extended\_noun\_phrase)$;
11 **end**
12 $abbreviations \leftarrow$ `AbbreviationDetector`$(simplified\_summary)$;
13 $medical\_terms \leftarrow extended\_noun\_phrases \cup abbreviations$;
14 **return** $medical\_terms$;

first sentence of the abstract summary of $d(wiki)$ using the MUSS model (Algorithm 4.2, lines 2). The next step involves extracting noun phrases from *simplied_summary* of $d(wiki)$ (Algorithm 4.2, lines 3). We then extracted the noun phrases and the extended noun phrases derived from its dependency tree (Algorithm 4.2, lines 4).

Here, *dependency_tree* represents the dependency tree of the simplified text $simplified\_summary$. This tree is important for determining the syntactic structure of the text, which is necessary for identifying the bounds of noun phrases as illustrated in Figure 4.5.

Using the dependency tree, the step of extracting extended noun phrases is shown in lines 5-10 from Algorithm 4.2.

- *extended_noun_phrase* represents the set of extended noun phrases extracted using the dependency tree *dependency_tree* from the $simplified\_summary$.
- *noun_indices* is the set of indices of nouns in the $simplified\_summary$ text.
- $FindLeftmostDescendant$ and $FindRightmostDescendant$ are functions that return the indices of the leftmost and rightmost tokens, respectively, connected to the noun at *noun_index* in the dependency tree *dependency_tree*.

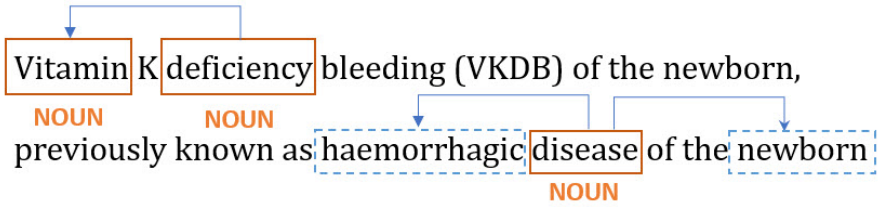


Figure 4.5: Example of an extended noun phrases extraction. The arrows show the leftmost and rightmost syntactic descendants of a token. The result of this extraction is (1) Vitamin (2) Vitamin K deficiency and (3) hemorrhagic disease of the newborn.

Abbreviation Detection: We observed that laypeople frequently use abbreviations. To capture these, we employ the *AbbreviationDetector* function from SciSpacy, which identifies abbreviations within a sentence (Algorithm 4.2, lines 12). For instance, as shown in Figure 4.5, "VKDB" is identified as an abbreviation for "Vitamin K deficiency bleeding." We incorporate these abbreviations into our dataset, forming *abbrevations*, the set of abbreviations identified in $simplified\_summary$.

Combining Extended Noun Phrases and Abbreviations: The combined set of popularized medical phrases extracted from the article summary, including both *extended_noun_phrases* and *abbreviations*, is defined as follows:

$$\text{medical\_terms} = \text{extended\_noun\_phrases} \cup \text{abbreviations}$$

medical_terms represents the union of the set of extended noun phrases *extended_noun_phrases* and the set of abbreviations *abbreviations*, thus encompassing a comprehensive set of popularized medical phrases present in the first simplified sentence of the Wikipedia article summary.

Extracting Medical Phrases from Wikipedia *Redirect* Pages and Wikilinks

As explained in Section 4.1.1, the Wikipedia dump contains Wikipedia's *redirect* pages and wikilinks. A Wikipedia *redirect* is a page that directs users to another page. For example, *Heart Attack* is redirected to the *Myocardial Infarction* article. Redirects involve Wikipedia article titles.

A Wikilink is an internal link to another Wikipedia page. For example, there are many hyperlinks in the Wikipedia article on *Chest Pain*, one of which is attached to the term *heart attack*. This term refers to the Wikipedia article on *Myocardial Infarction*. In this way, we take the *heart attack* phrase as a popularized medical phrase for *Myocardial Infarction*. We collected all of the redirects and wikilinks associated with $d(wiki), wiki \in Wiki$ and removed duplicates as they appeared.

Here, $RedirectTerms$ and $WikilinkTerms$ are associated with the Wikipedia data $d(wiki)$. $RedirectTerms$ refers to the data related to the redirect pages of the title of $d(wiki)$, while $WikilinkTerms$ refers to the data related to the wikilink of $d(wiki)$. To extract $RedirectTerms$ and $WikilinkTerms$ from $d(wiki)$, the Wikipedia dump was indexed, and dictionary/string matching techniques were employed using the title of $d(wiki)$ as an input. The combined set of medical terms extracted from both redirects and wikilinks is:

$$\text{RelatedMedicalTerms} = \text{RedirectTerms} \cup \text{WikilinkTerms}$$

Algorithm 4.3: Extract Related Medical Terms from Wikipedia Article

Input: WikipediaArticle
Output: RelatedMedicalTerms

Function `ExtractRelatedMedicalTerms`(*WikipediaArticle.title*)**:**
  RedirectTerms ← `ExtractRedirects`(*WikipediaArticle.title*)
  WikilinkTerms ← `ExtractWikilinks`(*WikipediaArticle.title*)
  RelatedMedicalTerms ← *RedirectTerms* ∪ *WikilinkTerms*
  return *RelatedMedicalTerms*
Function `ExtractRedirects`(*title*)**:**
  RedirectTerms ← *Retrieve redirect pages associated with title from Wikipedia dump*
  return *RedirectTerms*
Function `ExtractWikilinks`(*title*)**:**
  WikilinkTerms ← *Retrieve wikilinks associated with title from Wikipedia dump*
  return *WikilinkTerms*

The final set of lay medical terms associated with $d(wiki)$ concept is:

$$M = medical\_terms \cup RelatedMedicalTerms$$

where, $RelatedMedicalTerms$ represents the union of medical terms extracted from redirect pages of title of $d(wiki)$ and the wikilink of $d(wiki)$; and M is the final aggregated set, combining $medical\_terms$ (a previously defined set of medical terms extracted from Algorithm 4.2) with $RelatedMedicalTerms$ from Algorithm 4.3.

Table 4.5 displays the results of our distant supervision method, demonstrating the data extraction process. We utilized the Wikipedia Article extracted from Section 4.1.2 as the primary source for extracting popularized medical phrases, following the process explained in Section 4.1.4. Additionally, specialized medical terms were obtained from SNOMED-CT codes through the process detailed in Section 4.1.3.

4.2 Experiment Setup

We developed a dataset using a distant supervision method, referred to as *distant supervision data*. We conducted three experiments to assess the effectiveness of augmenting this distant supervision with three existing MCN (Medical Concept Normalization) datasets: CADEC, PsyTAR, and COMETA. This was done to create larger training sets for our MCN model. The process of creating training data involved several steps:

1. **Identifying Common Concepts**:
   - We compared the medical concepts in our *distant supervision data* with those in the original training sets of CADEC, PsyTAR, and COMETA.

Table 4.5: Example of the final results of our distant supervision approach for generating the Medical Concept Normalization (MCN) dataset using Wikipedia and Wikidata as the primary sources.

Wikipedia Article	**SNOMED-CT Code**	**Popularized Medical Phrases**
Appendicitis	74400008	{*appendicitis*, *appendix rupture*, *appendictic*}
Bone Fracture	125605004	{*bone fractures*, *bone breaking*, *broken bones*}
Ovarian Cyst	79883001	{*ovarian cyst*, *ovary cyst*, *an ovarian cyst*}

- We identified common medical concepts between our distant supervision data and each public MCN dataset. Table 4.6 displays the distribution of these overlapping concepts across each dataset. These common concepts are treated as class labels:
  - c_1 for CADEC
  - c_2 for PsyTAR
  - c_3 for COMETA

2. **Creating Combined Datasets (UD Sets)**:
   - We extracted subsets of our distant supervision data that contained the overlapping concepts identified in Step 1.
   - We merged these subsets with the corresponding public MCN datasets to create three new training sets:
     - UD_1 = CADEC Subset + Distant Supervision Data Subset for CADEC
     - UD_2 = PsyTAR Subset + Distant Supervision Data Subset for PsyTAR
     - UD_3 = COMETA Subset + Distant Supervision Data Subset for COMETA
3. **Extracting Original Data Subsets (CD Sets)**:
   - From each public MCN dataset, we selected only the training data that contained the overlapping concepts from Step 1.
   - These subsets serve as training sets to assess the effectiveness of training on data specific to the overlapping concepts:
     - CD_1 = CADEC Subset
     - CD_2 = PsyTAR Subset
     - CD_3 = COMETA Subset
4. **Training with Distant Supervision Data Subsets (DD Sets)**:
   - We created subsets from our distant supervision data containing the overlapping concepts identified in Step 1.
   - We trained our MCN model using these subsets, while validating and testing with the original public MCN datasets:

Table 4.6: Number of unique SNOMED-CT concepts in each dataset and the number of SNOMED-CT concepts that overlap with our distant supervision dataset, DD.

Dataset	Unique concepts	Concepts in DD
CADEC	1,029 [TMNM18]	58 (c_1)
PsyTAR	755 [ZFP+19]	195 (c_2)
COMETA	3,645 [BLSC20]	1,247 (c_3)

- DD_1 = Distant Supervision Data Subset for CADEC
- DD_2 = Distant Supervision Data Subset for PsyTAR
- DD_3 = Distant Supervision Data Subset for COMETA

The validation of our proposed distant supervision data was carried out by using it to train a model for the Medical Concept Normalization (MCN) task. This task involves mapping popularized medical phrases to specific SNOMED-CT concepts, which act as class labels in a multi-class classification setting.

Furthermore, we conducted an additional experiment to combine all available datasets for training the MCN model. This enables the MCN model to cover even more of the SNOMED-CT concepts. We combined the CADEC, PsyTAR, COMETA, as well and distant supervision data into one large corpus in the *Unify Datasets* step, as shown in Figure 1.2. Furthermore, we incorporated SNOMED-CT synonyms, to include all concepts found in SNOMED-CT. In this experiment, we created two sets of SNOMED-CT synonyms. First, we included all synonyms of SNOMED-CT concepts from all entity types within the SNOMED-CT (referred to *Full SCT*). Secondly, we selected synonyms of SNOMED-CT concepts specifically from certain types of entities, including *findings*, *substances*, *body structures*, *observable entities*, *procedures*, *disorders*, *organisms*, and *morphologic abnormalities* (referred to *Partial SCT*).

In the pre-processing stage of the distant supervision data, we identified popularized medical phrases that correspond to multiple SNOMED-CT concepts. For example, the phrase *baldness* appears in two medical concepts: *Alopecia* (SCUI: 56317004) and *Alopecia hereditary* (SCUI:201144006). While such ambiguity is common in multi-class classification tasks, we decided to remove these terms to focus on more precise mappings between popularized phrases and SNOMED-CT concepts for this particular study. We acknowledge that this decision may result in the loss of potentially valuable information for the classification task, and we plan to explore alternative approaches to handle ambiguous terms in future work. We also removed emoticons and excluded the SNOMED-CT concepts that we consider irrelevant for the MCN dataset, such as concepts with *geographical locations* type.

Then, we merge the training data from the CADEC, PsyTAR, and COMETA datasets, along with the distant supervision data and SNOMED-CT synonyms, to create our final train set, which is used in the complete pipeline of the informal medical entity linking model (see Figure 1.2). We augment medical concepts that are represented by a single popularized medical phrase by using the synonym replacement using WordNet (see Table 3.2 in Chapter 3). We use the synonym replacement technique due to the efficient computation and resources to maintain good performance. Table 4.7 presents the statistics of the training data used for training the MCN models. We utilized the original test sets from CADEC (CA), PsyTAR (PS), and COMETA (CO) to evaluate the performance of the MCN model.

Table 4.7: Number of unique terms and concepts in the datasets employed for training MCN models in different experiments. The names of the training datasets are explained as follows: DS = distant supervision dataset produced by our proposed approach, FULL = All synonyms in SNOMED-CT, PART = Synonyms in SNOMED-CT from several entity types.

Training data name	Unique Concepts	Unique Terms
CA ∪ PS ∪ CO ∪ DS ∪ FULL	351,042	818,702
CA ∪ PS ∪ CO ∪ DS ∪ PART	246,654	621,436

Our experiment included tests using the CADEC dataset, following the 5-fold data split method outlined by Tutubalina et al. [TMNM18]. For the PsyTAR [ZFP+19] and COMETA [BLSC20] datasets, we conducted them three times using the same train and test sets to observe consistent performance in our model with different random seeds. We evaluated our MCN model performance by using a micro-F1-score.

4.2.1 Model Training

Our MCN model has the same model architecture as the one detailed in Chapter 3. Our model utilizes a Gated Recurrent Units (GRU) algorithm [NHPA21], with each SNOMED-CT concept represented by a node in the model's final linear layer.

For feature representation, we employed two types of contextual embeddings. The first, Hunflair [WSM+20], is specialized in the biomedical domain and was trained on 23 different biomedical datasets. The second embedding, RoBERTa [LOG+19], is more general-purpose and was trained on a diverse set of five datasets: BookCorpus, English Wikipedia, CC-News, OpenWebText, and Stories. Both embeddings are implemented within the FlairNLP framework [ABB+19], which is also used for the text classification components of our model.

The model processes input text in three main steps. First, it computes embeddings for each word in the input sequence using both Hunflair and Roberta. Next, these embeddings are passed to the GRU layer, which includes reset and update gates. This allows the model to capture contextual information both preceding and following each word, creating a comprehensive sequence representation. Finally, a softmax layer is applied to perform multi-class classification, determining the appropriate SNOMED-CT code class for the input phrase (i.e., popularized medical term).

4.3 Results and Discussion

We extracted 12,101 Wikipedia articles using the method described in Section 4.1.2. Within these articles, we identified 9,759 distinct concepts and extracted about 122,628 popularized medical phrases, as described in 4.1.4. We identified that about 1,301 SNOMED-CT concepts overlapped with the union concepts of CADEC, PsyTAR, and COMETA. Table 4.6 shows the statistics of the overlap concepts (see Table 4.6, column: *Concepts in DD*).

The average of the F1-score for our MCN model trained on the evaluation scenarios described in Section 4.2 are shown in Table 4.8. The results show that the model performs better when trained with datasets UD_1, UD_2, UD_3 compared to those trained with CD_1, CD_2, CD_3 and

with DD_1, DD_2, DD_3(bold value in Table 4.8). Additionally, RoBERTa performed better in CADEC and PsyTAR, which have more popularized phrases. Meanwhile, Hunflair performed better in COMETA, which has more semi-formal phrases. RoBERTa, pre-trained on a large corpus of English text including books and Wikipedia articles [LOG+19], performs better on the CADEC and PsyTAR datasets, which contain informal language. However, its performance on the COMETA dataset, consisting of semi-formal phrases, is comparatively lower. This can be attributed to the fact that the COMETA dataset contains more semi-formal and structured medical language, which may not align as well with the non-specialized language RoBERTa was exposed to during pre-training.

In contrast, HunFlair, trained on a combination of medical and biomedical datasets [WSM+20], outperforms RoBERTa on the COMETA dataset. This performance difference can be attributed to the nature of the training data used for each model. RoBERTa's pre-training on diverse English text, including informal language, allows it to better handle the linguistic style found in CADEC and PsyTAR, capturing the nuances of informal language more effectively. On the other hand, HunFlair's training data, primarily focusing on medical and biomedical text, is more compatible with the semi-formal language style of COMETA. Table 4.10 demonstrates the difference in language style between the datasets.

In general, we observe that our distant supervision data could improve the model performance on the UD_i scenario on all existing data (CADEC, PsyTAR, and COMETA). According to these findings, we can conclude that components of Wikipedia articles (article summary, Wikipedia's *redirect* pages, and wikilinks data) contain popularized medical phrases that can be generalized to those stated in social media.

Table 4.8: F1-score comparison between our MCN model trained in overlap concepts with 3 different training set: (1) (UD_i) (2) the DD_i, and (3) the CD_i

Dataset	Hunflair			RoBERTa		
	CD_i	DD_i	UD_i	CD_i	DD_i	UD_i
CADEC ($i = 1$)	67.78	63.50	**78.10**	75.60	66.67	**81.04**
PsyTAR ($i = 2$)	87.93	79.28	**90.92**	89.18	79.44	**91.03**
COMETA ($i = 3$)	67.17	88.69	**92.81**	60.38	61.73	**70.02**

We discovered that our distant supervision method increased the number of medical terms in the available datasets (e.g., COMETA), as shown in Table 4.9. For instance, the medical concept *Crohn's disease of colon* is supported by one phrase (that is, there is one text segment mapped to it), which is *Crohn's colitis*. Our distant supervision method increased the number of medical phrase variations for *Crohn's disease of colon*, to *Crohn's disease, Lesniowski-Crohn disease, Crohn's disease of the esophagus, and granulomatous colitis*.

We must pay attention, though, to issues of topic shifting. We found topic-shifting issues in the previous example; the term *Crohn's disease of the esophagus* is a different disease than *Crohn's disease*. Nevertheless, based on the result, we can argue that our distant supervision has a positive impact on the model performance by increasing the number of popularized medical phrases. This improvement can be observed in the UD_i column result in Table 4.8.

Moreover, we observe that the model trained on DD_1 and DD_2 set performed poorly in comparison to the model trained on CD_1 (CADEC) and CD_2 (PsyTAR). However, the model trained on DD_3

Table 4.9: Three sample medical concepts in CD_3 and DD_3 (COMETA)

Concept	Example of Medical Phrases in Human Annotated Data CD_3	Example of Medical Phrases in Automatically Generated Data DD_3
Backache	*backaches, backpain, back pains, back ache*	*upper back pain, painful back, back problems, bad back, pain in the back*
Myocardial infarction	*myocardial infarction, heart attack, heart attacks*	*attack of heart failure, MI, heart infarction, severe heart attack, cardial infarction*
Crohn's disease of colon	*Crohn's colitis*	*Crohn's disease, Lesniowski-Crohn disease, Crohn's disease of the esophagus, granulomatous colitis*

(COMETA) produces the opposite result, indicating that DD_3 performs better than the model trained on CD_3. We discovered that the original test data of CADEC and PsyTAR appear to use more informal language than the informal medical which is not found in Wikipedia. Meanwhile, the COMETA dataset appears to be more consistent with what is written on Wikipedia.

The examples of these data characteristics can be seen in Table 4.10. We see that the concept *Alcohol intolerance* is supported with medical phrases *intolerance of alcohol* and *alcohol intolerance* in the DD_2 (PsyTAR) train set, meanwhile in the test set, it appears as *no longer to enjoy the occasional glass of wine or champagne b/c it makes me too drowsy* and *got really drunk really fast*. Due to the language gap between the medical phrases in the train and the test data, the model was probably unable to accurately classify the phrase. Furthermore, we discovered that common terms in our distant supervision data set could correspond to one or more SNOMED-CT concepts. For instance, the term *pain* can refer to a number of SNOMED-CT concepts, including *Pain (SNOMED-CT ID:22253000)*, *Abdominal pain (SNOMED-CT ID:21522001)*, *Suffering (SNOMED-CT ID:706873003)*, and *Neuropathic pain (SNOMED-CT ID:247398009)*. Thus we need to filter out ambiguous phrases to avoid injecting noise into the training data and models.

MCN Model Trained on Unified Data In this section, we present the model performance on the MCN model trained on the unified datasets from CADEC, PsyTAR, COMETA, distant supervision data, and synonyms of SNOMED-CT. The evaluation was conducted on the original CADEC, PsyTAR, and COMETA test sets, aiming to assess the model using ground truth data. Recall that the goal of the unified datasets is to broaden the concept coverage of the existing MCN datasets. Table 4.11 shows the MCN model performance based on each trained data explained in Section 5.3.

The model trained on the unified datasets with partial SNOMED-CT synonyms (PART) consistently achieves higher F1-scores across all three testing datasets than the model trained on the unified datasets with all SNOMED-CT synonyms (FULL). In particular, when evaluated with

Table 4.10: Comparison of a sample of medical terms from the provided test set to our distant supervision

Dataset	Concept	Sample Medical phrases in DD_i Train set	Sample Medical phrases in CD_i Test Set
CADEC	Menorrhagia	*Heavy menstruation, Heavy menstrual periods*	*heavy menstral bleeding with clots even though i had just finished my cycle a week before*
PsyTAR	Alcohol intolerance	*intolerance of alcohol, alcohol intolerance*	*no longer to enjoy the occasional glass of wine or champagne b/c it makes me too drowsy, got really drunk really fast*
COMETA	Anhedonia	*Social Anhedonia, lack of pleasure, decreased ability to feel pleasure, anhedonia*	*anhedonia*

Table 4.11: F1-score comparison between MCN models trained with 2 different training sets

Train Data Description	CADEC (CA)	PsyTAR (PS)	COMETA (CO)
CA ∪ PS ∪ CO ∪ DS ∪ FULL	53.57	69.06	**77.75**
CA ∪ PS ∪ CO ∪ DS ∪ PART	**56.35**	**69.57**	77.63

CADEC and PsyTAR, the model demonstrates improvements in F1-Score, achieving 56.35 on CADEC and 69.57 on PsyTAR. Nevertheless, there is a slight decrease in the F1-score when the model is tested on the COMETA dataset, where it achieves a score of 77.63. We observed that both models show a higher F1-score when tested on COMETA in comparison to their performance on the PsyTAR and CADEC test sets. This result is consistent with our previous experiment findings.

We conducted a manual evaluation of the predictions made by the model trained on PART when it was tested on the CADEC dataset. As an example, when presented with the popularized phrase *bad cramps*, the model predicted it to be mapped to *Cramps (SCUI:55300003)*, while the ground truth was *Spasms (SCUI:45352006)*. According to the definition from the National Library of Medicine (NLM) [9], *spasms* refer to any involuntary muscle contraction, while *cramps* denote episodic, involuntary, painful contractions of a muscle. These definitions share a common topic related to involuntary muscle activity, but the interpretation can vary depending on the context in which the terms are used by users.

[9] https://www.ncbi.nlm.nih.gov/books/NBK376/

Furthermore, due to the use of the newest version of SNOMED-CT, some of the concepts in the ground truth are no longer active. For example, for the term *flatulence*, the ground truth is *Flatulence/Wind (SCUI:267052005)* while the model's prediction is *Passing flatus (SUI:249504006)*, which refers to the same context. Another example is that the model sometimes predicts a more general SNOMED-CT concept compared to the specific one in the ground truth. SNOMED-CT concepts are organized in hierarchies. Within a hierarchy, concepts range from the more general to the more detailed[10]. For instance, the popularized phrase *belly weight gain* is mapped to *Weight gain (SCUI:8943002)*, while the ground truth is *Weight increased (SCUI:262286000)*. This also occurs in the PsyTAR dataset, particularly when dealing with popularized medical phrases expressed in very lay phrases. However, as we discussed earlier, the COMETA dataset contains popularized phrases that are not expressed in overly lay phrases.

When considering hierarchical structures like SNOMED-CT, predicting a broader concept rather than a specific one may not always be a misclassification. Instead, it can be seen as a difference in the level of granularity. To evaluate the model's performance accurately in such scenarios, it is crucial to adopt evaluation metrics that account for these hierarchical relationships. One approach is to use hierarchical evaluation metrics that measure how predictions align with the ground truth within the hierarchy. For example, hierarchical precision and recall consider the proximity of the predicted label to the ground truth within the hierarchy [RXA+22]. However, a limitation of our work is that we did not incorporate hierarchical evaluation metrics in our evaluation. Future studies could investigate the use of these metrics to provide a more thorough evaluation of model performance in hierarchical contexts.

Based on our findings, it is clear that the MCN task remains challenging, primarily due to the ambiguity inherent in the way users express their medical conditions. Further analysis is needed to understand the types of misclassifications our model produces beyond what has been discussed so far. Moreover, expert evaluation is required to assess how effectively the unified datasets map popularized medical phrases to SNOMED-CT terms.

4.4 Summary

In this chapter, we proposed methods to leverage Wikipedia as a source of popularized medical phrases to increase the SNOMED-CT concept coverage on three publicly available MCN datasets (CADEC, PsyTAR, and COMETA). We retrieved popularized medical phrases from Wikipedia articles' components (article summary, Wikipedia's redirect pages, and Wikilinks) and paired them with SNOMED-CT concepts.

Our distant supervision approach successfully mapped 9,759 SNOMED-CT concepts from 12,101 Wikipedia articles, of which 1,301 are SNOMED-CT concepts found in the public MCN datasets. The experimental results show that when we combine the data obtained by our distant supervision approach with each of the current MCN datasets, the model performance improves. Based on these findings, we argue that the Wikipedia components contain popularized medical phrases can be used as training data to improve the results for solving the MCN task. However, it is important to note that Wikipedia is a community-driven knowledge source with the potential for data inaccuracy.

We further developed our approach by training the MCN model on unified datasets. These datasets were created by merging publicly available datasets with our distantly supervised data

[10] https://confluence.ihtsdotools.org/display/docstart/4.+snomed+ct+basics

and SNOMED-CT synonyms, and then applying data augmentation techniques. Our analysis of the MCN model trained on unified datasets, specifically the partial (PART) synonym set, outperforms all the test datasets, including CADEC, COMETA, and PsyTAR than the model trained on unified datasets with all SNOMED-CT synonyms (FULL). Our findings indicate that the model excels when handling popularized medical phrases that are not expressed in overly lay terms.

In addition, manual investigation uncovered challenges related to the granularity levels in the MCN task. The model often struggles to normalize popularized medical terms to the appropriate level of specificity within the SNOMED-CT concept hierarchy, where the model tend to map the terms to broader concepts rather than more specific ones. This highlights the complexity of the MCN task, especially in managing concept granularity.

Tasks for future work are applying specific filtering and disambiguation processes to reduce interference from unrelated or common phrases. Additionally, we plan to conduct a comprehensive study to better understand and address the challenges posed by ambiguous popularized medical terms, particularly those that are sensitive to different levels of granularity within the SNOMED-CT hierarchy.

Furthermore, we investigate the potential of addressing the MCN task as a hierarchical classification problem rather than a multi-class classification, and incorporate hierarchical evaluation metrics to provide a more thorough evaluation of model performance in hierarchical contexts. Lastly, we are considering expert evaluation to assess how effectively our MCN model can classify popularized phrases into the correct SNOMED-CT concepts.

CHAPTER 5

Informal Medical Entity Linking

In this chapter, we introduce a model for informal medical entity linking (EL) aimed at helping laypeople understand medical terms. The model is built upon the foundations laid in Chapters 3 and 4, where we created models for Named Entity Recognition (NER), which is used to detect popularized medical phrases within user text, and Medical Concept Normalization (MCN), which normalizes medical concepts with the outputs from the NER model. Furthermore, we introduce a dedicated Entity Disambiguation (ED) model in this chapter. This model is responsible for selecting the most appropriate explanation of a concept, specifically from Wikipedia articles, drawing upon the results from the MCN. All these components build, together, an *Informal*

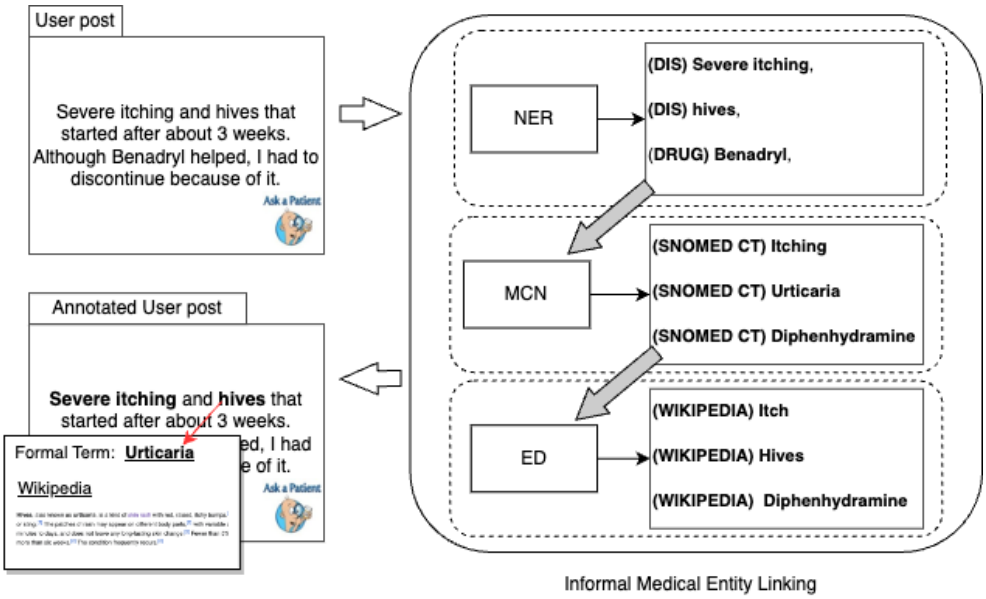


Figure 5.1: Overview of the informal medical entity linking workflow.

Medical Entity Linking model. Figure 5.1 gives a visual representation of the workflow of the model.

The contributions of this chapter are:

- We have developed an informal medical EL model designed to support laypeople in learning specialized medical terminology in social media settings. In addition, we created the interface to showcase how the model works.
- We conducted an evaluation by experts to assess the effectiveness of our informal medical entity linking model, focusing specifically on the MCN and Entity Disambiguation (ED) modules. This was done to determine whether the outputs from these modules are effective in helping laypeople understand medical terminology.

5.1 Informal Medical Entity Linking Model

In this section, we provide an overview of the Informal Medical Entity Linking (EL). As mentioned earlier, Figure 5.1 illustrates the components and workflow of Informal Medical EL model. The model consists of three phases: (1) *mention detection*, which involves identifying textual mentions that represent popularized medical phrases referring to a disease or a drug (referred to as the NER task (Section 3.1.3); (2) Mapping these identified mentions to corresponding concepts in SNOMED-CT (referred to as the Medical Concept Normalization (MCN) task); and (3) linking each of these SNOMED-CT concepts to the corresponding entities in Wikipedia to make it easier for laypeople to understand the specialized medical term (referred to as Entity Disambiguation (ED) task).

We denote with V the vocabulary of tokens in a social media post[1] A text sequence $\mathbf{t}$, consists of sequentially ordered tokens from V, forming a sentence in the original text. We write it as $\mathbf{t} = (t_i)_{i=1}^{n}$, where n is the length of the token sequence. From a token sequence, $\mathbf{t}$ we can extract a set of phrases (set of consequent tokens, also called *spans*), $\mathbf{p}(\mathbf{t})$:

$$\mathbf{p}(\mathbf{t}) = \{(p_j)_{j=1}^{m} | p_j \in \mathcal{L}(\mathbf{t})\}$$

where $\mathcal{L}(\mathbf{t})$ denotes the set of candidate spans over $\mathbf{t}$.

> **Example**: consider the user text "*Severe itching and hives that started after about 3 weeks. Although Benadryl helped, I had to discontinue because of it.*"

For example, given a sample user text, we extract two sequences (as it contains two sentences), one of which is $\mathbf{t}$ = (Severe, itching, and, hives, that, ..., weeks). Then, $\mathcal{L}(\mathbf{t})$ includes possible candidate spans such as {(Severe, itching), (Severe, itching, and), (Severe, itching, and, hives), (hives), ...}.

[1]Tokens are extracted user text posts by the word tokenization.

Phase 1: Named Entity Recognition Let $T = \{\mathbf{t}\}$ be the set of all sentences or text sequences in the vocabulary V, and let P be the list of identified popularized medical phrases. The Named Entity Recognition (NER) process operates on T and returns some popularized medical phrases P from $\mathcal{L}(T)$. This process is represented as $\text{NER} : T \rightarrow \mathcal{L}(T)$.

The process of identifying textual mentions in social media is challenging. One of the reasons is that people can express medical concepts in various ways. For instance, the term *Vertigo (SNOMED-CT Code: 399153001)* could be described by *head spinning*, *waves like vertigo but whilst sitting*, or *vertigo like attacks*. Given the variation in expression, to identify the popularized medical phrases, we trained a NER model using a deep learning technique for sequence labeling. Building upon our previous work in [NHPA21], we use a BILSTM-CRF architecture, trained on the CADEC [KMJKW15] and the MedRed [SMLQB20] datasets, to create a larger training dataset (see Section 3.3).

> **Example**: Considering T from the earlier example, the NER identifies the following set of popularized medical phrases $P = \{$(severe, itching), (hives), (benadryl)$\}$

Phase 2: Medical Concept Normalization This phase involves mapping each identified popularized medical phrase p from the set P, derived from the NER phase, to a corresponding specialized medical concept in SNOMED-CT. Let C represent the set of specialized medical concepts in SNOMED-CT. The Medical Concept Normalization (MCN) model performs this mapping, represented as $MCN : P \rightarrow C$, with:

$$c_p = MCN(p) \quad \forall p \in P$$

Here, c_p denotes the specialized medical term in SNOMED-CT associated with the popularized phrase p.

> **Example**: The phrase *(severe, itching)* is mapped to the SNOMED-CT code: 418290006, representing *Itching*.

We trained the MCN model using supervised learning by treating it as a multi-class classification task. The model architecture employed a gated recurrent unit (GRU). The training data was a combination of existing datasets such as CADEC, PsyTAR, and COMETA, as well as automatically labeled data generated by the approach proposed in Chapter 4. We refer to the auto-labeled data as *distant supervision data*. To reduce the noise in our auto-labeled data, we applied certain filtering processes. We also enriched the training dataset by adding the SNOMED-CT synonyms (see Section 4.3, the best-performing model was used).

Phase 3: Entity Disambiguation In the Entity Disambiguation (ED) phase, we employed the GENRE model [DIRP21]. Here is an explanation of how the GENRE model works. GENRE takes an input text and generates a Wikipedia entity name in an autoregressive manner. It employs a Beam Search algorithm within a prefix tree structure, where each node represents tokens from the vocabulary, primarily Wikipedia titles [DIRP21]. This structure allows GENRE to propose potential Wikipedia entities[2] based on the input sequence. It then connects these proposed entities to actual Wikipedia entities by using a scoring and ranking mechanism. This

[2]Wikipedia entities refer to the article titles on Wikipedia

mechanism evaluates the likelihood of each proposed entity name being a valid Wikipedia entity, considering the context provided by the input text.

For the GENRE model to function effectively, the input must include entities or concepts enclosed within special tokens [START] and [END]. This approach allows GENRE to treat the encapsulated mention as a query for retrieving and linking to relevant Wikipedia entities.

During the Entity Disambiguation (ED) phase, GENRE operates on two distinct inputs: popularized medical phrase, denoted as p, and its specialized medical term, denoted as c_p.

1. **Tagging Popularized Medical Phrases** (p): The process begins with the original text, where the popularized medical phrase (p) is identified. This phrase is tagged with placeholders, marking the start and end of the tagging. This step is formalized as:

$$E_p = \text{GENRE}(\text{original text with [start]} + p + \text{[end] tagging})$$

   Here, E_p represents the set of top five Wikipedia entities identified for the tagged text based on the popularized medical phrase.

2. **Tagging with Specialized Medical Terms** (c_p): In the subsequent step, the text modified with placeholders for the popularized phrase (p) is further processed to tag these placeholders with the specialized medical term (c_p), also marked by placeholders to highlight the tagging. This process is formalized as:

$$E_{c_p} = \text{GENRE}(\text{original text with [start]} + c_p + \text{[end] tagging})$$

   E_{c_p} denotes the set of top five Wikipedia entities identified for the text based on the specialized medical term.

Considering the two different types of inputs, this could lead to overlapping entities, resulting in a total of up to ten unique entities. The outputs from the ED module consist of two separate sets of Wikipedia entities: $E_p = \{e_{p1}, \ldots, e_{p5}\}$ for the popularized phrases and $E_{c_p} = \{e_{c1}, \ldots, e_{c5}\}$ for the specialized medical terms. Within each set, the entities are ordered by their relevance as determined by the GENRE score, with e_1 being considered the most relevant entity for its respective input.

Example: Consider text = *Severe itching and hives that started after about 3 weeks.* and the popularized medical phrase p is *(severe, itching)*.

$$\begin{aligned} E_p &= \text{GENRE}(\text{[START]Severe itching[END] and hives that started} \ldots 3 \text{ weeks.}) \\ &= \{\text{Itch, Anorexia (symptom), Arthralgia, Allergic rhinitis, Erythema}\} \end{aligned}$$

urthermore, with the specialized medical term, c_p, *itching*:

$$\begin{aligned} E_{c_p} &= \text{GENRE}(\text{[START]Itching[END] and hives that started} \ldots 3 \text{ weeks.}) \\ &= \{\text{Itch, Allergic rhinitis, Arthralgia, Herpes labialis, Infectious mononucleosis}\} \end{aligned}$$

To further refine the entity selection, a learning-to-rank model, denoted as *Rank*, is introduced. This model performed a re-ranking of entities based on their relevance to popularized and

specialized medical inputs. The re-ranking process for each medical concept c_p combines the entities identified for the popularized phrase (E_p) and its corresponding specialized medical term (E_{c_p}). The details of this re-ranking model are explained later in the chapter. Formally, this process is represented as:

$$R_{c_p} = Rank(E_p \cup E_{c_p})$$

Here, R_{c_p} signifies the ranked output for a given concept c_p, obtained by applying the *Rank* model to the union of the entity sets from both E_p and E_{c_p}. This approach ensures that the most relevant entities for both popularized and specialized medical terms are identified and prioritized effectively.

> **Example**: The output of the $R_{c=itching}$ is re-ranked Wikipedia entities as follows: (e_1) *Itch*, (e_2) *Allergic rhinitis*, (e_3) *Herpes labialis*, ($e4$) *Arthralgia*, (e_5) *Erythema*, (e_6) *Anorexia (symptom)*, and (e_7) *Infectious mononucleosis*. Recall, that the first rank is assumed as the most relevant entity given to the $c_{itching}$.

The informal medical entity linking model produces its final output as a structured set of tuples. Each tuple in this set comprises three distinct elements: (1) a predicted informal medical phrase from the text, (2) its normalized counterpart in SNOMED-CT, and (3) a relevant Wikipedia entity. These elements are formalized as (p, c, e_1), where p denotes the popularized medical phrase, c represents the corresponding specialized medical term in SNOMED-CT, and e_1 is the most relevant Wikipedia entity associated with the specialized term. This tuple structure aids in enhancing the comprehensibility of medical terminology for laypeople, as depicted in Figure 5.1.

> **Example**: In the context of the informal medical Entity Linking (EL) process, consider a tuple formatted as (p, c, e_1). For the popularized phrase p = *severe itching*, the corresponding tuple would be (p = *severe itching*, c = *itching*, e_1 = *"Itch"*). Here, p is the popularized medical phrase found in the text, c is its specialized medical term normalization in SNOMED-CT, and e_1 is the associated Wikipedia entity, which comprises the title, url, and summary that provides further understanding of the specialized medical term.

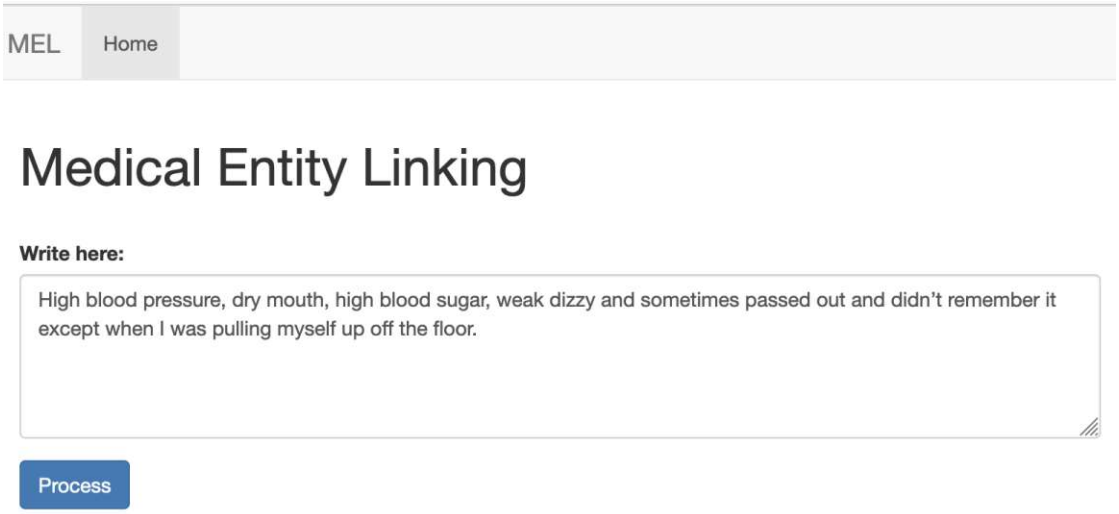
5.2 Informal Medical EL Interface Design

We have designed and implemented a simple web-based application to display the outputs of our informal medical Entity Linking (EL)[3]. The main objective of this application is to enable testing and evaluation of the model performance. The system has been designed and implemented to facilitate an expert evaluation, which will be discussed in the next section. The system displays the EL model's output from the ED Phase up to the ED-GENRE model's results, allowing users to see the top-ranked Wikipedia articles linked to the identified medical entities. Note that our learning-to-rank model is not included for the expert evaluation.

The interface includes a text area where users can insert content, such as a post from health forums. The user text will be processed by the EL model and the output of this process will be shown to the user in a text display field where detected entities are highlighted and linked to wiki entries.

[3]https://gitlab.com/annisamaulidaningtyas/medlingo.git

Figure 5.2a presents a sample post from AskAPatient, discussing a user's experience with the adverse effects of the *Antihistamine* drug. Figure 5.2b displays this post after processing through the informal medical EL model. To view detailed information in Figure 5.2b, a user should click on the highlighted terms. For example, when the popularized medical phrase *dry mouth*



(a) A user post from AskAPatient.com

(b) Processed post with highlighted popularized medical phrases along with detailed descriptions of their corresponding specialized medical concepts and Wikipedia candidates.

Figure 5.2: Interface Design for Informal Medical EL

is selected in the figure, it shows its corresponding SNOMED-CT code *87715008*, which is described as the specialized term *Xerostomia*, and a list of relevant Wikipedia entities, such as *Xerostomia*, *Laryngitis*, *Hemorrhoid*, *Laryngoscopy*, and *Laryngospasm*, generated through the Entity Disambiguation (ED) phase. A feature we implemented in the interface is a slide bar where users can set the prediction confidence level threshold for the outputs generated by the medical concept normalization (MCN) model. The slider determines how the annotations are visually represented based on the model's confidence scores.

When the MCN model predicts an annotation with a confidence level higher than the threshold

set by the user, the corresponding term is highlighted in green. On the other hand, if the prediction score falls below the specified confidence level, the term is highlighted in yellow. For example, if the confidence level is set to 90%, any term with a confidence score below 90% will be highlighted in yellow. An example of this would be the term *high blood pressure* appearing in yellow, indicating that the model's confidence in its prediction is below the 90% threshold. Notice that our informal medical EL model does not change the original post. It only provided the additional information used to support laypeople in learning medical terminology.

5.3 Evaluation of Informal Medical EL Model

In this section, we present how we conducted the expert evaluation to verify the output of the end-to-end informal medical EL model, particularly focusing on the outputs from the Medical Concept Normalization (MCN) and Entity Disambiguation modules. The objective of this evaluation was to demonstrate that the informal medical entity linking model is capable of generating reliable information that can be used by laypeople to empower them in learning medical terminology. It's important to note that we did not evaluate the overall performance of the informal medical EL model against a standard dataset (ground truth data) in this study. However, we did conduct the model performance evaluation of NER and MCN model performance in Chapters 3 and 4.

In addition, we faced the challenge that we could not automatically validate the performance of the output of the GENRE model used in the ED phase. To our knowledge, there exists no ground truth data for the ED task that is suitable for our research purposes. Moreover, we had a similar problem in finding suitable available data to perform the model performance evaluation for the complete pipeline of the informal medical EL model. Therefore, we conducted an expert evaluation to determine the accuracy of the MCN modules in mapping popularized phrases to specialized medical terms and whether the ED module could locate the appropriate Wikipedia entities or articles based on the predicted specialized medical terms.

Initially, we assumed that the appropriate explanation for specialized medical terminology would be the highest-ranked Wikipedia article provided by the GENRE model used in the ED module. However, our manual review revealed that the most relevant article is not always the top-ranked one. We address this challenge through expert evaluation. We did not perform an evaluation for the output of the Named Entity Recognition (NER) model. This decision was based on the findings from our failure analysis in Chapter 3, which indicated that the predictions of popularized medical phrases by the NER model were reliable enough to be used by the MCN model to predict specialized medical terms. However, we acknowledge that this decision may have overlooked potential errors in NER tasks that could affect overall performance. We've noted this as an area for future research. The details of this expert evaluation are explained below.

5.3.1 Participants

Initially, our goal was to recruit a medical expert who is a general practitioner and familiar with SNOMED-CT, and who is also proficient in English to minimize language biases. However, finding experts is a challenge due to the limited time and financial resources. Therefore, we only have one inclusion criteria to choose the experts: they must be medical professionals. We worked with three experts, each with distinct medical specializations: a general practitioner, a midwife, and a medical coder, all of whom are our colleagues from Indonesia.

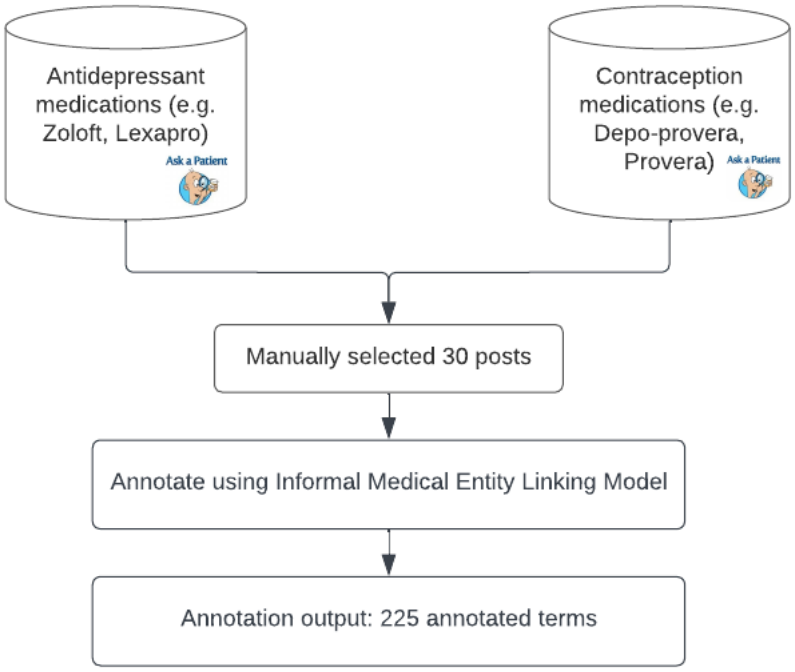


Figure 5.3: The diagram illustrates the process of selecting data from two categories of drug reviews on the AskAPatient forum - *antidepressants* and *contraceptives*. Thirty user posts were manually selected and annotated using our informal medical entity linking model to create an evaluation dataset. In total, 225 annotated terms were identified by the model in the selected posts.

5.3.2 Data Selection for the Expert Evaluation

Figure 5.3 illustrates the data selection process used in the expert evaluation. We manually selected 30 posts from AskAPatient.com. These posts were specifically selected from two categories: *antidepressant medication* and *contraception medication*. We included *antidepressant medication* posts because most of the existing datasets primarily focus on reviews of antidepressant drugs. Additionally, we included *contraception medication* posts to assess the performance of our pipeline in the context of different types of drug reviews. Once we selected these user posts, we annotated each of them using our informal medical EL model pipeline. As a result of this annotation process, the model was able to extract 225 annotated terms. These annotated terms are the predicted popularized medical phrases along with their corresponding specialized medical terms and associated Wikipedia candidates.

Figure 5.4 presents a visualization of a user post annotated by our informal medical entity linking model, specifically for expert evaluation. In this visualization, the highlighted terms represent the popularized medical phrases identified during the Named Entity Recognition (NER) phase. These phrases are then mapped to their corresponding specialized medical terms, the output of the Medical Concept Normalization (MCN) phase, along with Wikipedia candidates identified

in the Entity Disambiguation (ED) phase. We modified the ED phase process to align with the objectives of the expert evaluation, focusing specifically on the outputs from the GENRE model. As previously mentioned in Section 5.1, we employed two types of inputs, popularized medical phrase (E_p) and specialized medical term (E_{c_p}), to retrieve the top five Wikipedia candidates for each input. This approach allowed a thorough analysis of how well GENRE accurately retrieved Wikipedia candidates that are most relevant to both popularized medical phrases and their specialized medical terms.

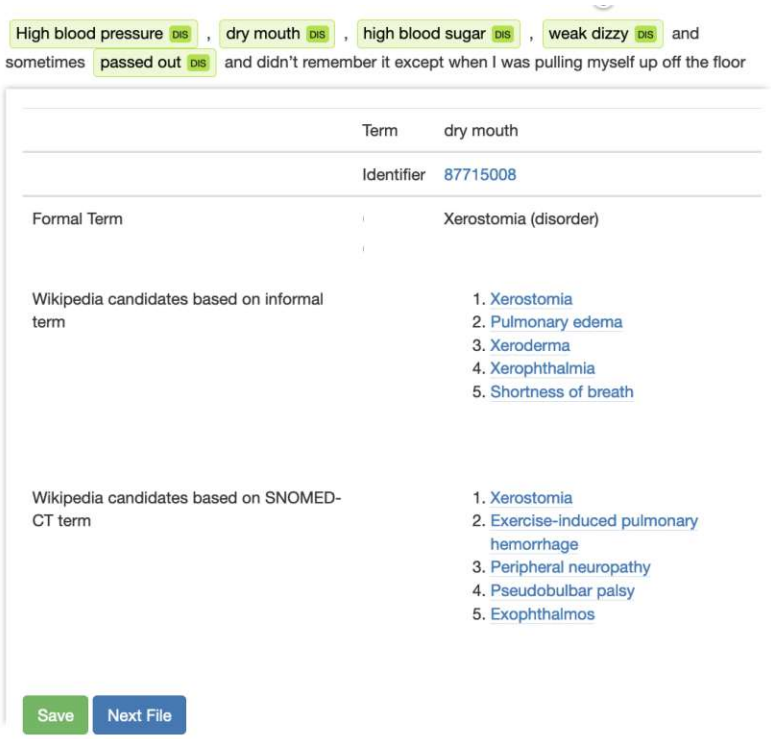


Figure 5.4: The figure shows an annotated sentence from a user post, where popularized medical phrases are highlighted alongside their specialized medical terms and Wikipedia candidates. In this example, clicking on the predicted popularized medical phrase *dry mouth* displays detailed information. This information is organized into five rows. The first row, labeled *Term*, shows the popularized phrase itself, *dry mouth*. The second row, *Identifier*, presents the SNOMED-CT code associated with *dry mouth*. In the third row, *specialized Term*,the specialized medical term derived from this identifier is listed; in this case, *Xerostomia* is the specialized term for *dry mouth*. The fourth row details Wikipedia candidates that the GENRE model retrieved using the popularized medical phrase *dry mouth* as input. Finally, the fifth row displays Wikipedia candidates retrieved by the GENRE model using the specialized medical term *Xerostomia* as input.

5.3.3 Expert Evaluation Design

The evaluation of the informal medical entity linking (EL) model involved three experts. Each expert annotated a set of 30 posts, which collectively contained 225 terms identified by the named entity recognition (NER) model. The medical concept normalization (MCN) model provided the corresponding specialized medical terms for these entities, while the entity disambiguation (ED) module, utilizing the GENRE model, retrieved relevant Wikipedia candidates. The experts were tasked with evaluating two aspects:

- The correctness of the specialized medical terms predicted by the MCN model.
- The appropriateness of the Wikipedia articles retrieved by the ED module, which were associated with both the informal medical phrase and specialized medical terms.

Figure 5.5 illustrates the user interface of the expert evaluation system employed in this process.

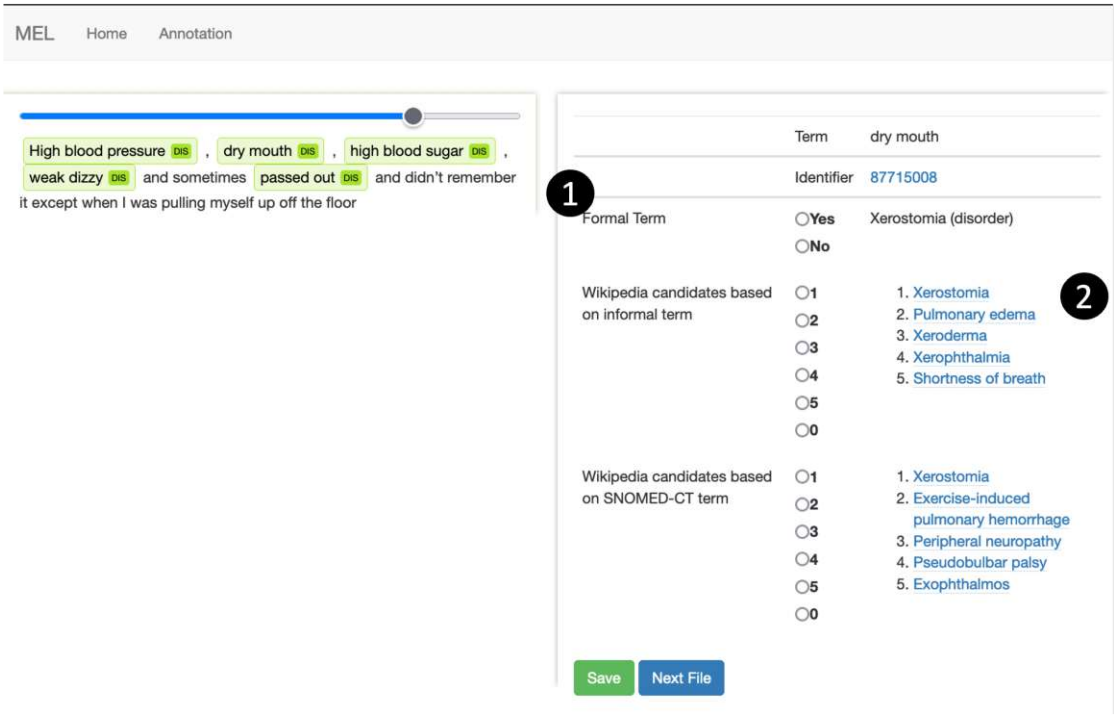


Figure 5.5: Expert Evaluation System

Task 1: Correctness of Predicted Specialized Medical Terms

Verify the correctness of the MCN model output (the box marked with 1): Indicate whether the displayed specialized term shown in “Formal term“ row is a correct mapping to the popularized phrase by selecting “Yes” or “No” accordingly.

Task 2: Selection of Appropriate Wikipedia Articles

The second part of the evaluation focused on the selection of relevant Wikipedia articles. The articles selected by the GENRE model needed to effectively explain both popularized phrases and specialized medical terms, aiming to enhance laypeople's understanding of medical terminology. Experts were required to choose the appropriate Wikipedia based on their relevance to the terms: *popularized medical phrases* and *specialized medical terms* by marking the most relevant Wikipedia article from a list of the top 5 ranked articles, given by the GENRE model, for both popularized phrases and specialized medical terms. When no Wikipedia articles were considered relevant, the experts selected the option "0". For instance, if the most suitable article for a popularized phrase like "dry mouth" (formally, *Xerostomia*) was listed first, experts would select the option *1*. This process corresponds to the second objective of our expert evaluation: determining the suitability of Wikipedia articles. Detailed instructions and guidelines were provided to the experts for this evaluation (see Appendix 7.3.3).

5.3.4 Evaluation Metrics

This section presents the metrics used to assess the outcomes of the expert evaluation. To evaluate the agreement between the experts on each evaluation task, we use the *Inter-Annotator Agreement* metric. For assessing the accuracy of the MCN model in Task 1, we used the *Accuracy* metric. Finally, to determine the effectiveness of the ED module in retrieving relevant Wikipedia articles for Task 2, we utilized the *P@1 (Precision at 1)* and *MAP (Mean Average Precision)* metrics.

Inter-Annotator Agreement To evaluate how consistently different annotators understood and completed the evaluation tasks, we computed Cohen's Kappa. This metric, also termed an Inter-Annotator Agreement (IAA), measures the extent to which annotators agree on the labels given for the items, beyond what would be expected by chance [Coh60]. Cohen's Kappa is calculated as:

$$Cohen'sKappa = \frac{P_o - P_e}{1 - P_e}$$

Here, P_o is the proportion of observed agreement among annotators, and P_e is the expected agreement by chance.

Accuracy To evaluate the accuracy of the mapping of popularized medical phrases to their corresponding specialized medical terminology, this research specifically defines *accuracy* as the ratio of popularized phrases that were accurately associated with the correct specialized medical terms to the total number of mappings assessed by the annotators. This metric was determined based on annotations provided by expert annotators. The formula for calculating accuracy is as follows:

$$accuracy = \frac{number\ of\ correctly\ mapped\ phrases}{total\ number\ of\ phrases\ evaluated}$$

Specifically, the *number of correctly mapped phrases* is the count of terms for which a majority of annotators agreed upon a correct popularized-to-specialized medical term corresponds. The *total number of phrases evaluated* refers to the overall number of terms reviewed by the annotators.

P@1 and Mean Average Precision (MAP) The effectiveness of the Expert Disambiguation (ED) module in retrieving relevant Wikipedia articles was evaluated as an Information Retrieval problem. The relevance of the top-ranked Wikipedia article to the query was assessed using $P@1$ (Precision at 1). In this context, a query refers to the input phrases provided to GENRE, which include both *popularized medical phrases* and *specialized medical terms*.

$$P@1 = \frac{\textit{Number of relevant articles in top 1}}{\textit{Total number of articles in top 1}}$$

Additionally, we used Mean Average Precision (MAP) to assess the overall accuracy of retrieved articles across multiple queries.

$$MAP = \frac{1}{N}\sum_{i=1}^{N} \mathrm{AP}_i$$

Here, AP_i represents the Average Precision for a single query, and N is the total number of queries.

5.4 Results and Discussion

This section presents the results of our evaluation, structured into three main parts: (1) We assessed the agreement among experts on two tasks by calculating the mean Inter-Annotator Agreement (IAA). These tasks were: (a) evaluating the accuracy of specialized medical term predictions, and (b) selecting the most relevant Wikipedia articles as retrieved by the Entity Disambiguation (ED) module. (2) We measured the precision with which the Medical Concept Normalization (MCN) model normalized popularized medical phrases to specialized medical terms. (3) Lastly, we evaluated the performance of the ED model's article retrieval function, using Precision at 1 (P@1) and Mean Average Precision (MAP) metrics.

5.4.1 Expert Agreement

We distributed all 30 posts, containing a total of 225 evaluated terms, to each of the experts for assessment. Table 5.1 presents the average of *Cohen's Kappa* score with standard deviation, minimum and maximum score of the pair-wise agreement.

Table 5.1: The pairwise agreement between experts

Description	Cohen's Kappa Score		
	Avg. Pairwise	Min	Max
MCN Output Correctness	0.37 ± 0.07	0.32	0.46
Wikipedia Selection (Specialized Term)	0.51 ± 0.01	0.50	0.52
Wikipedia Selection (Popularized Term)	0.53 ± 0.01	0.52	0.55

Based on the results, we noticed that the agreement of verifying the correctness of the mapping between popularized to specialized medical terms as the MCN output is *fair*. This result highlights that verifying the correctness of predicted specialized medical terms from popularized terms is a challenging task. To gain further insights, we, at the end of the evaluation, re-discussed the matter with the experts, to understand the challenges they faced during this evaluation process.

One of the experts highlighted the different handling of popularized medical phrases indicating varying degrees of disease severity, like 'mild' or 'severe'. In addition, the verification of the normalization task is very subjective, a finding that aligns with previous research [KMJKW15]. For example, the MCN model mapped the popularized phrase of *feeling useless* to *feeling hopeless (SCUI:307077003)*. However, it's difficult for the experts to decide whether it's correct or not because the term *feeling useless* is general: it can lead to the *Depression mood (SCUI: 366979004)*. However, the experts did not find problems when the informal medical entity linking is most likely or the same with the specialized medical terms, for example *terrible pain in feet* mapped to *Foot pain (SCUI:47933007)*, or *insomnia* mapped to *Insomnia (SCUI:193462001)*.

Furthermore, as discussed in Section 5.3.2, we processed two distinct input queries through the ED model. These included the output from the NER model, indicating the popularized medical term, and the output from the MCN model, representing the specialized medical term. Furthermore, the task of selecting a suitable Wikipedia article as a source of explanation for a specific medical term demonstrated *moderate* agreement for both types of input queries provided to the ED model. These scores are higher than the IAA for the other evaluated tasks. We interpret these results as an indication that the task of selecting appropriate Wikipedia articles for both popularized and specialized medical terms is more straightforward than the first task. The access to information from Wikipedia likely provides experts with valuable assistance in their selection process, leading to higher levels of agreement observed.

We examined the overall level of agreement among all the experts for each task. The agreement summaries are presented in Table 5.2. For Task 1, only 2 out of the experts agreed on 39 out of the 225 annotated terms (16.9%). For Task 2, which involved selecting Wikipedia articles for both popularized and specialized medical terms, there was no agreement among all experts for 7 out of the 225 annotated terms (3.1%). However, 2 experts agreed on 55 out of the 225 terms (24.4%). Although most experts agreed on the same label, the inter-annotator agreement (IAA) score might be lower due to chance agreement, especially in Task 1. The possible reasons for low IAA scores are annotation guidelines, personal background knowledge, and personal assumptions [APW18]. The guidelines regarding correctness criteria may have caused different interpretations by the experts, as there was a lack of ground truth annotation examples to reference. Furthermore, our experts come from different medical backgrounds, which may have affected their responses. Additionally, varying degrees of disease severity found in the annotated terms may result in inconsistent interpretations among the experts. We acknowledge the importance of further analysis to gain a deeper understanding of the complexity of each task, enabling us to refine the tasks or annotation guidelines to increase the agreement levels for both tasks.

5.4.2 Model Performance Analysis Based on the Annotation

We present the evaluation of model performance analysis based on the evaluation tasks done by experts. As discussed in Section 5.4.1 we used the 225 annotated terms for the analysis. These terms represent the predicted popularized medical phrases derived from the Named Entity Recognition (NER) output. The following explains the details of this evaluation, and we report the aggregate results derived from all the annotated terms.

MCN Model Accuracy To assess the accuracy of the MCN model in correctly classifying popularized medical phrases into specialized medical terms, we gathered data annotated by experts. For each popularized phrase, we used majority voting to determine whether the MCN model accurately predicted the correct specialized term. We used the accuracy metrics explained

Table 5.2: Agreement summaries based on the experts response in evaluation tasks

Description	Label	all different agreements	2 experts in agreement	all experts in agreement
MCN Output Correctness	Y	0.0	11.1	78.2
	N	0.0	5.8	4.9
	Overall	0.0	16.9	83.1
Wikipedia Selection (Popularized Term)	0	0.9	8.0	6.7
	1	0.9	8.9	57.3
	2	0.4	4.0	4.9
	3	0.0	1.8	2.2
	4	0.4	1.8	0.4
	5	0.4	0.0	0.9
	Overall	3.1	24.5	72.4
Wikipedia Selection (Specialized Term)	0	0.9	9.8	11.1
	1	0.9	12.0	54.7
	2	0.0	0.4	1.8
	3	1.3	1.8	1.3
	4	0.0	0.4	2.2
	5	0.0	0.0	1.3
	Overall	3.1	24.4	72.4

Table 5.3: Example of incorrect MCN model output

Popularized Term	Incorrect Specialized Term	SNOMED-CT Code (SCUI)
poor bowel movement	Infectious diarrheal disease	19213003
want to cry all the time	Hypersomnia	77692006

in Section 5.3.4. Our results show that the MCN model accurately classified 89.3% (201 out of 225 terms) of the popularized medical terms into their specialized terms.

Table 5.3 presents examples where the model incorrectly classified popularized terms, as identified by expert evaluations. From the example, *poor bowel movement* is a very broad term and could mean many different bowel problems. It is not clear if it means *constipation* or *diarrhea*[4]. The predicted specialized term is *Infectious diarrheal disease (SCUI:19213003)* which has the more specific meaning of diarrheal disease caused by viruses[5]. While it correctly identifies the topic (bowel problems), it inaccurately narrows down the broad popularized term to a specific type of diarrheal disease. Furthermore, the terms *want to cry all the time* can be interpreted as a constant feeling of sadness, however, the model classified it as *Hypersomnia (SCUI:77692006)* (a person feels excessively tired during the day)[6]. This misclassification is considered to be a significant error, as the two conditions are not even closely related.

In cases where a popularized term refers to a particular health issue, the model is capable of correctly identifying the corresponding specialized medical terminology. For instance, the phrase *chest tightness* is accurately associated with the specialized term *tight chest (SCUI:23924001)*. Similarly, popularized terms are linguistically close to their specialized terms, for example, the phrase nausea is accurately associated with the specialized medical term *Nausea (SCUI:422587007)*, making it easier for the model to make a correct prediction. Based on these findings, we can argue that the MCN model has the capability to accurately classify popularized terms into their correct specialized medical terms. However, we acknowledge that the model faces challenges when dealing with highly popularized or colloquial medical phrases.

ED Model Evaluation Similarly, with the MCN model evaluation, we construct the data to evaluate the ED model performance by considering the majority voting of the post assigned by the experts. In cases of a tie, we assign a value of *0*, indicating that there is no relevant Wikipedia article. For each query input (both popularized and specialized terms), the top-5 retrieved documents were taken into account. In this evaluation process, there was only one relevant document identified per query. This was because the experts were asked to select the most appropriate Wikipedia article that best explained a medical concept, corresponding to the pair of popularized-specialized terms across the retrieved results (see Section 5.3.3. We then calculated the P@1 and MAP score (see Section 5.3.4). As previously discussed, we assume the first-ranked article is the most appropriate article as a source of explanation for a specific medical concept.

Table 5.4 shows the proportion of the relevant documents ranked first and the accuracy of the model. The results show that the ED model where we took specialized medical terms as the input has a P@1 score of 0.67, slightly higher than the P@1 score for the model that took popularized

[4]https://medlineplus.gov/bowelmovement.html
[5]https://www.cdc.gov/disasters/disease/infectevac.html
[6]https://www.ninds.nih.gov/health-information/disorders/hypersomnia

medical phrases as input. Additionally, the Mean Average Precision (MAP) score is higher when the model uses popularized medical phrases as queries, at 0.73, compared to 0.7 when using specialized medical terms. In assessing the correctness of the MCN prediction results, if experts choose 'No', indicating an incorrect prediction, the option for Wikipedia selection with the specialized term input is automatically set to *0*, we assume that if the specialized terms are incorrect, it becomes challenging to find a suitable Wikipedia article that aligns with both the popularized and specialized terms. We assume this adjustment lowers the accuracy of the ED model, resulting in a lower MAP score for specialized terms compared to popularized terms.

Table 5.4: ED model performance based on the experts evaluation

Query Input	$P@1$	MAP
Specialized Medical Terms	0.67	0.70
Popularized Medical Terms	0.66	0.73

We are interested in having the relevant candidate from Wikipedia articles at the top of the list. The P@1 shows that 66% of the candidates in the top-ranked. To improve this, we used the information collected from the annotation to train a re-ranking algorithm, referred to as Learning-To-Rank (LeToR) from the output of GENRE-ED module which is presented in the following section.

5.4.3 Learning-To-Rank for Reranking the ED Module

We formulate the re-ranking problem as the LeToR task. To develop a LeToR model for re-ranking Wikipedia candidate articles, we follow two key steps: (1) determining the relevance score between the query (the input to the GENRE-ED module, including both popularized and specialized terms, and the user posts), and documents (union set of Wikipedia articles output by the GENRE-ED module), and (2) performing feature engineering. Once these two steps are completed we apply various algorithms from each LeToR category from the RankLib library[7].

Making the dataset LeToR is designed to re-rank the previously retrieved Wikipedia articles to select the most relevant article to serve as the explanation source for a specific medical concept. Since the process is supervised, we need to define a ground truth that presents how relevant each retrieved Wikipedia article is to the popularized-specialized medical terms. To build the LeToR model, we filtered the expert-annotated data to only include popularized terms that were correctly mapped to specialized medical concepts based on the majority votes of the experts in Task 1. We then used these 201 correctly mapped terms to train the LeToR model for re-ranking Wikipedia articles retrieved in Task 2 (see Section 5.4.2, with a 75/25 split for training and testing data respectively. The relevance of each Wikipedia article was determined by counting the number of times it was selected by the experts for each query (including both popularized and specialized terms). This count served as the relevance score for each article. If a Wikipedia article appeared in both the popularized and specialized term sets, we used the higher count from either set as the final relevance score for that article. The detailed process is shown in Figure 5.6.

[7]https://sourceforge.net/p/lemur/wiki/RankLib

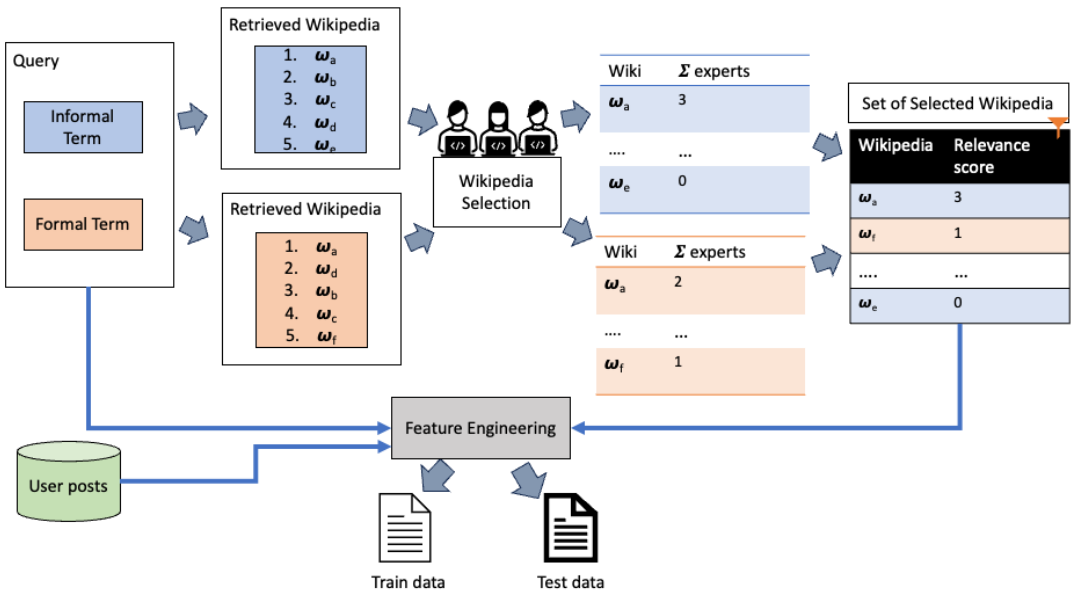


Figure 5.6: A process of data creation used to build a LeToR model

Figure 5.6 illustrates the process of creating a dataset for training the Learning to Rank (LeToR) model. We used the output of Task 2 in Section 5.3.3. The relevance of each Wikipedia article was determined by counting the number of times it was selected by the experts. This count was used as the relevance score for each article. To create a final dataset, we pooled the sets of Wikipedia articles from both query-Wikipedia pairs. In cases where the same Wikipedia article was found in both sets, we took the maximum count from either set as the final relevance score for that article. Moreover, we have also developed a Wikipedia corpus as part of the LeToR dataset. This corpus consists of a collection of the introduction paragraph of all the Wikipedia articles retrieved by the ED model in response to the given queries.

Feature Engineering The effectiveness of a LeToR model significantly depends on the quality of its feature engineering. Our approach incorporates features adapted from the Microsoft LeToR dataset [HL18] and is categorized into two levels: query-level and interaction-level features.

Query-level features are derived directly from input queries of ED model, namely popularized and specialized. We categorize queries into three types: popularized terms, specialized terms, and user posts. The features at this level include:

1. **Covered Query Term Number**: This feature calculates the count of words in query terms that found Wikipedia titles.
2. **Covered Query Term Ratio**: This is the proportion of the query terms that are found in the Wikipedia titles, calculated as the number of covered query terms divided by the total number of terms in the query. It is expressed as a decimal between 0 and 1, where 0 means no query terms are covered, and 1 means all query terms are covered.
3. **Number of Characters in Queries**: The total character count in the query terms.

4. **Query IDF (Inverse Document Frequency)**: This measures the importance of a term in a query, relative to its frequency across Wikipedia titles, calculated as the reciprocal of the number of titles containing the term.
5. **Query TF-IDF**: A set of features combining term count and IDF for each query term, computed using the sum, minimum, maximum, average, and variance of the product of term count and IDF for each term.

Interaction-level features are derived from the correlation between queries and Wikipedia articles. We focus on the first paragraph of each article for feature extraction. The features at this level include:

1. **The terms frequency (TF)**: A set of features that quantifies the count of each term in a query within the corresponding Wikipedia introduction paragraph in the Wikipedia corpus. These features can be calculated by taking the sum, minimum, maximum, average, and variance of the term count.
2. **Okapi BM-25**: Relevance score that measures the suitability of a Wikipedia article for a given query. It is based on a ranking function, which takes into account the term frequency and inverse document frequency of the query and document [RZT04].
3. **Semantic similarity**: Measure of how closely related queries and Wikipedia introduction paragraphs are in terms of their meaning. We use semantic similarity by calculating the cosine similarity of the embedding vectors for the texts, which can be obtained using a sentence transformer [RG19]. Pre-trained sentence transformers are used, namely all-MiniLM-L6-v2[8] and PubMedBERT (abstracts + full text)[9].

Experiment setup We use RankLib [Dan13] to build and evaluate our LeToR model, utilizing the available algorithms, which are based on point-wise, pair-wise, and list-wise approaches. These algorithms include AdaRank, Coordinate Ascent, LambdaMART, ListNET, MART, RandomForests, RankBoost, and RankNet. We split the dataset into 75% train data and 25% of test data.

We evaluate our LeToR model using two different metrics. The first is the Precision@k (P@1), and the second evaluation metric is the Mean Average Precision, as explained in Section 5.3.3. In addition, we evaluate the feature importance for the best-performing model using the forward feature selection method [DA22].

Results and Discussion The comparison of the MAP and P@1 is presented in the Table 5.5. Coordinate Ascent shows the best overall performance across all metrics. Upon further analysis, it is observed that all listed learning-to-rank algorithms have higher Mean Average Precision (MAP) and Precision@1 (P@1) scores when applied to retrieving Wikipedia articles using both popularized and specialized medical terminology than the original results obtained with ED-GENRE (*Baseline*). Specifically, the *Baseline* Wikipedia retrievals had MAP and P@1 scores of 0.85 and 0.83 for popularized terms, and 0.79 and 0.80 for specialized medical terms. This improvement demonstrates the effectiveness of learning-to-rank algorithms in re-ranking the most appropriate Wikipedia articles to appear at the top of the rankings.

[8] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2
[9] https://huggingface.co/microsoft/BiomedNLP-BiomedBERT-base-uncased-abstract-fulltext

Table 5.5: Performance of Learning-To-Rank Algorithms

Algorithm	*P*@1	*MAP*
MART	0.92	0.91
RankBoost	0.90	0.90
AdaRank	**0.94**	0.92
Coordinate Ascent	**0.94**	**0.93**
LambdaMART	0.88	0.90
RandomForests	0.86	0.89
Baseline* (Popularized Terms)	0.83	0.85
Baseline* (Specialized Medical Terms)	0.80	0.79

* Wikipedia retrieved using the ED-GENRE model.

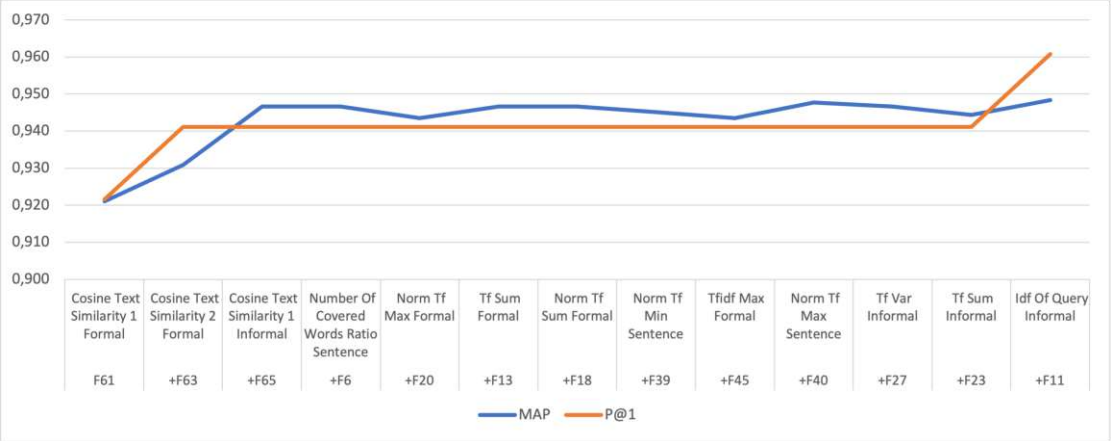


Figure 5.7: Forward feature selection to understand the feature importance of the Coordinate Ascent algorithm. The x-axis shows the sequence of feature addition until the maximum performance.

We investigate the importance of the features for the Coordinate Ascent. We used the forward feature selection [DA22] process to determine which of the contributed features have the greatest impact on the prediction result. The process begins by selecting a feature and calculating the metric value; in this case, we select the feature with the highest MAP and P@1. The feature with the best metrics is selected and appended to the list. This iterative process is continued until the coordinate ascent no longer shows any significant improvements. Figure 5.7 shows the list of features that have improved the performance of the Coordinate Ascent model.

In total, there are 13 important features out of 72 that improve the performance of the model. Among them, the three most important features were based on the semantic similarity features. These features could find the contextual relationships between queries (specialized medical terms, popularized medical terms, user posts) with the Wikipedia introduction paragraph. The remaining important features are based on a statistical feature (i.e. the ratio of covered words to sentences), and lexical features. Notably, the IDF feature associated with popularized medical term queries,

which measures the reciprocal of the number of titles containing the query term, proved to be the most effective feature and led to an increase in performance.

We acknowledge that limited topic coverage, particularly related to variations in medical terms within the corpus, can hinder the performance of Learning-To-Rank (LeToR) algorithms. Despite the improved performance compared to baselines, the effectiveness of LeToR algorithms may still be limited by potential issues such as insufficient coverage of relevant topics, lack of diversity in the corpus, and incomplete feature extraction. We are aware that limited topic coverage, which in this case is related to the variations of the medical terms of the corpus can indeed hinder the performance of Learning-To-Rank (LeToR) algorithms. Semantic similarity features and other statistical and lexical features help mitigate these issues to some extent, but they may not be sufficient to completely overcome the limitations imposed by the limited topic coverage. We propose addressing these challenges in future work.

5.5 Summary

We have developed an end-to-end informal medical entity linking (EL) model tailored to support laypeople in learning specialized medical terminology in social media settings. This end-to-end model comprises three primary modules:

1. Named Entity Recognition (NER) Module: This module is designed to identify spans of text that represent popularized medical phrases.
2. Medical Concept Normalization (MCN) Module: The aim here is to classify or normalize each popularized medical phrase, linking it to a specific specialized medical terminology.
3. Entity Disambiguation (ED) Module: The final step involves selecting the most relevant Wikipedia entry for each specialized term, providing a clear and understandable explanation of these medical terms.

A web-based system demonstrates the model's functionality, enabling users to input user text and receive detailed information processed by the model. The evaluation by experts was conducted to assess the end-to-end informal medical EL model, particularly focusing on the MCN and ED modules. We created two evaluation tasks: (1) evaluate the correctness of the MCN model in classifying the popularized medical phrase to its corresponding specialized medical terms from SNOMED-CT, and (2) the selection of appropriate Wikipedia articles, which correspond to both popularized and specialized medical terms.

We worked with three experts who have a medical background to evaluate our informal medical EL model. To measure the level of understanding on each task, we calculated the Inter-Annotator Agreement (IAA) for each task. The average Cohen's Kappa score indicated that the agreement among the experts was fair for the first task. For the second task, the level of agreement across all experts was moderate. Moreover, in the first evaluation task, the MCN model correctly classified 89% (201 out of 225 terms) popularized medical phrases into their specialized medical terms. The GENRE model utilized in the ED module demonstrated its effectiveness by successfully predicting relevant Wikipedia articles for both popularized and specialized medical term queries. Additionally, the LeToR model proved more efficient at identifying the most relevant Wikipedia articles compared to the original articles retrieved by the GENRE model.

While our expert evaluation yielded promising results, it also revealed some limitations and areas for future improvement. The MCN model encountered challenges when processing highly

popularized or colloquial expressions. A detailed analysis of these errors could provide valuable insights for refining and improving the model. Additionally, the current evaluation process utilizes binary answers, which may limit the precision of the annotation. In future work, we plan to explore the incorporation of a finer granularity of answers in the annotation process. This would allow for a more nuanced assessment of the specialized predicted terms, even if they are not the most precise but still relate to the popularized ones. Moreover, we aim to investigate methods for annotators to indicate the existence of related Wikipedia pages that may not have been retrieved and ranked by the model. This could help identify potential gaps in the model's performance and guide future improvements.

Another limitation of our current model is that it does not allow for multiple labels to be assigned to a popularized phrase. As mentioned in Section 5.4.2, the normalization task is highly subjective, and multiple labels could correspond to a single popularized phrase. This is particularly relevant for concepts that appear several times in different "branches" of the thesaurus or popularized phrases that could have multiple meanings. However, we chose to use a single-label approach in our current model for two main reasons. First, the single-label or multi-class classification approach simplifies the pipeline to the entity disambiguation (ED) module, as it allows for a more straightforward mapping between the popularized phrase, its corresponding specialized term, and the relevant Wikipedia article. Second, we believe that using a single label can make learning medical terms easier for laypeople, as it provides a clear and concise definition for each popularized phrase without overwhelming the laypeople with multiple meanings or related concepts.

Despite these reasons, we acknowledge that allowing for multiple labels could better capture the complexity and subjectivity of the normalization task. In future work, we plan to explore the possibility of incorporating multiple labels for a popularized phrase while still maintaining the simplicity and ease of use for laypeople in learning medical terminology. This could involve developing a pipeline that can handle multiple labels and their corresponding Wikipedia articles, as well as designing user interfaces that present the multiple labels in a clear and accessible manner.

Furthermore, we aim to expand our LeToR datasets to better generalize the performance of the LeToR model in re-ranking relevant Wikipedia articles. This will be achieved by evaluating their performance on a diverse set of queries and comparing the results with a more comprehensive corpus that covers a wider range of topics related to medical terms. By addressing these limitations and exploring the proposed future directions, we aim to enhance the robustness and effectiveness of our informal medical EL model. Additionally, incorporating contextual information (such as surrounding sentences) into the model could potentially improve its performance, particularly in handling highly popularized or colloquial expressions. We plan to explore this possibility in future studies.

CHAPTER 6

Informal Medical Entity Linking for Learning Medical Terminology

In this chapter, we evaluated the effectiveness of the informal medical entity linking model in helping laypeople learn specialized medical terms knowledge from social media. As previously discussed in Section 2.1, functional literacy encompasses the basic reading and writing skills necessary to understand medical information. This includes the comprehension of medical terminology.

In real-world scenarios, we aim to integrate the informal entity linking model with social media platforms. As such laypeople could be exposed to become familiar with the specialized medical terms. Our proposed model will annotate each user's social media posts by identifying the relevant specialized medical terms that correspond to the popularized medical phrases used in the post. The goal is to bridge the gap between popularized medical phrases and specialized medical terminology by automatically providing the specialized medical terms and their definitions alongside the popularized medical phrases within user-generated social media content. Due to the expense and time required for real-world evaluation, we assessed the model in controlled settings.

To evaluate the model, we developed an evaluation design inspired by the approach outlined in Lalor et al. [LWY19]. In their study, they used ComprehENote, a test instrument to assess a patient's ability to comprehend Electronic Health Record (EHR) notes [LWC+18]. This instrument was used to explore whether integrating NoteAid [CDPR+18], an NLP tool that links medical terms in EHR with simplified explanations from external sources, enhances the comprehension of EHR notes.

We structured our evaluation design with an educational approach. We created an evaluation instrument derived from vocabulary knowledge tasks. Vocabulary knowledge poses a difficulty in language learning and plays an important role in the overall language competence of a second language [EAZG18]. Nation [Nat01] categorizes vocabulary knowledge into three aspects: form, which includes spoken and written forms , and word parts; meaning, includes form and meaning connections, concepts and referent, and word associations; and use, refers to grammatical functions, collocations with other words, and constrains on use (i.e when and how often to use word).

Drawing on these frameworks, we developed an instrument to assess two key aspects of vocabulary knowledge:

1. **Surface-level Term Familiarity**: This refers to recognizing the word form of a specialized medical term that corresponds to a popularized medical phrase, with context.
2. **Concept-level Term Familiarity**: This extends beyond just recognizing medical terms. It assesses whether laypeople can understand the underlying meaning of a medical term.

Measuring the *surface-level* and *concept-level* term familiarity aligns with the process of vocabulary acquisition, where laypeople encounter new terms and begin to establish connections between the word form and its meaning [Nat01]. This stage is consistent with the receptive level of vocabulary knowledge [Nat01], which involves recognizing and understanding words (in terms of both form and meaning) when they are encountered [Sch14].

The contributions of this chapter are as follows:

- We created an evaluation design and instruments to assess whether our informal medical entity linking model can help laypeople learn medical terminology in a social media setting.
- We conducted a user study to assess how effectively the informal medical EL model.

The structure of this chapter is as follows: In Section 6.1, the design and the implementation of our evaluation tool are described. Section 6.2 describes the setup of the experiment. The results and discussions are presented in Section 6.3. Finally, Section 6.4 provides a summary of the chapter.

6.1 Experiment Planning

In this section, we present a detailed explanation of the experiment planning to evaluate the effectiveness of the informal medical EL model in supporting laypeople to improve their medical terminology knowledge on both *surface* and *concept-level* term familiarity.

We hypothesize that by integrating this model into social media platforms, it could enhance laypeople's comprehension of medical terms they encounter in their everyday use of these platforms. However, due to resource and time limitations, we were not able to conduct this evaluation by integrating the model into a real social media platform and observing its effectiveness on real users. Instead, our evaluation was conducted in a controlled setting that may not fully represent the complexity and variability of user behavior and interaction in real-world social media environments.

6.1.1 Goal

Our first goal is related to the effectiveness of the model in assisting laypeople with *surface-level* term familiarity. This goal (G1) is to *analyze* the difference in the level of *surface-level* term familiarity among laypeople. This evaluation focuses on how the informal medical EL model affects their ability to memorize and retain the word-form of specialized medical terminology.

Our second goal (G2) is to analyze the difference in the level of *concept-level* term familiarity among laypeople. This goal focuses on the effectiveness of the informal medical EL model, specifically how it influences laypeople to identify the definitions or meanings behind specialized

medical terminology. This evaluation builds on the first goal, shifting the focus from memorization to a deeper understanding of the terms. These analyses are conducted through an experiment involving recruited laypeople.

6.1.2 Participants

In this experiment, we recruited participants from the Prolific platform. We selected Prolific due to its compliance with the General Data Protection Regulation (GDPR)[1]. Additionally, Prolific is known for providing higher data quality compared to other platforms [PRGD22]. We created inclusion criteria requiring participants to be more than 18 years old, fluent in English, and located in an English-speaking country to prevent language bias. The Prolific participants were split into two groups: *non-intervention* and *intervention*. This setup allowed us to evaluate how our informal medical EL model supports participants' understanding of medical terminology.

- The *non-intervention* group completed evaluation tasks **without** any support from the informal medical EL model.
- In contrast, the *intervention* group completed evaluation tasks **with** any support from the informal medical EL model.

The workflow of the experiment is illustrated in Figure 6.1. In total, we recruited 150 participants with 50 assigned to the *non-intervention* group and 100 assigned to the *intervention* group. The reason for the larger number of participants in the *intervention* group was to gain a more comprehensive understanding of the impact of the informal medical EL model on a greater number of participants. Each participant in both groups completed a series of tasks, including *(i) demographic survey, (ii) surface-level task, (iii) concept-level task, and (iv) a feedback survey*. We collected scores from both the *surface* and *concept-level* tasks for further analysis. *Surface-level scores* measure the ability to recognize the word-form of specialized medical terms, while *concept-level scores* assess the ability to identify the meaning of these terms. Detailed descriptions of these tasks and the formula for score calculation will be provided in the later sections.

6.1.3 Evaluation Design Workflow

We developed two distinct evaluation designs for each group: one incorporating the intervention from our informal medical entity linking, and the other without it.

Intervention-Based Evaluation Design Workflow

The intervention-based evaluation design workflow aims to support the hypothesis that exposure to the model when accessing user posts helps participants gain knowledge of medical terminology. The workflow, illustrated in Figure 6.2, involves the following steps: (1) User posts from the AskAPatient forum are annotated using an informal medical entity linking (EL) model, providing lists of highlighted popularized medical phrases and their specialized counterparts with definitions from Wikipedia. (2) The annotated user posts are utilized for the *surface-level* term familiarity task, which employs the spaced repetition technique to enhance memory retention [SM16]. The layperson is prompted to recall specific medical concepts, with increasing time gaps between presentations. To align with practical constraints, this method is adapted to a more feasible format by alternating between different medical concepts for varied yet repeated exposure within

[1] https://researcher-help.prolific.com/hc/en-gb/articles/360009094594-Data-protection-and-privacy

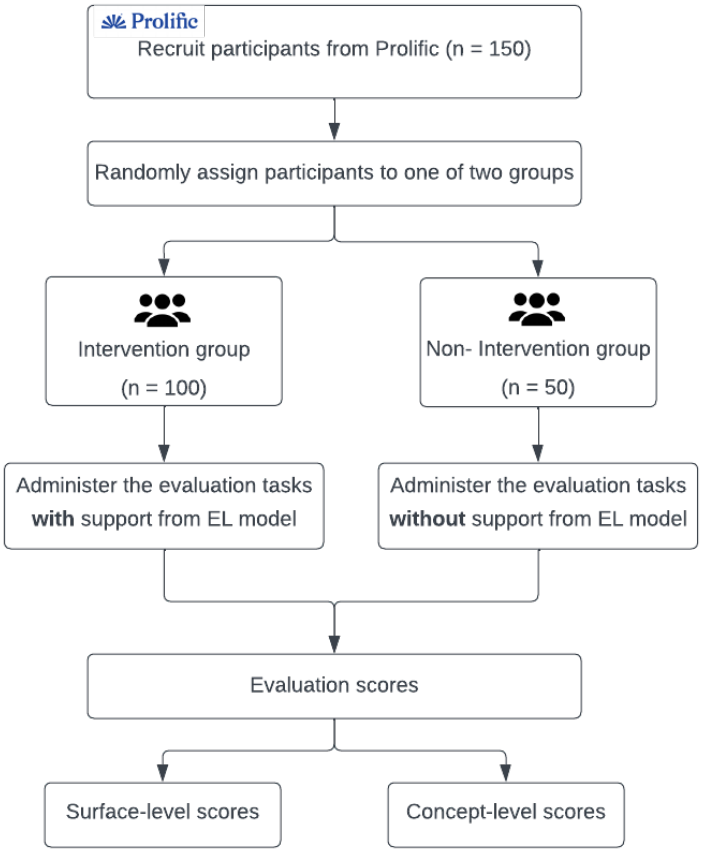


Figure 6.1: Participant Recruitment and the Experiments Description

a shorter time frame. Despite the model providing a list of annotated terms in each user post, the evaluation process presents only one popularized medical phrase per post, leveraging a microlearning approach to ensure focused learning. Microlearning involves presenting *micro*, *bite-sized* content to teach specific concepts step-by-step [Hug12]. The *surface-level* term familiarity is structured as follows:

2.A. *Surface-level* Term Familiarity *Learning*: Introduce specialized medical terms corresponding to popularized medical phrases found in user posts, providing hints to assist laypeople in familiarizing the concepts (controlled condition).

2.B. *Surface-level* Term Familiarity *Testing*: Evaluate memory retention of specialized medical concepts by assessing whether laypeople can recognize the word-form of the concepts at certain spaced intervals. We repeated this process three times.

After completing *surface-level* term familiarity task, participants continue to (3) *Concept-level* term familiarity task that assesses their ability to identify the meanings of specialized medical terms encountered in the *surface-level* task. It serves as a follow-up, where participants identify

the definitions of specialized terms using explanations provided by the informal medical entity linking model.

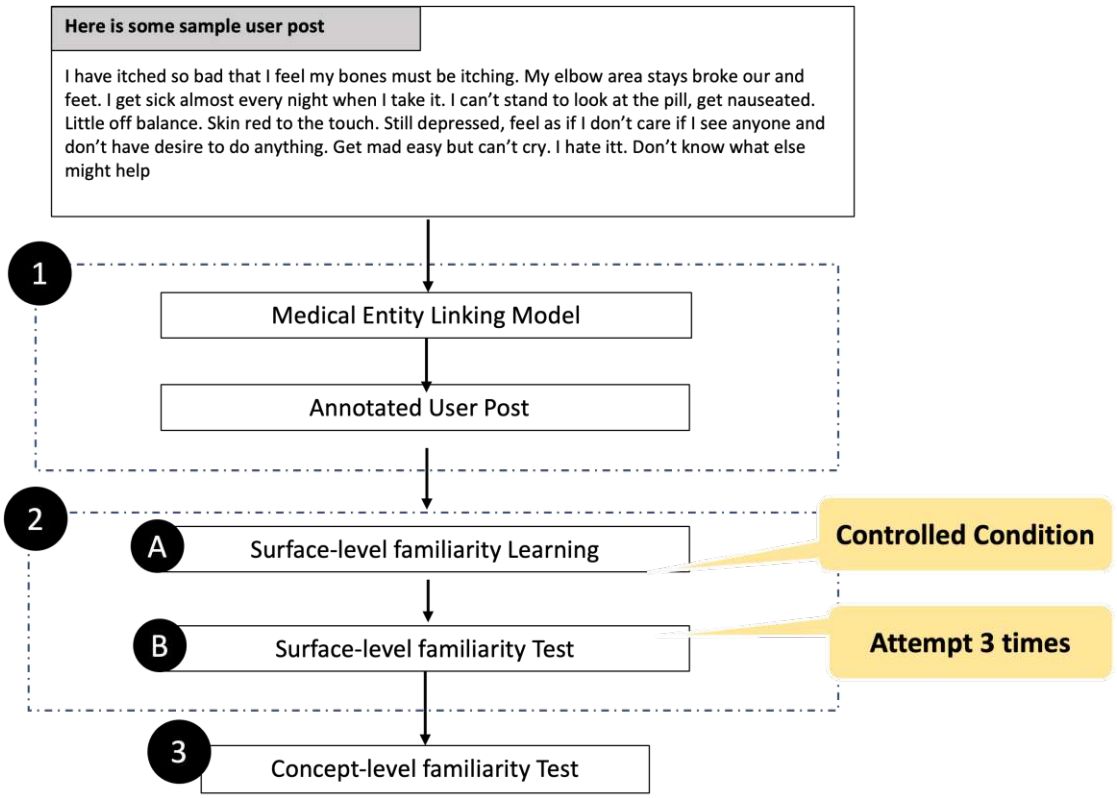


Figure 6.2: Intervention-Based Evaluation Design Workflow for Intervention Group

Figure 6.3 shows where participants were introduced to "Dyspnea", the medical term for the feeling of inadequate breathing[2]. They were shown user posts with related popularized phrases, such as "inability to take a deep breath" from the Step (1). For the *surface-level* we created a multiple-choice question to assess their ability to recognize the word-form of specialized medical term. The correct answer was taken from our EL model output, while the other options were manually added by the researcher. Participants encountered the term "inability to take a deep breath" for the first time, we provided the correct term "Dyspnea" as a hint (this refers to Step (2A)). Should the participants choose the correct term, we then display detailed information, the correct medical term and its explanation from Wikipedia are displayed, based on our informal medical EL model (see Figure 6.4), which provides immediate feedback and definition of the term. This helped participants understand the meaning of a specific medical term.

In subsequent steps, referred to as the *surface-level* term familiarity testing phase (Step (2B)), this process was repeated three times for each concept. The participants again were introduced to the new user posts with the same or different popularized medical phrases for the same medical concept. If the participant's response is correct, their mastery level progresses through four stages

[2] $https://en.wikipedia.org/wiki/Shortness_of_breath$

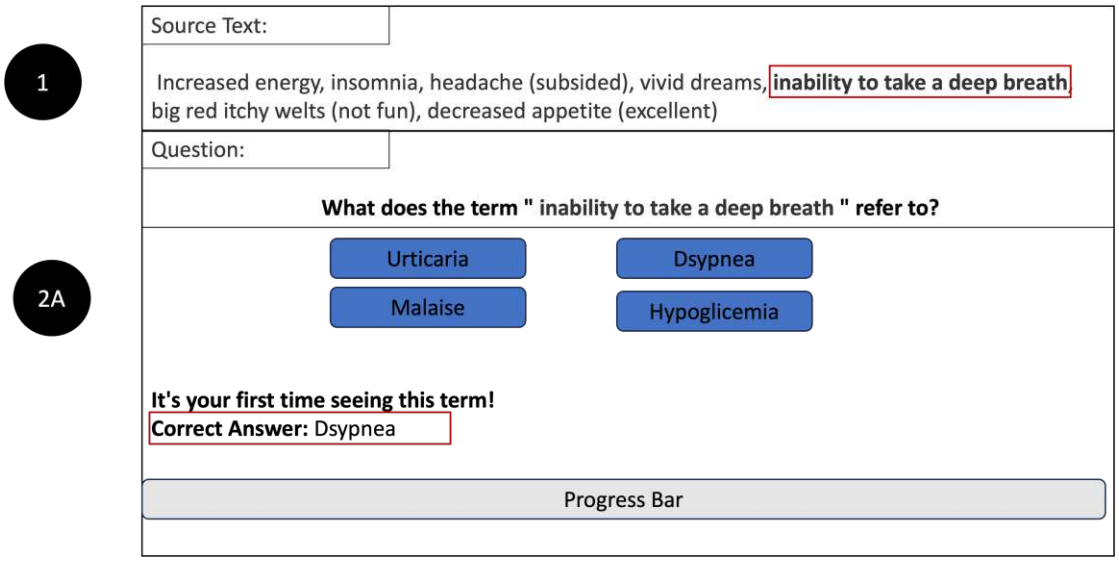


Figure 6.3: *Surface-level* term familiarity learning step: (1) User posts annotated by an informal medical entity linking model, e.g., the popularized phrase "**inability to take a deep breath**" linked to the specialized term "**Dyspnea**". (2A) For *surface-level* term familiarity learning, laypeople provided hint for "**Dyspnea**" corresponding to popularized phrase (controlled condition)

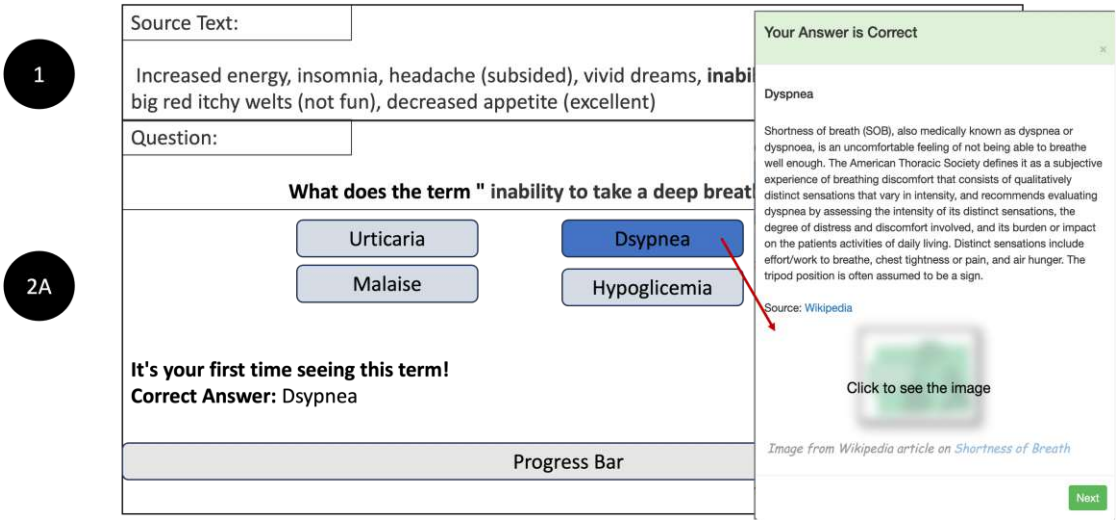


Figure 6.4: *Surface-level* term familiarity learning step: Detailed information on the correct/incorrect answer

(25%, 50%, 75%, and 100%), with a score of 1 awarded for each level achieved. Incorrect answers resulted in a score of -1 and the participant moving back one level.

The mastery level was determined by the total number of repetitions for each concept, combining one repetition from the learning phase and three from the testing phase. The purpose of this repetition is twofold: first, to demonstrate that exposing users to various popularized medical phrases associated with their specialized medical terminology and explanations could enhance their medical terminology knowledge. Second, it is aimed to prevent users from randomly guessing the correct answer by chance.

During this phase, participants were shown a new user post with the same or different popularized medical phrase for the same medical concept (see Figure 6.5). They were expected to select the correct answer without any hints. After each selection, detailed information about the correct medical term and its explanation from Wikipedia was provided (Figure 6.6).

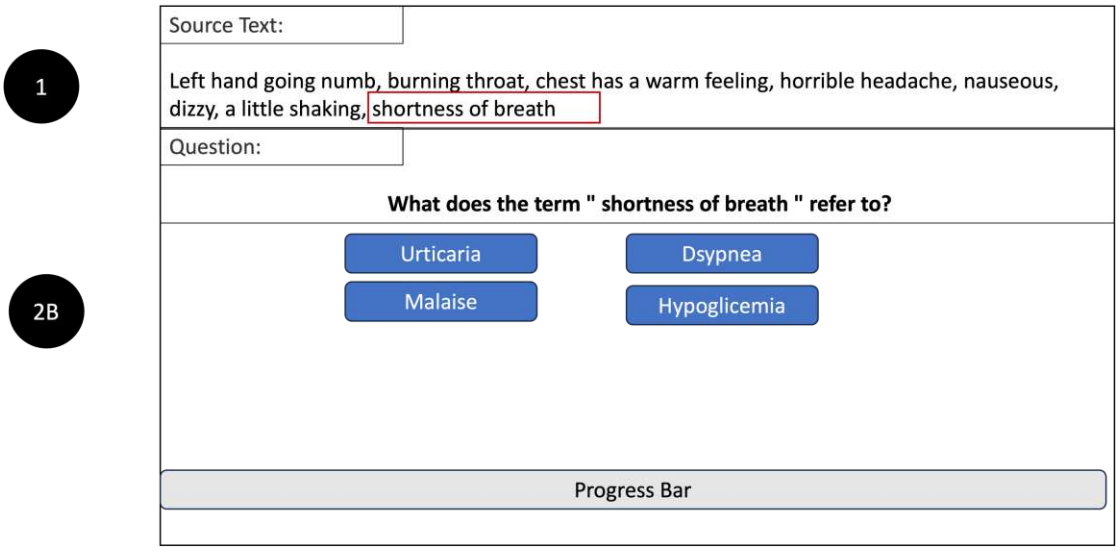


Figure 6.5: *Surface-level* term familiarity testing: Next repetition - a participant directly responds to the question

The purpose of the *surface-level* term familiarity testing phase was to assess the participant's ability to recall and identify each medical concept accurately by exposing them to the information from our entity linking (EL) model. Once all concepts have been presented, we calculate the *surface-level score*, which will be explained in the following section. The details of the *surface-level* term familiarity workflow are presented in Figure 6.7. In contrast to the *surface-level* term familiarity task, we designed a *concept-level* term familiarity task where participants identify the definitions of medical concepts they previously encountered. The purpose of this task is to assess their ability to identify the meanings of medical terminology. Following the method used by Puspitasari et al. [PMFN15], we use multiple-choice questions with one correct answer and two distractors. If participants are unsure, they can select "I don't know." We then ask them to choose the correct definitions for the medical concepts they learned earlier (see Figure 6.8). This task evaluates whether the information provided by our EL model during the *surface-level* term familiarity task helps them understand the meaning of each specialized medical term. We

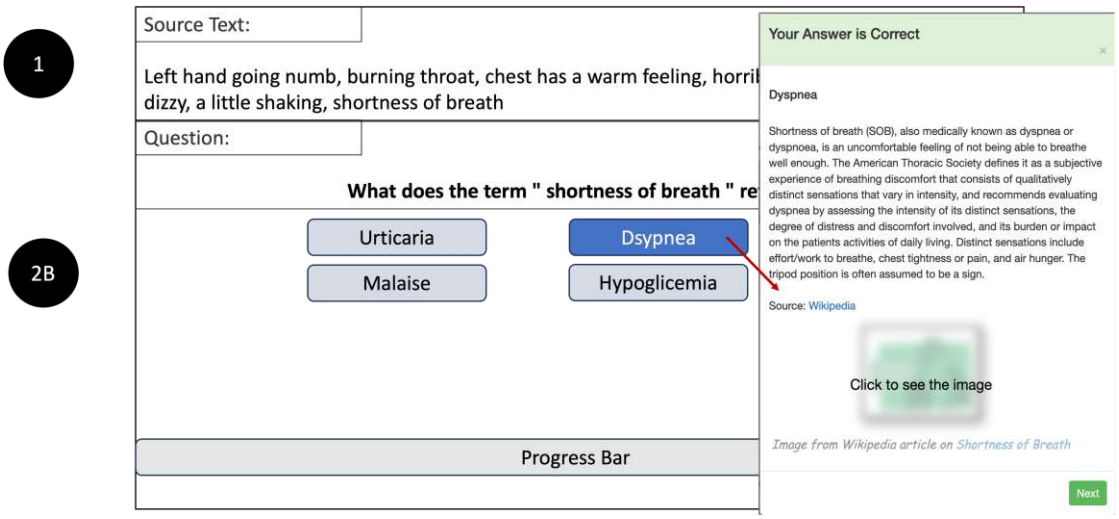


Figure 6.6: *Surface-level* term familiarity testing: Detailed information on the correct/incorrect answer

then calculate the *concept-level score* for each participant, which we will explain in the following section.

Non-Intervention-Based Evaluation Design Workflow

This evaluation design was developed for the *non-intervention* group, which did not receive support from the EL model to administer the evaluation task. Figure 6.9 illustrates the workflow for the non-intervention-based design. The overall workflow is similar to the intervention-based design workflow, utilizing the same data. It involves the following steps: (1) User posts annotated by an informal medical EL model are presented. (2) These annotated user posts are utilized for the surface-level term familiarity task. In contrast to the intervention group, the *surface-level* term familiarity task given to the *non-intervention* group was more basic. (2B) Participants are shown annotated posts containing a popularized phrase and asked to identify the corresponding specialized medical term. After completing the surface-level term familiarity task, participants proceed to (3) the concept-level term familiarity task.

Figure 6.10 shows where participants in *non-intervention* group were asked to find the specialized counterpart of the "shortness of breath". After participants responded, no detailed feedback was given that would display the correct answers, such as the corresponding specialized medical term and its definition (see Figure 6.11).

This applied even when the medical concept was introduced for the first time. Therefore, participants did not receive any assistance from our informal medical EL model to administer the surface-level term familiarity task. For each medical concept, we repeated two times. The repetition is to prevent the participants from selecting the correct answer randomly and also to keep a similar format with the intervention group. Other than that, we did not want the participant to have a chance to learn when we repeated more than two times. Once all concepts have been presented, we calculate the *surface-level score*, which will be explained in the following section. The details workflow of the *surface-level* term familiarity in non-intervention-based evaluation design is presented in Figure 6.12. Furthermore, for the *concept-level* term familiarity

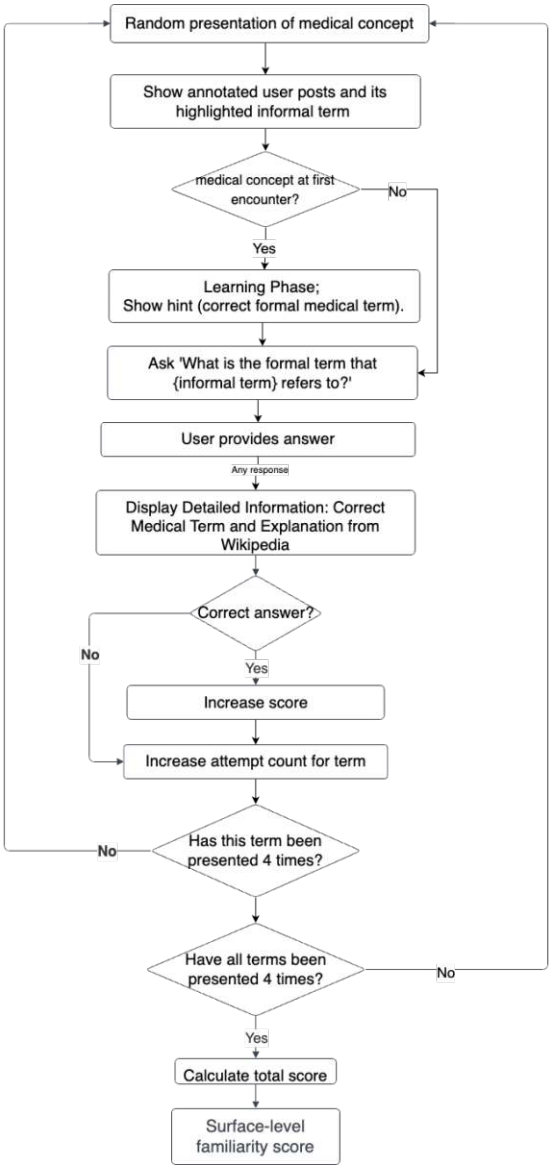


Figure 6.7: Workflow of the *surface-level* term familiarity in the *intervention* group

task where participants were asked to identify the definition of medical concepts encountered in *surface-level* term familiarity. We utilized the same design and questions with *intervention* group, which is illustrated in Figure 6.8.

6.1.4 Medical Concepts Selection

As mentioned in the previous section, the first phase of our evaluation design scenario is data preparation. In this section, we explain our method for selecting medical concepts for the user experiment, which aims to create evaluation data to assess our informal medical entity linking

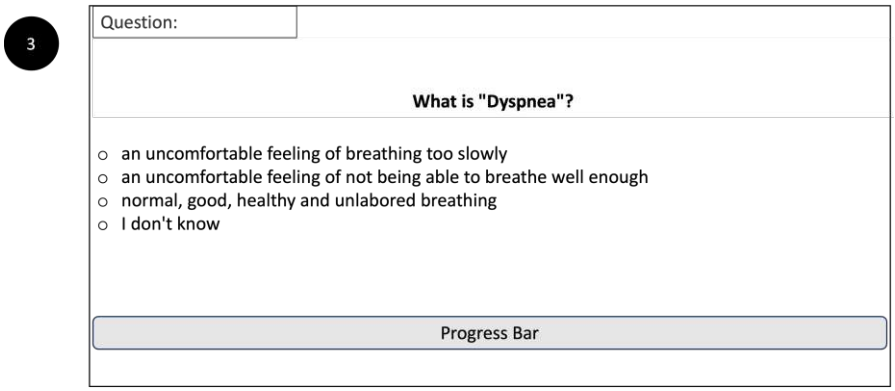


Figure 6.8: An overview of the approach used to assess the *concept-level* familiarity task

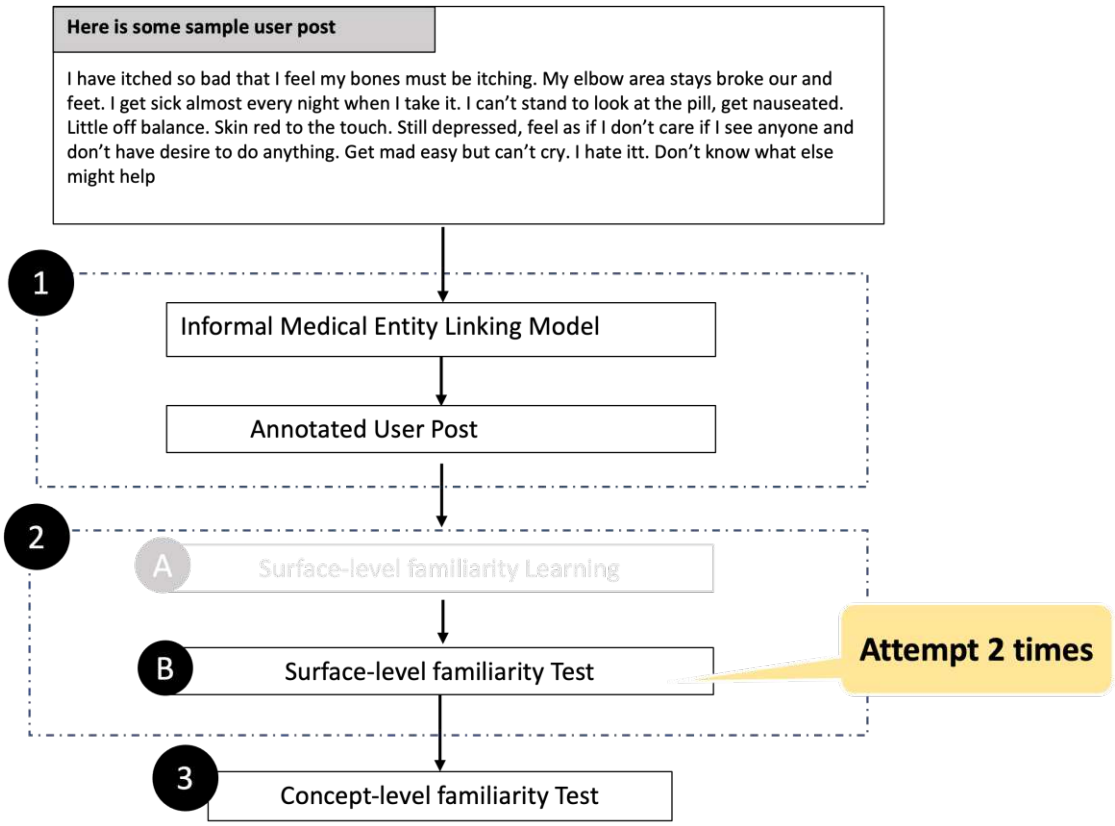


Figure 6.9: An overview of the approach used to assess the *concept-level* familiarity task

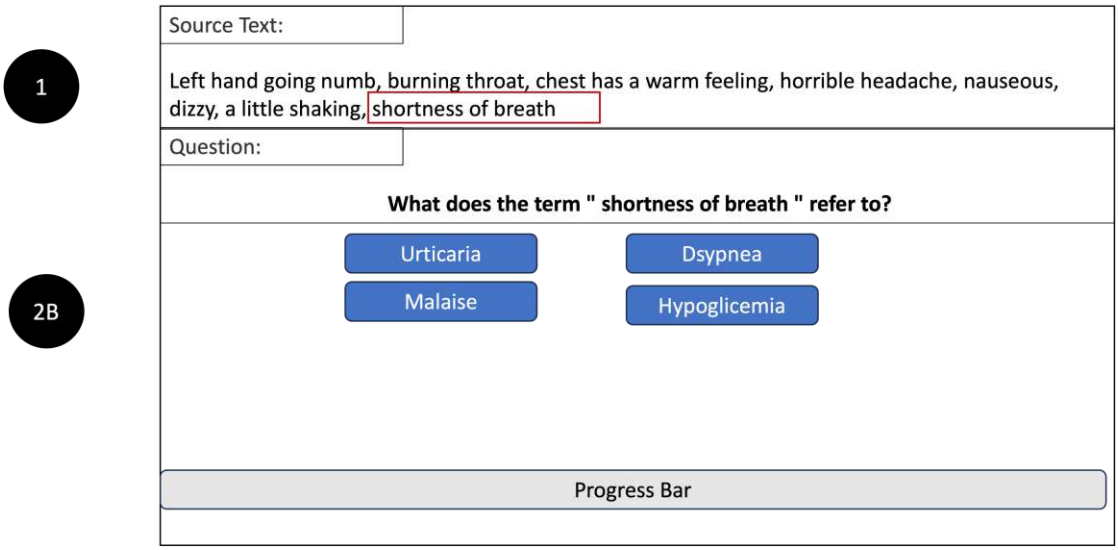


Figure 6.10: *Surface-level* term familiarity testing: a participant in *non-intervention* directly answer the question

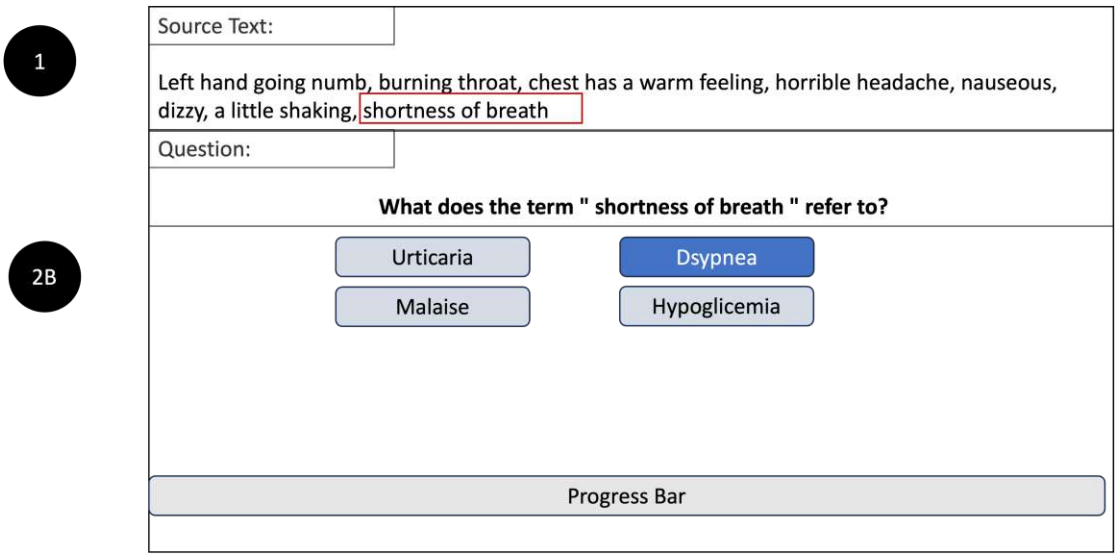


Figure 6.11: *Surface-level* term familiarity testing: a participant answered the question

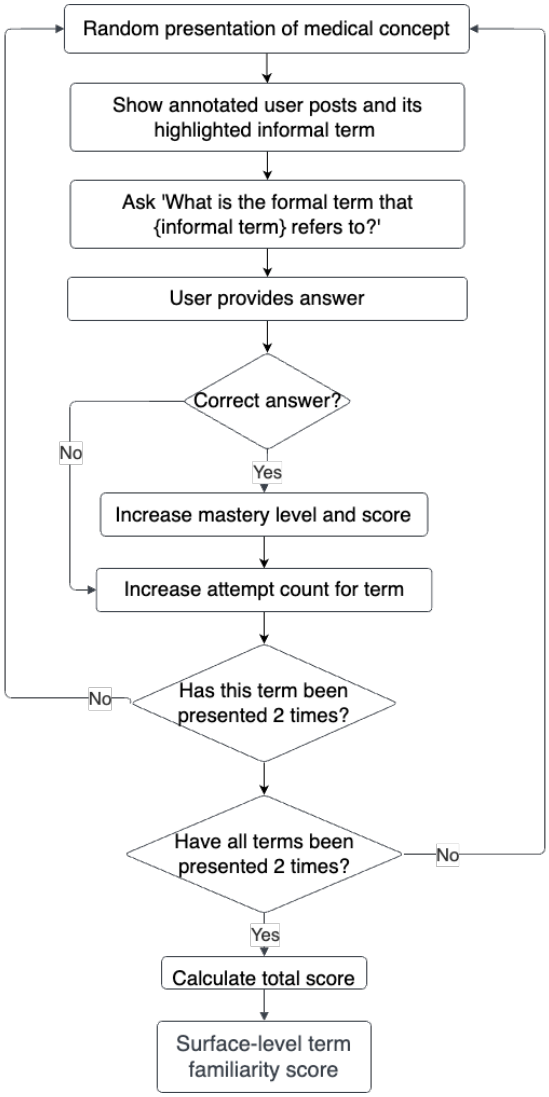


Figure 6.12: Workflow of the *surface-level* term familiarity in the *non-intervention* group

model. This involved a systematic process for extracting medical concepts from user-generated content, as illustrated in Figure 6.13. We collected user posts from AskAPatient.com. The training dataset for our informal medical entity linking model, particularly the medical concept normalization stage, relies on publicly available datasets [KMJKW15, ZFP+19, BLSC20], that primarily focus on antidepressant medications. Therefore, we specifically selected posts related to *Wellbutrin SR* (580 posts), which received a high number of ratings and positive opinions from consumers, with average ratings of 3.5.

In addition to antidepressants, we also included *Glucophage* (557 posts), a medication used to treat type 2 diabetes, in our data collection. *Glucophage* achieved higher ratings among blood regulator drugs, and its inclusion is relevant due to the increasing prevalence of diabetes. According to the World Health Organization (WHO) the prevalence of *diabetes mellitus* increased from 108

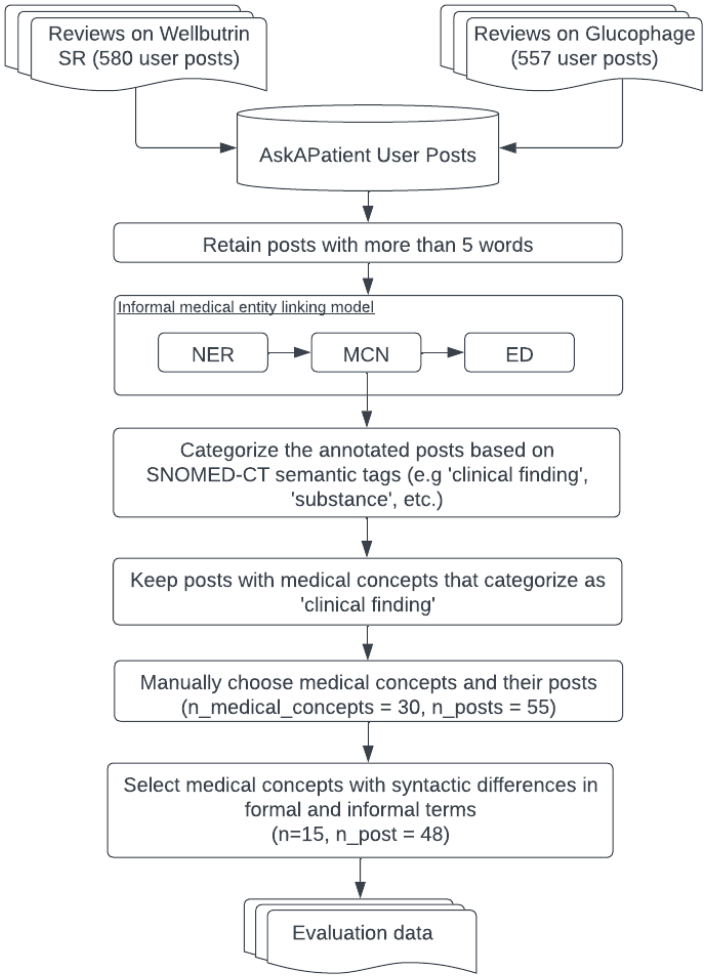


Figure 6.13: Medical concept selection process

million (4.7%) in 1980 to 425 million (8.5%) in 2017, and it is estimated to be 629 million by 2045 [ZLH+16]. By incorporating reviews of different medications, we aim to access a diverse range of medical concepts for learning purposes that reflect real-world layperson language and experiences.

To provide additional context for laypeople during the learning phase of the user experiment, we kept user posts containing more than 5 words. Each post was automatically annotated using our proposed informal medical entity linking model, which identifies and links popularized medical phrases mentioned by laypeople to their specialized counterparts along with an explanation. Following the annotation, the posts were categorized based on the specialized medical terms derived from the annotations. The medical terminology predicted from the MCN output can cover various SNOMED-CT categories, including substances, medical procedures, diseases, findings, and more. Therefore, we focus on "clinical finding (finding and disorder)" category. This category

Table 6.1: The Example of Evaluation Data

User Post	Medical Concept	Popularized Phrase	Wikipedia
Nausea, stomach cramps, bad headaches, fatigue, drowsiness, pain in upper right quadrant, generally felt unwell.	Malaise	"felt unwell"	Malaise
Sudden and urgent need to have bowel movements; nausea, stomach cramps– felt increasingly unwell.	Malaise	"felt increasingly unwell"	Malaise
Left hand going numb, burning throat, chest has a warm feeling, horrible headache, nauseous, dizzy, a little shaking, shortness of breath.	Dyspnea	"shortness of breath"	Shortness of Breath
Increased energy, insomnia, headache (subsided), vivid dreams, inability to take a deep breath, big red itchy welts (not fun), decreased appetite (excellent)	Dyspnea	"inability to take a deep breath"	Shortness of Breath

includes symptoms and disease (disorder)[3], which is one of the common categories searched by laypeople [MBS24].

We manually selected 30 medical concepts from 55 distinct posts. Each of these concepts is represented by at least 2 distinct popularized phrases from the dataset. The purpose is to demonstrate the diversity of medical concepts as perceived by laypeople. Additionally, we focused on medical concepts where there is a difference between the specialized medical terminology and the popularized phrases used by laypeople. This demonstrates the communication gap between medical professionals and laypeople. As a result, we refined our selection to 15 medical concepts from 48 posts, including *Urticaria, Tachycardia, Emesis, Malaise, Arthralgia, Pollakisuria, Paresthesia, Sleep deprivation, Tinnitus, Clouded consciousness, Xerostomia, Somnolence, Dyspnea, Hypoglycemia, and Pyrosis.*

This selection process aimed to ensure the evaluation covered a diverse range of medical concepts expressed by laypeople and targeted concepts with notable differences between specialized medical terminology and popularized phrases. The goal was to demonstrate the model's effectiveness in supporting laypeople's comprehension of medical terminology. Table 6.1 shows extracts of the evaluation data used in our user experiment, including user posts that contain selected medical concepts, their popularized phrases, and corresponding Wikipedia articles, as identified by our informal medical entity linking model.

6.1.5 Evaluation Metrics

This section outlines the metrics used to assess the effectiveness of the informal medical Entity Linking (EL) model.

[3]https://confluence.ihtsdotools.org/display/DOCEG/Clinical+Finding+and+Disorder

Surface-level Score: The *surface-level score* is a score to measure the ability of participants to recognize the word-form of specialized medical terms corresponding to popularized medical phrases found in user posts. To measure the *surface-level score*, we propose using three different metrics: (1) $familiarity$, which measures the mastery levels of the medical concepts; (2) Answer Accuracy ($answer\_accuracy$), which measures how well the participants can recall the correct answer across multiple attempts; and (3) Learning Rate Estimation (LRE), which measures the ability to provide correct answers in final attempts.

Familiarity Score: The score evaluates the ability of the participants to memorize and remember specialized medical terms given to popularized medical phrases derived from the annotated user post. The calculation of this metric is based on the mastery level attained by the participants for each medical concept presented to them in the surface-level term familiarity assessment. As outlined in Section 6.1.3, a score of 1 is awarded for a correct participant response, and the mastery level progresses to the next level. Conversely, an incorrect answer results in a score of -1, with a minimum score of 0 if incorrect responses are provided at all levels, and the participant is pushed back to one mastery level.

We proposed two evaluation metrics to measure the $familiarity$ score for both *intervention* and *non-intervention* groups. For the intervention group, the $familiarity$ score is calculated as follows:

$$\text{familiarity}_{\text{intervention}} = \frac{0.0s_0 + 0.25s_1 + 0.5s_2 + 0.75s_3 + 1.0s_4}{\sum_{i=0}^{4} s_i} \times 100 \tag{6.1}$$

Where:

- s_i $(i = 0, 1, 2, 3, 4)$ represents the number of concepts a participant answered correctly i times.
- The denominator $\sum_{i=0}^{4} s_i$ is the total number of concepts attempted.
- Weights (0.0, 0.25, 0.5, 0.75, 1.0) correspond to mastery levels:
  - 0% (fully incorrect)
  - 25% (correct once)
  - 50% (correct twice)
  - 75% (correct three times)
  - 100% (fully correct)

For the non-intervention group, the $familiarity$ score is calculated as:

$$\text{familiarity}_{\text{non-intervention}} = \frac{0.0s_0 + 0.5s_1 + 1.0s_2}{\sum_{i=0}^{2} s_i} \times 100 \tag{6.2}$$

Where:

- s_i $(i = 0, 1, 2)$ represents the number of concepts a participant answered correctly i times.
- The denominator $\sum_{i=0}^{2} s_i$ is the total number of concepts attempted.

- Weights (0.5, 1.0) correspond to mastery levels:
  - 0% (fully incorrect)
  - 50% (correct once)
  - 100% (fully correct)

However, we acknowledge that these different evaluation metrics could lead to incomparable results among the groups due to the varying number of mastery levels. To address this issue, we extended the *surface-level* score by introducing additional evaluation metrics that consider only the first two attempts in the task for the *intervention* group, making it comparable to the *non-intervention* group.

We then propose using a new set of evaluation metrics to measure ability in recognizing *surface-level* term familiarity (*surface-level score*). We measure the participant's ability in two folds: (1) *Answer Accuracy (answer_accuracy)*, and (2) *Learning Rate Estimation (LRE)* scores.

Answer Accuracy Score: The answer accuracy evaluates how well participants can remember or memorize the correct word-form of specialized medical terms given the popularized medical phrases extracted from annotated user posts. For each correct answer, participants will receive a score of 1, while an incorrect answer will receive 0 points. The difference between the answer accuracy and $familiarity$ is we did not punish the participants if their answers were incorrect, and focus on the the *testing* phase of the *surface-level task*.

Let $\mathcal{C} = \{c_1, c_2, \ldots, c_n\}$ be the set of all medical concepts under evaluation. For each concept $c_i \in \mathcal{C}$, the answer accuracy score is defined as:

$$ans\_acc(c_i)@k = \frac{N_c(c_i)}{N_c(c_i) + N_i(c_i)} \tag{6.3}$$

Where:

- $ans_acc(c_i)@k$ is the answer accuracy score for concept c_i after k attempts.
- $N_c(c_i)$ is the number of correct responses for concept c_i out of k attempts.
- $N_i(c_i)$ is the number of incorrect responses for concept c_i out of k attempts.
- $k = N_c(c_i) + N_i(c_i)$, representing the total number of attempts.

The aggregate answer accuracy score across all concepts, denoted as answer_accuracy, is calculated as:

$$\text{answer\_accuracy} = \frac{1}{|\mathcal{C}|} \sum_{c_i \in \mathcal{C}} ans\_acc(c_i)@k \times 100 \tag{6.4}$$

Where:

- $answer\_accuracy$ is the aggregate answer accuracy score accross all concepts.
- $|\mathcal{C}|$ is the total number of concepts (cardinality of set $\mathcal{C}$).

- $ans\_acc(c_i)@k$ is the answer accuracy score for concept c_i after k attempts as defined in the previous equation.

Learning Rate Estimation (LRE) Score: The Learning Rate Estimation (LRE) score quantifies a participant's ability to improve their performance over multiple attempts. It assesses whether participants provide more accurate answers in later attempts compared to earlier ones. For instance, learning progress is demonstrated if a participant's performance improves from their second to their third attempt, or if they consistently give correct answers in subsequent attempts after initially answering correctly.

The LRE score calculation draws inspiration from the Average Precision (AP) metric, commonly used in information retrieval. However, the LRE score reverse the priority of AP to focus on learning progress. Traditional AP assigns higher importance to relevant answers at higher ranks (i.e., top k positions). In contrast, the LRE score focuses on learning progress by giving more weight to correct answers that occur later in the sequence of attempts. This prioritization of later correct responses enables the LRE score to effectively capture and quantify improvements in a participant's performance over time.

For a concept $c \in \mathcal{C}$ with n attempts, define:

$$P@k = \frac{1}{k} \sum_{i=1}^{k} I(a_{(n-i+1)}), \quad 1 \leq k \leq n \tag{6.5}$$

where $I(a_i)$ is an indicator function:

$$I(a_i) = \begin{cases} 1 & \text{if attempt } a_i \text{ is correct} \\ 0 & \text{otherwise} \end{cases}$$

This equation calculates the precision at a given attempt k. It represents the proportion of correct answers out of the k most recent attempts, where a_n is the latest attempt.

The Average Precision for concept c is:

$$AP_c = \frac{1}{n} \sum_{k=1}^{n} P@k \tag{6.6}$$

where n is the total number of attempts. This equation computes the Average Precision for a specific concept by averaging the precision values across all attempts, from the most recent to the earliest. A higher AP_c suggests that participants tend to give more correct answers in their later attempts, indicating positive learning. This implies that a participant has the potential to memorize the correct medical concepts, and we hypothesize that the model could support this memorizing of the word-form process. In contrast, a lower AP_c suggests less improvement over time. This metric helps evaluate how effectively the entity linking model supports learning of medical terminology.

The Learning Rate Estimation (LRE) score across m concepts is:

$$\text{LRE} = \frac{1}{m} \sum_{j=1}^{m} AP_{c_j} \tag{6.7}$$

This equation calculates the average AP score (AP_c) across multiple concepts. Here, m represents the total number of concepts being evaluated.

Concept-level Score: The *concept-level score* measures participants' ability to identify the correct definition of specialized medical terms without any context. For each correct answer, participants receive a score of 1, while an incorrect answer receives 0 points. The formula to calculate the *concept-level score* is:

$$conceptLevel = \frac{\text{Number of correct answers}}{\text{Total number of concepts}} \times 100 \tag{6.8}$$

where *Number of correct answers* is the count of correctly answered concepts by the participant, and *Total number of concepts* is the total number of medical concepts.

Threshold Determination: To categorize participants' scores into "high" and "low" levels, we used a threshold determined by calculating the weighted average of the mean scores from the two groups, considering their sample sizes. This cutoff point allowed us to classify the scores into predefined levels based on the weighted average to calculate the overall average of all participants.

Let $\mathcal{G} = \{g_1, g_2\}$ be the set of participant groups, where g_1 is the non-intervention group and g_2 is the intervention group.

For each $g_j \in \mathcal{G}$, let N_j be the number of participants and $\bar{S}_j$ be the mean score in group g_j.

The threshold is:

$$T = \frac{\sum_{j=1}^{2} N_j \bar{S}_j}{\sum_{j=1}^{2} N_j} \tag{6.9}$$

Define the level function:

$$\text{Level}(S) = \begin{cases} \text{High} & \text{if } S \geq T \\ \text{Low} & \text{if } S < T \end{cases}$$

where S is either the $familiarity$ score or Concept-level score.

6.1.6 Demographic and User Feedback Survey

In addition to the *surface-level* and *concept-level* term familiarity tasks, we also gathered demographic information and conducted a user feedback survey. The demographic survey, which participants are required to complete before beginning the assessment tasks, includes basic details such as gender, age, education level, English proficiency, annual income, and experience with searching for medical information.

Furthermore, participants were asked to read two health-related articles on "Ulcerative Colitis" sourced from *Healthline.com* and an abstract from a research publication [FC14]. Subsequently, we asked participants to self-assess their comprehension of the articles and their familiarity with the medical terminology used in both texts, using a scale ranging from 1 (indicating 'do not understand at all') to 5 (indicating 'understand very well'). Although we did not formally measure

their health literacy level, our intention was to gain insights into their ability to understand health-related articles.

As for the user feedback survey, it is divided into two parts. The first part focuses on the learning experience during the learning phase of the surface-level term familiarity task. In the second part, we collect user feedback regarding the idea of implementing informal medical EL in social media. The questions primarily ask about the usefulness of features such as specialized medical terms and their explanations. Additionally, it collects data on which medical topics participants are most interested in and what other sources of information they find necessary for understanding these concepts. Participants are required to complete the feedback survey in the last session after finishing the assessment tasks. Both surveys can be found in Appendix 7.3.3.

6.1.7 Experimental Procedure

In this section, we outline the experimental procedure. This includes a detailed description of the sequence of tasks presented to the participants, the specific data used for each task, and the time allocated for each task. The procedure consists of four distinct tasks, each structured as follows:

Demographic Survey

To begin, participants in the *non-intervention* and *intervention* groups are asked to complete a demographic survey. As previously mentioned, the main goal of this survey is to collect general information while ensuring that no personally identifiable information is collected. The detailed demographic questions can be found in the Appendix 7.3.3.

Time Allocation: The survey is designed to be completed in approximately 5 minutes.

Surface-level term familiarity task

After completing the demographic survey, the participants move on to the *surface level term familiarity*. The aim here is for participants to identify specialized medical terms that correspond to the popularized medical phrases. As previously mentioned, those in the *intervention* receive assistance through learning material derived from our informal medical entity linking model. In contrast, the *non-intervention* group does not receive such support from our model.

Data: For the *surface-level term familiarity* task, we selected 15 medical concepts from 48 user posts as our primary data (see Section 6.13). These concepts were used to assess participants from both the *intervention* and *non-intervention* groups.

The *non-intervention* group had fewer attempts and the absence of a learning process, we introduced 15 additional medical concepts for this group only. These included concepts like *asthenia, suicidal feelings, chest tightness, lightheadedness, and hypersomnia*. The purpose of including these additional terms was to ensure that participants in both groups spent a *comparable amount of time* on the *surface-level* term familiarity task.

While these additional terms were presented to the *non-intervention* group, they were *excluded* from our *final analysis*. Our evaluation focused solely on the original 15 medical concepts derived from user posts, which were common to both groups.

Time Allocation: The *surface-level term familiarity* task is designed to be completed in approximately 30 minutes.

Concept-level term familiarity task

Once the participants have completed the *surface-level term familiarity task*, they move on to a series of multiple-choice questions. These questions aim to evaluate their understanding of the terms covered in the *surface-level term familiarity*.

Data: For this task, we employ the same set of medical concepts introduced in the surface-level term familiarity task. This consistency applies to both the intervention and non-intervention groups.

Time Allocation: The *concept-level term familiarity* task is designed to be completed in approximately 15 minutes.

Feedback Survey

Finally, we requested all participants to complete a feedback survey, as detailed in Section 6.1.6. This survey is divided into two parts: *Section A* and *Section B*. We provided two different versions of *Section A*, one for the intervention group and another for the non-intervention group. The *Section A* survey for the intervention group pertains to their learning experiences during the *surface-level term familiarity* task. In contrast, the *Section A* survey for the non-intervention group centers on their interest in learning medical terminology. *Section B*, the second part of the survey, is focused on the application of the informal medical EL model in real-world settings. Both groups received the same survey for *Section B*.

Time Allocation: The feedback survey is designed to be completed in approximately 10 minutes.

Overall Time Allocation

Participants are expected to spend around 1 hour in total to complete all assigned tasks.

6.1.8 Hypotheses and variables

For each one of the goals presented in Section 6.1.8, we define the null hypothesis (H_0) and the alternative hypothesis (H_1). For G1, concerning the differences in the level of *surface-level* term familiarity of laypeople, measure by the *surface-level score*, we have the following hypothesis:

- Null Hypothesis (H_0): The medical entity linking model **is not** effective in increasing the *surface-level* term familiarity of laypeople.
- Alternative Hypothesis (H_1): The medical entity linking model is effective in increasing the *surface-level* term familiarity of laypeople.

In this context, the **independent** variable is the *presence of the medical entity linking model* (intervention group vs. non-intervention group), and the **dependent** variable is the level of *surface-level* term familiarity, quantified by the *surface-level score*.

We follow the same approach to define the hypothesis for G2, concerning the difference in the level of *concept-level* term familiarity of laypeople, measured by the *concept-level score*. We have the following hypothesis:

- Null Hypothesis (H_0): The medical entity linking model **is not** effective in increasing the *concept-level* term familiarity of laypeople.

- Alternative Hypothesis (H_1): The medical entity linking model is effective in increasing the *concept-level* term familiarity of laypeople.

The **independent** variable remains the **presence of the medical entity linking model** (intervention group vs. non-intervention group), while the **dependent** variable is the level of *concept-level* term familiarity, measured by the *concept-level score*.

6.2 Experiment Setup

We created two separate advertisements for the *Prolific* platform. The first advertisement was designed to enlist 100 participants for the *intervention* group, while the second advertisement aimed to recruit 50 participants for the *non-intervention* group using the inclusion criteria we mentioned in Section 6.1.2. After agreeing to participate in our study, participants were redirected to one of the two evaluation systems via the Prolific platform, depending on their assigned group.

Participants in the *intervention* group registered at [https://ir-group.ec.tuwien.ac.at/mel/], while those in the *non-intervention* group registered at [https://ir-group.ec.tuwien.ac.at/mediquiz/]. During registration, participants reviewed and agreed to an informed consent form, detailing how their data would be used in the study. After successful registration, participants accessed the system with a unique username and password, where they were presented with the experiment's objectives and guidelines. Participants then followed instructions for the experimental procedure described in Section 6.1.7.

The experimental process began with a demographic survey, followed by a *surface-level* term familiarity task, then a *concept-level* term task. The final task involved providing feedback through surveys, after which participants received a completion code to confirm their participation. They entered this code into Prolific to verify task completion. Figure 6.14 presents the workflow of the evaluation systems.

The goal of this experiment is to assess the effectiveness of an informal medical EL model in supporting laypeople to improve their medical terminology knowledge. To measure this effectiveness, we conducted two types of evaluations:

1. **Objective evaluation:** This focused on both *surface* and *concept-level* term familiarity among laypeople, as described in Section 6.1.3. This evaluation addressed the evaluation goal and hypotheses testing outlined in Sections 6.1.1 and 6.1.8, respectively. The evaluation scores for each task were calculated (see Section 6.1.5). The *surface-level* term familiarity score included: familiarity score (Eq. 6.1 and Eq. 6.2), answer accuracy score (Eq. 6.4) and learning rate estimation (LRE) (Eq. 6.7). The *concept-level* term familiarity score was calculated using Eq. 6.8. These scores were used to answer the evaluation goal.
2. **Participant Feedback:** As described in Section 6.1.7, this involved a short questionnaire filled out by participants. The *intervention* group participants were asked about their learning experience during the *surface-level* term familiarity task, using a 1-5 Likert-type scale (e.g. 1 = strongly disagree, 2 = disagree, 3 = neutral, 4 = agree, and 5 = strongly agree). The *non-intervention* group participants filled out a questionnaire focused on user information and were asked about the perceived importance of having a basic understanding of medical terminology, using the same 1-5 Likert-type scale. The second part of the survey for both groups focused on the idea of implementing the model in social media, with agreement expressed on a 1-5 Likert-type scale.

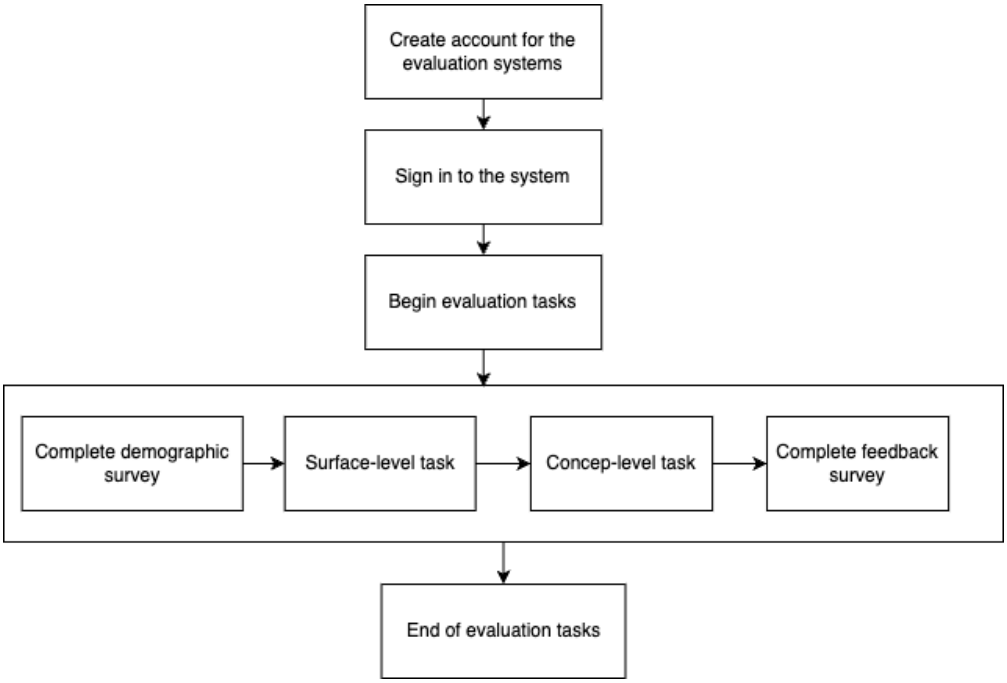


Figure 6.14: Evaluation System Workflow.

6.3 Results and Discussion

6.3.1 Participant Demographics

Out of the 150 participants initially recruited from Prolific, we had to exclude 9 participants (5 participants from the *non-intervention* group and 4 from the *intervention* group). There were several reasons for these exclusions. First, during the initial phase of our user experiments, our evaluation system experienced slow performance. Since we had set a time limit of 30 minutes for the *surface-level term familiarity* task, some participants were unable to complete it within this time frame. Consequently, we decided to exclude these participants from our analysis. This exclusion affected 1 participant from the *non-intervention* group and 4 from the *intervention* group. Second, we identified that 4 participants, who had initially joined the *intervention* group, subsequently participated in the *non-intervention* group. To mitigate potential bias, these participants were excluded from the analysis of the *non-intervention* group. As a result, our final analysis included 141 of 150 participants. Table 6.2 provides the demographic information gathered from both the *non-intervention* and *intervention* groups. The majority of our participants from both groups are young males who possess a higher level of education. Furthermore, 135 out of 141 participants (94%) do not work in healthcare or medical services. In terms of income, most earn an annual income either under $15,000 (18%) or between $30,000 and $49,999 (35 out of 141 participants). The majority of participants (69%) are from the UK and are native English speakers (116 out of 141). Furthermore, 89% of participants are highly engaged with the internet, and 95% perform online searches daily. This high frequency of online searches and reading information in English is also similar among the participants.

Table 6.2: Demographic Information of Non-Intervention and Intervention Groups

Category	Non-Intervention, n(%)	Intervention, n(%)	Total, n(%)
Age			
18 - 21 years old	2 (4)	4 (4)	6 (4)
22 - 34 years old	27 (60)	52 (54)	79 (56)
35 - 44 years old	7 (16)	21 (22)	28 (20)
45 - 54 years old	6 (13)	13 (14)	19 (13)
55 - 64 years old	2 (4)	5 (5)	7 (5)
65 and over	1 (2)	1 (1)	2 (1)
Gender			
Female	18 (40)	19 (20)	37 (26)
Male	25 (56)	76 (79)	101 (72)
Other	1 (2)	0 (0)	1 (1)
Prefer not to answer	1 (2)	1 (1)	2 (1)
Education Level			
Middle school/Junior high school	0 (0)	1 (1)	1 (1)
High school/Senior high school	10 (22)	27 (28)	37 (26)
Diploma/Vocational/Technical school	8 (18)	9 (9)	17 (12)
Bachelor degree	20 (44)	47 (49)	67 (48)
Master degree	7 (16)	10 (11)	17 (12)
Doctoral/PhD	0 (0)	2 (2)	2 (1)
Country			
Australia	1 (2)	1 (1)	2 (1)
South Africa	9 (20)	17 (18)	26 (18)
United Kingdom	30 (67)	67 (70)	97 (69))
United States	5 (11)	11 (12)	16 (11)
Working Area			
Business Management and Administration	5 (11)	3 (3)	8 (6)
Education and Teaching	3 (7)	9 (9)	12 (9)
Engineering	4 (9)	8 (8)	12 (9)
Finance and Accounting	4 (9)	9 (9)	13 (9)
Healthcare and Medical Services	1 (2)	5 (5)	6 (4)
Information Technology (IT)	7 (16)	14 (15)	21 (15)
Media and Entertainment	0 (0)	3 (3)	3 (2)
Sales and Marketing	7 (16)	7 (7)	14 (10)
Other	14 (31)	38 (40)	52 (37)
Income			
Under $15,000	12 (27)	24 (25)	36 (26)
Between $15,000 and $29,999	7 (16)	19 (20)	26 (18)
Between $30,000 and $49,999	11 (24)	24 (25)	35 (25)

Continued on next page

Table 6.2 – *Continued from previous page*

Category	Non-Intervention, n(%)	Intervention, n(%)	Total, n(%)
Between $50,000 and $74,999	7 (16)	14 (15)	21 (15)
Between $75,000 and $99,999	4 (9)	5 (5)	9 (6)
Over $100,000	1 (2)	2 (2)	3 (2)
I do not know	0 (0)	2 (2)	2 (1)
Prefer not to answer	3 (7)	6 (6)	9 (6)
English Proficiency			
Fluent English	8 (18)	16 (17)	24 (17)
Good knowledge of English	0 (0)	1 (1)	1 (1)
Native English	37 (82)	79 (82)	116 (82)
Internet use			
Yes	40 (89)	85 (89)	125 (89)
I do not work/study at the moment, I use Internet only for my personal inquiries	3 (7)	10 (10)	13 (9)
No	2 (4)	1 (1)	3 (2)
Frequency of online search			
A few times a week	1 (2)	5 (5)	6 (4)
Every day	44 (98)	89 (94)	133 (95)
Once a month	0 (0)	1 (1)	1 (1)
Frequency of online search in their own language			
A few times a month	0 (0)	2 (2)	2 (1)
A few times a week	2 (4)	8 (8)	10 (7)
About once a week	0 (0)	1 (1)	1 (1)
Every day	38 (84)	79 (82)	117 (83)
Less than once a month	5 (11)	6 (6)	11 (8)
Frequency of online search or reading information on the Internet in English			
A few times a week	1 (2)	5 (5)	6 (4)
About once a week	1 (2)	0 (0)	1 (1)
Every day	43 (96)	91 (95)	134 (95)

Non-Intervention Group (N=45), Intervention Group (N=96), and Total (N=141).

We also collected additional data on how often participants searched for health information and what kinds of health topics they were interested in. According to Figure 6.15, the *intervention* group has more participants who *never* or *sometimes* search for health information, while the *non-intervention* group more frequently searches *often* and *sometimes*. Furthermore, Figure 6.16 indicates that participants in both groups were mostly seeking information about *Specific diseases or medical problems* and *Certain medical treatments and procedures*. This finding aligns with a survey conducted by the Pew Research Center [Pew11], which measured internet users' interest in health information. Furthermore, participants were asked to read two health-related articles on "Ulcerative Colitis" sourced from *Healthline.com* and an abstract from a research publication [FC14]. Subsequently, we asked participants to self-assess their comprehension of the articles

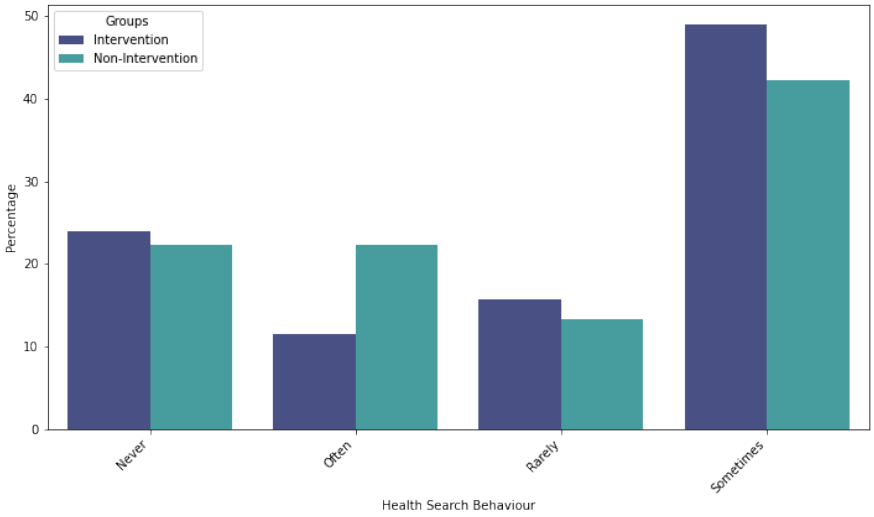


Figure 6.15: Distribution of responses to 'How Often Do You Search for Online Health Information Regarding Your/Your Family's/Friends' Health?'

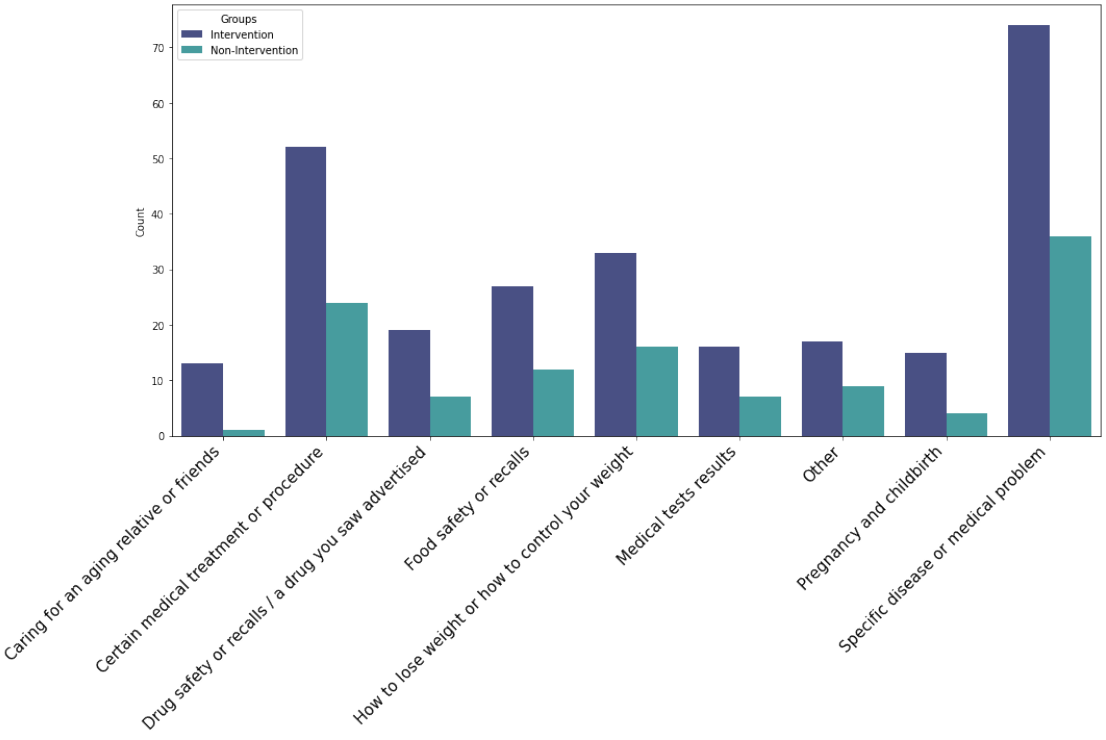


Figure 6.16: Distribution of Type Health Information Search among both groups

and their familiarity with the medical terminology used in both texts, using a scale ranging from 1 (indicating 'do not understand at all') to 5 (indicating 'understand very well'). Although we

did not formally measure their health literacy level, our intention was to gain insights into their ability to understand health-related articles.

The majority of participants in both groups showed that they could understand the article content and medical terminology presented in the "Ulcerative Colitis" articles sourced from Healthline, as shown in Figures 6.17 and 6.18. This result is to be expected as Healthline provides articles with less medical terminology, which makes it easier for participants to understand the content of the articles. In contrast to the first article, the participants from both groups assessed their reading

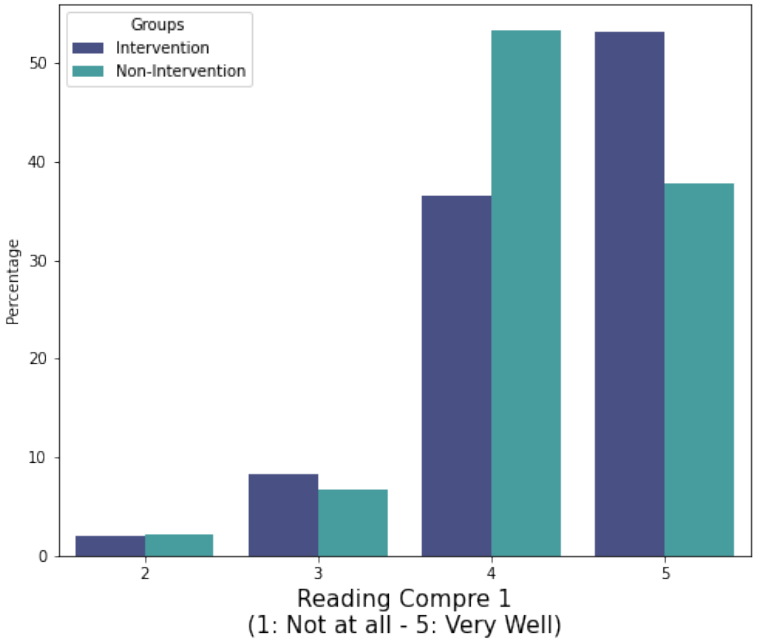


Figure 6.17: Distribution of Reading Comprehension of *Ulcerative Colitis* sourced from Healthline.com on both groups

comprehension as varying from 'do not understand' to 'understand'. In the *intervention* group, nearly 50% of participants reported a good to very good understanding of the article, while almost half of the participants indicated a moderate or limited understanding. In the *non-intervention* group, the results differed slightly, with more than half of the participants expressing a moderate or limited understanding of the article, and approximately 29% reporting a good understanding. Meanwhile, with regard to medical terminology comprehension, as shown in Figure 6.20, nearly 70% of the participants showed either a limited or moderate level of understanding of the medical terminology presented in the abstract publication. We acknowledge that our participants' demographics are not representative of the broader population, especially when considering characteristics associated with low health literacy as described by [HMCRT+18]. The authors [HMCRT+18] stated that people with low health literacy are typically older, possess limited education, have lower incomes, suffer from chronic conditions, and are often non-native English speakers. Furthermore, the majority of participants in both groups showed good comprehension skills when reading health-related articles on medical websites and had moderate levels of understanding of medical terminology in abstract publications.

Nevertheless, our main goal in this user experiment is to evaluate whether our informal medical

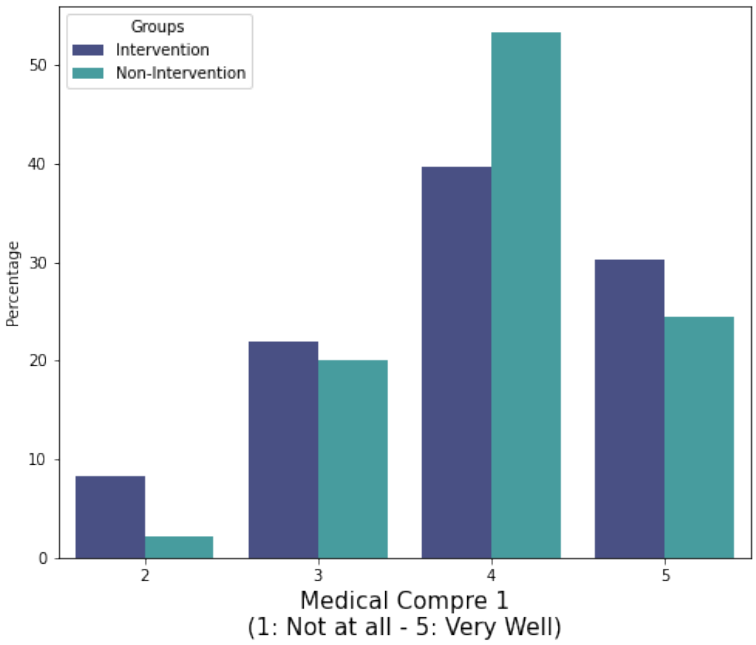


Figure 6.18: Distribution of Medical Terminology Comprehension of *Ulcerative Colitis* sourced from Healthline.com on both groups

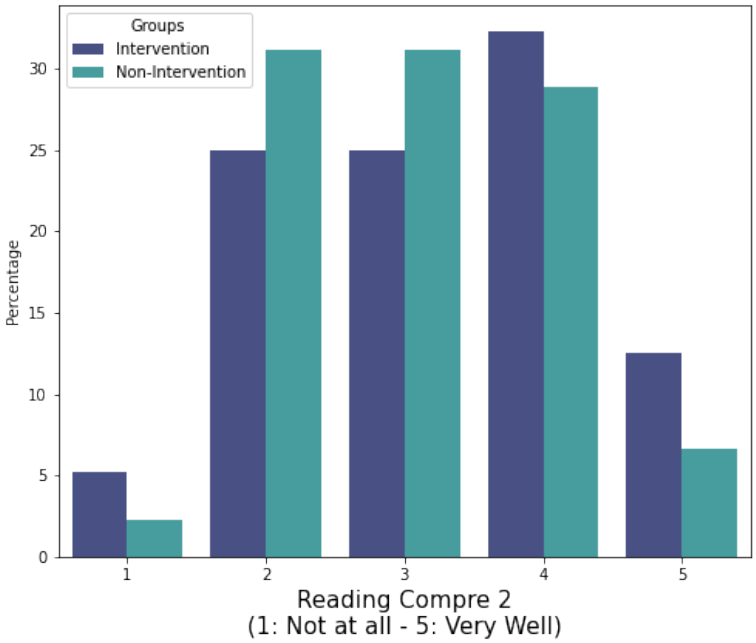


Figure 6.19: Distribution of Reading Comprehension of *Ulcerative Colitis* sourced from the abstract of publication [FC14] on both groups

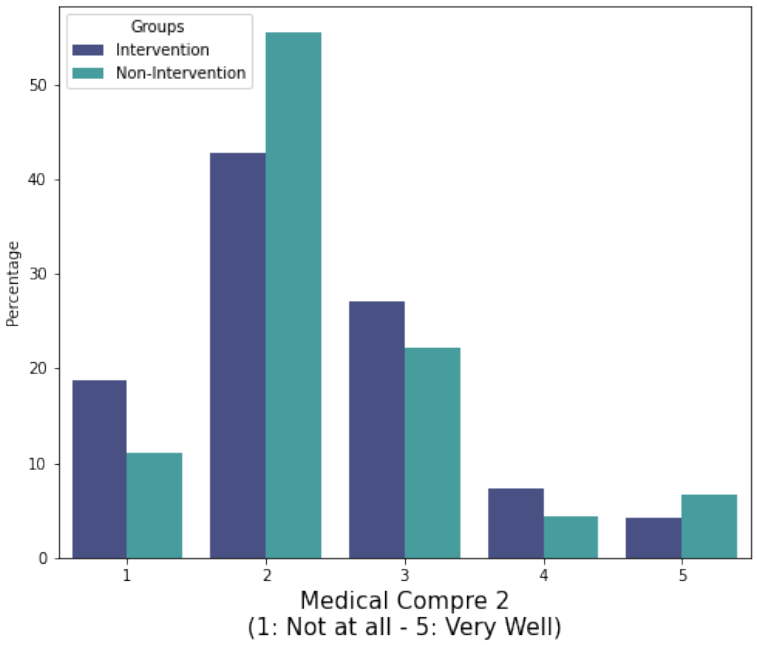


Figure 6.20: Distribution of Medical Terminology Comprehension of *Ulcerative Colitis* sourced from the abstract of publication [FC14] on both groups

EL supports participants in familiarizing themselves with and understanding specialized medical terminology based on the popularized medical phrases commonly used in social media. To summarize, we would like to highlight that the participants in both groups are similar in terms of their characteristics.

6.3.2 Analysis of the Effectiveness of Informal Medical EL Model

In this section, we present the results from objective tests that assess the effectiveness of an informal medical EL model in supporting laypeople to improve their medical terminology knowledge on both *surface-level* and *concept-level* term familiarity. We measured this effectiveness using two evaluation scores: the *surface-level score* and the *concept-level score*. The *surface-level score* measures participants' ability to recognize specialized medical terms based on their popularized medical phrases used in social media. Meanwhile, the *concept-level score* assesses how well participants are able to identify the correct meanings of these terms based on prior tasks.

Analysis Procedure

We started by collecting descriptive statistics on our variables to get an overview of their distribution. We collected the *min, max, mean, median, and standard deviation*. We also used *box plots* to help with the visual analysis of the distribution. We then applied *Levene's test* for homogeneity of variance to assess if each group of the independent variable had the same variance. If Levene's test is significant at $p < .05$, we reject the null hypothesis that the groups have equal variances. For testing our hypothesis, we used *Welch's t-test* for comparing the distributions to

detect statistically significant differences between groups. We are using $p < .05$ for the level of significance and thus reject the null hypothesis if this threshold is met.

Descriptive Statistics

Tables 6.3 present the descriptive statistics for the metrics collected. These metrics were collected for each task (i.e., *surface-level* and *concept-level* term familiarity) across different groups. *Surface-level* and *concept-level* scores were calculated for each participant in both the *intervention* and *non-intervention* groups. These scores were then utilized to assess the effect of informal medical EL to answer our goal mentioned in Section 6.1.1.

The $familiarity$ score is based on the mastery level achieved by participants in both groups. The *mean* $familiarity$ score for the *intervention* group is 86.23 (SD = 14.46), which is higher than the *mean* of 78.67 (SD = 17.28) for the *non-intervention* group. Similarly, the *median* $familiarity$ score for the *intervention* group is 90.0, higher than the *median* of 83.0 for the *non-intervention* group.

We recognize that the $familiarity$ score yields incomparable results due to the different numbers of attempts between groups. To address this, we introduced an additional measure of the *surface-level score*, which allows for comparison between groups. As mentioned in Section 6.1.3, the *intervention* group had three attempts during the *surface-level* term familiarity *testing* phase, while the *non-intervention* group had two. To ensure a fair comparison between groups, we made adjustments to our analysis of *surface-level* term familiarity. We considered only the *first two* attempts of the *intervention* group, matching the number of attempts for the *non-intervention* group. Additionally, we excluded the first attempt of the *intervention* group from our analysis, as it was conducted under controlled conditions where participants received hints to help them grasp the word form of specialized medical terms. These adjustments allowed us to compare *surface-level* scores based on two equivalent attempts for each group, ensuring a more equitable assessment of term familiarity between the *intervention* and *non-intervention* groups.

The *intervention* group performed better than the *non-intervention* group, indicating higher recall ability for correct answers. For the *answer_accuracy*@2 score, the *mean* of the *intervention* group was 89.62 (SD = 11.13), higher than the *non-intervention* group's *mean* of 82.44 (SD = 14.91). The *median answer_accuracy*@2 score of the *intervention* group was 93.0, also higher than the *non-intervention* group's *median* of 83.0. Regarding the *LRE*@2 score, higher scores reflect positive learning rates, indicating that participants were able to correct their answers in the final attempts. The *mean LRE*@2 score of the *intervention* group was 90.64 (SD = 11.55), higher than the *non-intervention* group's *mean* of 83.52 (SD = 14.10). This implies participants in the *intervention* group were more successful in correctly answering *surface-level term familiarity* tasks within two attempts, particularly by getting the correct answer on the last attempt. Furthermore, for the *concept-level score* reveals that the *mean* score for the *intervention* group is 82.11, which is higher than the *non-intervention* group's *mean* score of 62.58. Figure 6.21 presents the visual analysis of the distribution of the scores among the groups.

Hypotheses Testing

We present the results for the hypotheses testing given the hypotheses provided in Section 6.1.8.

RQ1: *Is the medical entity linking model effective in increasing the* surface-level *term familiarity of laypeople?*

Table 6.3: Descriptive statistics for the *Surface-level* and *Concept-level* Scores

Group	Metric	Min	Max	Median	Mean	SD
Intervention	Surface-level					
	$familiarity$	38.0	100.0	90.0	86.23	14.46
	$answer\_accuracy$@2	53.0	100.0	93.0	89.62	11.13
	LRE@2	50.0	100.0	94.0	90.64	11.55
	$answer\_accuracy$@3	56.0	100.0	94.0	90.81	10.44
	LRE@3	56.0	100.0	96.0	92.15	10.42
	$Concept-level$	13.0	100.0	87.0	82.11	14.51
Non-Intervention	Surface-level					
	$familiarity$	23.0	100.0	83.0	78.67	17.28
	$answer\_accuracy$@2	33.0	100.0	83.0	82.44	14.91
	LRE@2	37.0	100.0	88.0	83.52	14.10
	$Concept-level$	13.0	100	67.0	62.58	23.24

*In metrics like $answer\_accuracy$@k and LRE@k, @k denotes the number of attempts.

We employed the *surface-level* score to evaluate if the model improved *surface-level* term familiarity of laypeople. The *surface-level score* consisted of three components: $familiarity$, $answer_accuracy$, and LRE (Learning Rate Estimation) scores. As mentioned earlier, for the *intervention* group, we only considered the answer accuracy and LRE scores from their first two attempts. This allowed us to compare the *intervention* group's performance with the *non-intervention* group, which had only two attempts for testing surface-level term familiarity, denoted as $answer\_accuracy$@2 and LRE@2.

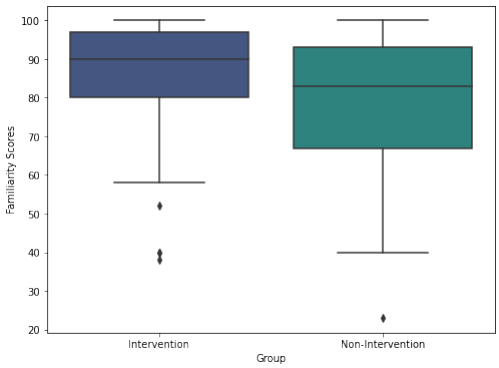
Table 6.4 presents the results of Levene's test and Welch's t-test. For the $familiarity$, Levene's test showed no significant difference, indicating equal variance among groups. However, Welch's t-test revealed a statistically significant difference in $familiarity$ among the groups. For $answer\_accuracy$@2 and LRE@2, both Levene's test (indicating unequal variance) and Welch's t-test demonstrated statistically significant differences. These results suggest a significant difference in *surface-level* term familiarity among the groups for these metrics.

Table 6.4: Levene's and Welch's t-test for the *surface-level score*

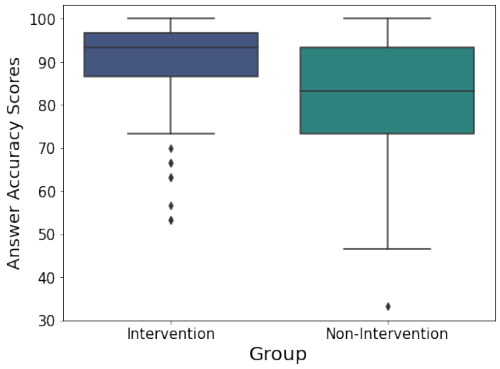
Metric	Levene Sig.	Welch Sig.
$familiarity$	.1107	.00788
$answer\_accuracy$@2	.0128	.0058
LRE@2	.0499	.0045

RQ2: *Is the medical entity linking model effective in increasing the* concept-level *term familiarity of laypeople?*

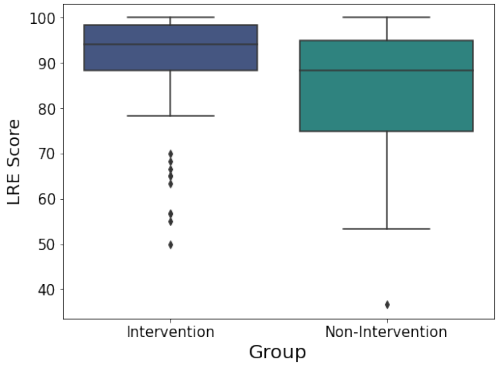
We also report the results of Levene's test and Welch's t-test for the *concept-level score*. Levene's



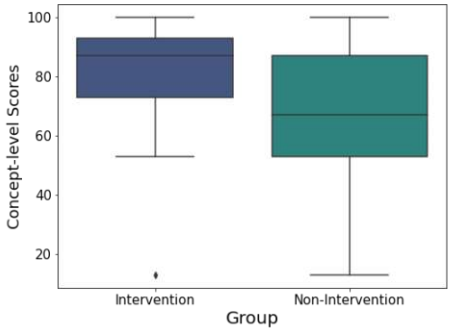
(a) Familiarity Score



(b) Answer Accuracy Scores



(c) LRE Scores



(d) *Concept-level* Scores

Figure 6.21: Distribution of Scores Among The Groups

test showed a significant difference in variances (p = .0009), indicating unequal variances. Welch's t-test revealed a statistically significant difference in the mean values of the *concept-level score* (p = 3.22e-06), indicating a significant mean difference among the groups.

Discussion of RQ 1

In this section, we present a discussion of the evaluation of the results to assess the effectiveness of our informal medical entity linking model for laypeople in supporting laypeople in *surface-level* term familiarity.

There was a statistically significant mean difference in the $familiarity$ score with participants in the *intervention* group than participants in *non-intervention* group ($p - value$ = .00788). These results suggest that the participants in the *intervention* group achieved mastery of medical concepts, which can be considered to improve the specialized medical terminology retention than the one in the *non-intervention* group.

Furthermore, we evaluate whether our informal medical EL model contributed to the participants in the *intervention* group obtaining higher $familiarity$ score in term familiarity tasks than the participants in the *non-intervention* group. We first calculated the threshold as a cutoff point score to categorize each of the participant's $familiarity$ score into two categories: "high level"

and "low level" scores using Eq. 6.9, which the threshold = 84. It means that the mean average of all participants' *familiarity* score across groups is 84. Table 6.5 presents the distribution of the familiarity scores based on the calculated threshold across the *intervention* and *non-intervention* groups. This distribution suggests that approximately 64% (62 out of 96) of the participants in the *intervention* group achieved high-level *familiarity* score compared to those in the *non-intervention* group (19/45, 42%).

Table 6.5: Distribution of High and Low *familiarity* score Among Intervention and Non-Intervention Groups

	Intervention	**Non-Intervention**	**Total**
High Familiarity Score	62	19	81
Low Familiarity Score	34	26	60
Total	96	45	141

A threshold score of 84 was used to classify familiarity scores as high or low.

A chi-square test was conducted to assess the association between the effectiveness of the informal medical EL model and participants' ability to achieve high familiarity and understanding scores. The association was found to be statistically significant $X^2(1, N = 141) = 5.4, p = .02$, indicating that participants in the *intervention* group were 2.5 times more likely to have higher *familiarity* score than those in the *non-intervention* group.

The statistically significant difference in *familiarity* score between the *intervention* and *non-intervention* groups is suggesting that the intervention may have been effective in improving participants' familiarity with medical concepts. However, we acknowledge potential biases and reliability issues due to the different mastery level scales used in the two groups. The *intervention* group had four mastery levels, while the *non-intervention* group had only two, making a direct comparison of mastery between the groups less reliable.

Furthermore, the *threshold* score used to separate the scores into "high" and "low" levels has limitations. We used the weighted average to calculate the mean average of all the participants' *familiarity* score. This formula is often used in aggregating students' scores [MKA13]. In the future, the standard threshold value for contrasting score levels, such as the Angoff method [Ang71], which uses a group of experts to judge the difficulty of each exam item to determine the cut-off score. Alternatively, Bloom's cut-off point [Blo68] can be used as the standard value. It is worth noting that, according to Barman [Bar08], referring to the work of Kane et al. [Kan94], there is no perfect method to determine a cut score on a test.

To overcome the reliability issues of the *familiarity* score arising from the different mastery level scales used in the *intervention* and *non-intervention* groups, we propose to use answer accuracy and LRE as alternative metrics. These metrics offer a more reliable comparison of *familiarity* score between the groups by only considering the scores from the first two attempts for both groups, thus ensuring a consistent basis for comparison.

There was a statistically significant mean difference in the *answer_accuracy*@2 scores. These results suggest that participants in the *intervention* group were able to recall the correct specialized medical terms from the corresponding popularized medical terms found in the annotated user posts (Mean = 89.62, SD = 11.13), compared to participants in the *non-intervention* group (Mean = 82.44, SD = 14.91). Based on the box plots 6.21b, the *answer_accuracy*@2 scores for the *intervention* group are distributed above the median, whereas the scores for the *non-intervention*

group are more spread out around the median. This suggests that the informal medical EL model may assist laypeople in recalling specialized medical terms associated with popularized medical phrases, compared to those who did not receive the intervention.

Moreover, we quantified participants' learning estimation rate using the LRE score. As we did not formally assess functional health literacy using established measures such as REALM [DLJ+93] or TOFHLA [PBWN95], we assumed participants had no prior medical knowledge. The LRE score ranges from 0 to 100, with 100 indicating optimal learning (correct answers for all 15 concepts on both attempts) and 0 indicating no learning (incorrect answers after multiple attempts). Scores between these extremes represent varying degrees of correct responses across concepts and attempts.

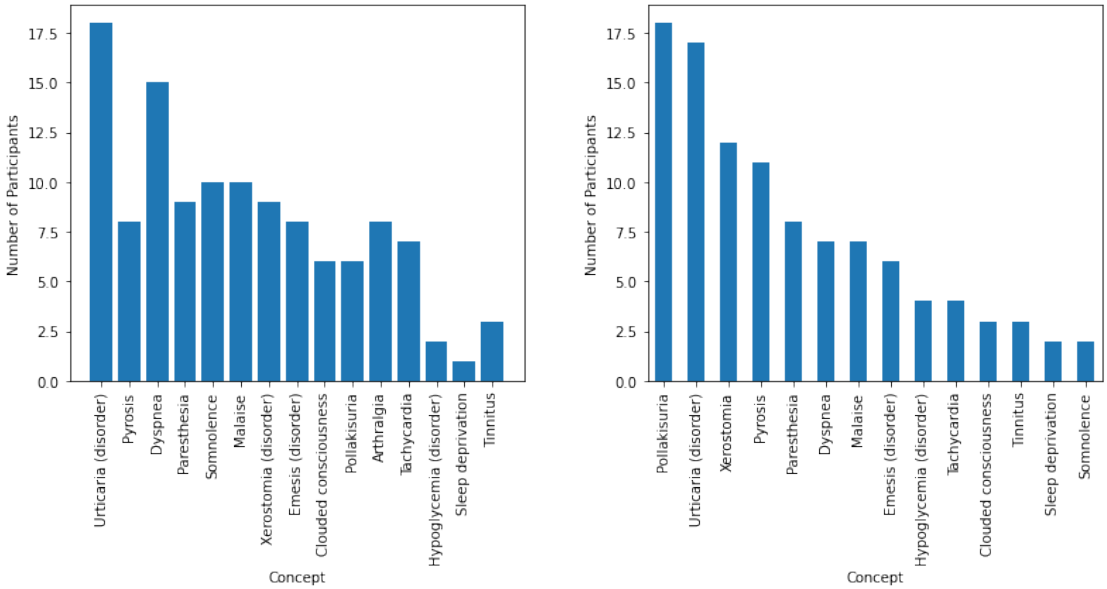
Statistical analysis demonstrated a significant difference in mean (LRE@2) scores between groups. The *intervention* group demonstrated a higher mean (LRE@2) score (M = 90.64, SD = 11.55) compared to the *non-intervention* group (M = 83.52, SD = 14.10), indicating that our medical EL model more effectively improved participants' ability to provide correct answers in their final attempts. The *intervention* group's minimum score (Min = 50.0) exceeded that of the *non-intervention* group (Min = 37.0), indicating the lowest-performing participant(s) in the *intervention* group able to answer correctly in their last or both attempts. This suggests our medical EL model effectively supports the acquisition of medical terminology knowledge, as measured by participants' performance in their response on the last attempts, compared to *non-intervention* group.

Figure 6.22 illustrates the distribution of medical concepts with lower (LRE@2) scores across both groups. (LRE@2) scores below 0.5 indicate concepts for which participants were unable to provide correct answers in their final or both attempts, highlighting where learning medical terminology remained challenging.

The top 5 medical concepts where participants in the *intervention* group demonstrate lower LRE scores are: *Urticaria, Pyrosis, Dsypnea, Paresthesia, and Somnolence*. On the other hand, for the *non-intervention* group, the top 5 medical concepts where participants have lower LRE scores are: *Pollakisuria, Urticaria, Xerostomia, Pyrosis, and Paresthesia*. *Urticaria* and *Pyrosis* were the concepts with the highest number of participants showing low LRE scores in both groups. The other top concepts with lower LRE scores differed between the *intervention* and *non-intervention groups*.

Outliers in *Intervention* Group: Although the overall $answer_accuracy$@2 and LRE@2 scores of the *intervention* group was higher than the *non-intervention* group, there were some participants identified as outliers. However, about 8 participants were consistently identified as outliers, scoring below the lower quartile on both measures. This prompted an investigation into why the model failed to improve their ability to recognize specialized medical terms. Our observation revealed that during the initial learning phase, participants could answer correctly with the hints provided. However, some participants struggled with certain terms even during this phase.

In the subsequent testing phase, these same participants had difficulty recalling certain medical concepts, despite having answered correctly during the learning phase. Additionally, some participants were unable to answer correctly even after multiple attempts (i.e. 2 attempts). Based on their LRE@2 scores, we found that some participants were also outliers with scores below the lower quartile, indicating they could answer some medical concepts correctly on the last attempts.



(a) Medical Concepts with Low LRE score from *intervention* group

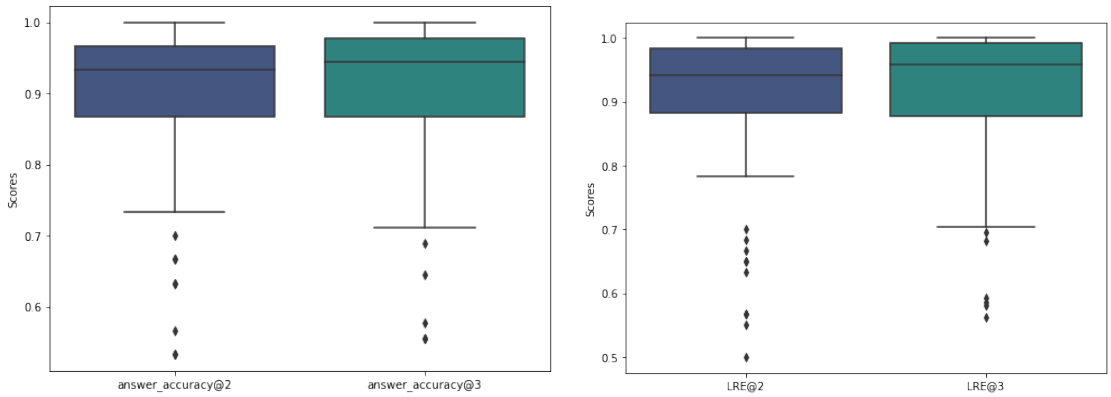
(b) Medical Concepts with Low LRE score from *non-intervention* group

Figure 6.22: Distribution of Medical Concepts with Low LRE Score Among the Groups

We also examined their demographic information. According to [HMCRT+18], laypeople with low health literacy tend to be older, have limited education, have lower incomes, have chronic conditions, and are often non-native English speakers. However, our participants varied in their demographics. Their education levels ranged from high school to master's degrees, their ages from 22-34 to 35-44 years old, and all were native English speakers. Despite this, two participants self-assessed their reading and medical comprehension abilities as moderate for both provided articles during the survey.

Impact of Number of Attempts Within *Intervention* Group: We conducted an additional analysis on the intervention group to determine if increasing exposure to the medical entity linking model was effective in increasing *surface-level* term familiarity of laypeople. We used a paired t-test for hypothesis testing. There was a statistically significant mean difference (p value = 1.086e-06) within participants between the *answer_accuracy*@2 scores and *answer_accuracy*@3. *answer_accuracy*@3 scores (Mean = 90.81, SD = 10.44) were higher than *answer_accuracy*@2 (Mean = 89.62, SD = 11.13). This findings suggests that increasing exposure (i.e., allowing more attempts) to the medical entity linking model improved participants' ability to correctly identify the word-form of specialized medical terms.

Additionally, there was also a statistically significant difference in the *LRE*@3 scores compared to *LRE*@2 scores (p value = 0.0002). The mean score of *LRE*@3 was 92.15 (SD = 10.42), higher than the *LRE*@2 mean score of 90.64 (SD = 11.55). This means that there was an improvement in participants' ability to answer with the correct specialized medical terms in the last attempts. These findings suggest that increasing the exposure from the medical entity linking model may enhances laypeople's familiarity of medical terminology. Figure 6.23 shows the visual analysis of the score distribution within *intervention* group.



(a) *ans_acc*@2 vs *ans_acc*@3 Score (b) *LRE*@2 vs *LRE*@3 Score

Figure 6.23: Distribution of Scores with Different Number of Attempts for *intervention* group

Discussion of RQ 2

In this section, we present a discussion of the effectiveness of our informal medical entity linking model for laypeople in supporting laypeople in *concept-level* term familiarity.

There was a statistically significant difference in the *concept-level scores* among the groups. The *mean* score of the *intervention* group is 82.11 (SD = 14.51), which is higher than the *non-intervention* group's *mean* score of 62.58 (SD = 23.24). This indicates that the scores for both tasks in the *non-intervention* group are more widely spread around the mean when compared to the participants in the *intervention* group.

We argue that participants who received support from the informal medical EL model achieved significantly higher scores than those who did not receive support from the model. The effect of the model is particularly evident in the *concept-level* term familiarity task. This means that the intervention of the informal medical EL model during the *surface-level* term familiarity task was able to improve knowledge of medical terminology so that participants in *intervention* group achieved significantly higher *concept-level scores* compared to participants in the *non-intervention* group. In other words, the participants who received support from the model were better able to identify the meaning of the introduced terms from the *surface-level* term familiarity task. This helped the participants in *intervention* group improve their vocabulary retention not only at the word-form level but also at the concept level, demonstrating their ability to grasp the meaning of their specialized counterparts.

Furthermore, similar to the $familiarity$ score in *surface-level* term familiarity, we also investigate whether the model could lead to higher *concept-level score*. Following the same step, the *concept-level score* is categorized into "high" and "low" levels based on the Eq. 6.9. This threshold is determined by calculating the mean average of all participants' concept-level scores across both groups, which equals 76.

Table 6.6 presents distribution of the *concept-level scores* between the *intervention* and *non-intervention* groups (*threshold = 76*). This distribution shows that approximately 70% (68/96) of the participants in the *intervention* group achieved a high level of identifying the definition of the medical terms, compared to 31% (14/45) in the non-intervention group.

Table 6.6: Distribution of High and Low *concept-level score* Among Intervention and Non-Intervention Groups

	Intervention	**Non-intervention**	**Total**
High Understanding Score	68	14	82
Low Understanding Score	27	31	58
Total	95	45	140

A threshold score of 76 was used to classify understanding scores as high or low.

A chi-square test was conducted to assess the association between the effectiveness of the informal medical EL model and participants' ability to achieve high *concept-level score*. The statistical results demonstrated that those in the *intervention* group were 5.4 times more likely to have higher *concept-level score* compared to those in the *non-intervention* group, $X^2(1, N = 141) = 18.3, p = 1.92e - 05$.

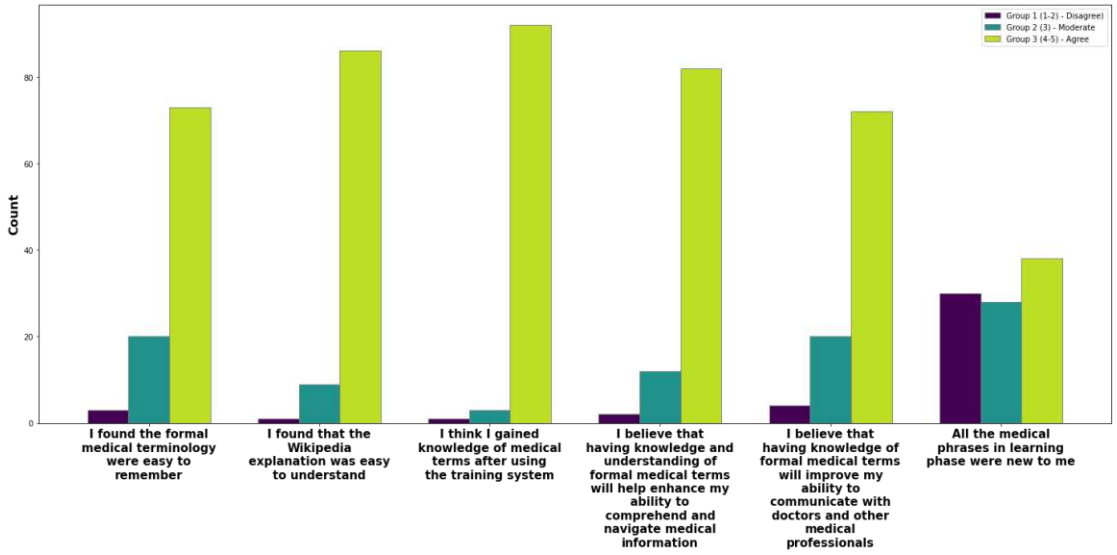
These results suggest there is evidence for an association between the *intervention* from the informal medical EL model and the improvement in *concept-level scores*. The findings demonstrate that the model was able to support participants in the *intervention* group in memorizing and grasping the meaning of the specialized medical terms while administering the *surface-level* term familiarity task, compared to the participants in the *non-intervention* group.

6.3.3 Participant Feedback

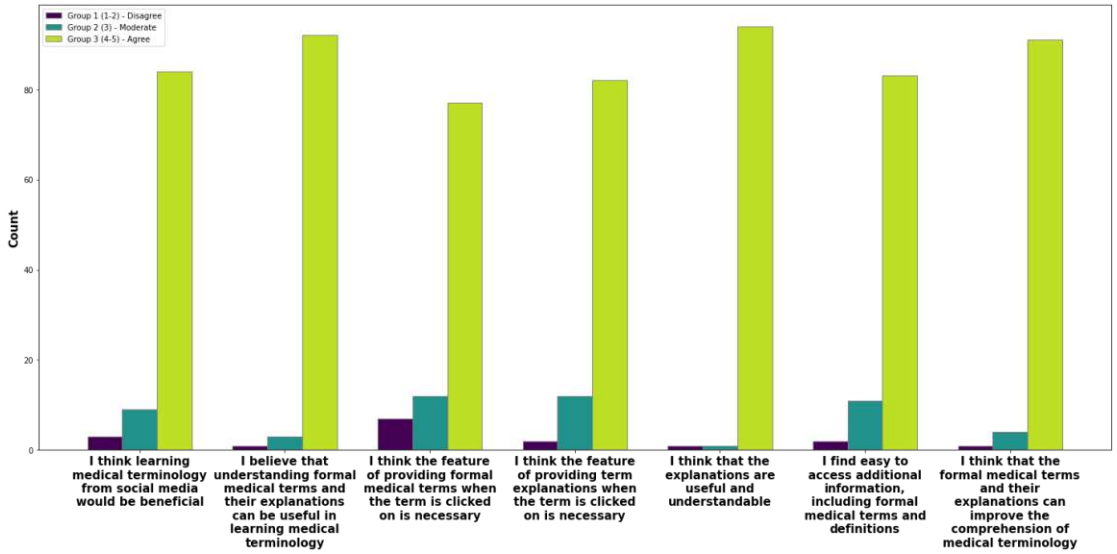
We conducted a two-section feedback survey. In the first section (**Section A**), distinct questionnaires were given to the intervention and non-intervention groups. The intervention group's survey aimed to evaluate their learning experience. In contrast, the non-intervention group's survey assessed their interest in learning formal medical terminology. The feedback from each group is detailed below. In the second section (**Section B**), similar questionnaires were given to both groups. **Section B** aimed to collect feedback on the idea of implementing our informal medical entity linking model in social media, showcasing its real-world application. In **Section B**, potential future work was introduced, such as the application of the proposed system to real-world scenarios. As a feature in social media platforms like AskAPatient, participants could access detailed information about everyday medical terms found in posts. This feature would showcase the formal medical terminology and its explanation, aiding in better understanding.

***Intervention* Group Feedback Survey** In **Section A**, participants shared their feedback on their learning experiences during a task session. They rated their learning experiences on a scale: 1 (Strongly Disagree) to 5 (Strongly Agree). We categorized the response options into three groups: *1-2 (Disagree)*, *3 (Moderate)*, and *4-5 (Agree)*.

The feedback results (shown in Figure 6.24a) indicate that approximately 76% (73 out of 96) of participants found the specialized medical terminology easy to remember, and 89% (86 out of 96) found Wikipedia's explanations easy to understand. Furthermore, nearly all participants (96%, 92 out of 96) believed that acquiring knowledge of medical terms would enhance their ability to comprehend medical information. Moreover, around 85% (82 out of 96) of participants felt that this knowledge would improve their ability to communicate effectively with healthcare professionals. Additionally, 75% (72 out of 96) of participants indicated that having a grasp of



(a) Feedback Survey **Section A**



(b) Feedback Survey **Section B**

Figure 6.24: Distribution of Responses from Feedback Survey from the *Intervention* Group

specialized medical terminology would enhance their ability to navigate medical information. Lastly, around 39% (38 out of 96) of participants were unfamiliar with the medical terms presented, suggesting that these terms were new to them. About 29% had some familiarity, while the rest were confident in their knowledge. In **Section B** (shown in Figure 6.24b), participants were asked their opinion about the integration of the informal medical EL in social media (i.e. AskAPatient).

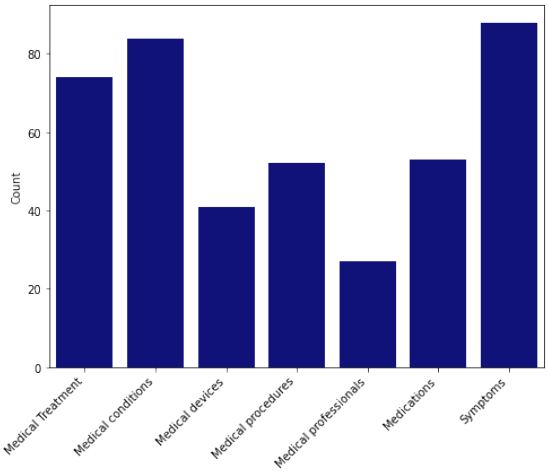


Figure 6.25: Feedback Survey **Section B** - Distribution of responses to the question 'Which types of entities would you be interested in learning more about to enhance your understanding of medical terminology' from the *Intervention* Group

Participants reported that incorporating specific popularized medical phrases with detailed information, such as their corresponding specialized medical term and its explanation, could enhance their knowledge of medical terminology, and found the incorporation of this learning process in social medical settings beneficial for them. Participants appreciated this feature, finding it useful and easy to navigate. Figure 6.25 further details the types of medical topics the participants were most interested in learning about, with symptoms, medical conditions, and treatments being the top three choices among many participants. Finally, Figure 6.26 shows that many participants selected *WebMD, National Institutes of Health (NIH), Centers for Disease Control and Prevention (CDC)* as additional sources to help them learn medical terminology.

***Non-Intervention* Group Feedback Survey** A different feedback survey on **Section A** was given to the *non-intervention group* (results shown in Figure 6.27a and Figure 6.28). The format of the feedback remained similar to that of the other group. However, we modified **Section A** because participants in this group did not receive assistance from our system. For this group, we gathered information on their opinions about the importance of having a basic knowledge of medical terminology and identified their needs to enhance their medical terminology knowledge. Most of the participants expressed that understanding medical terminology is important for them. Nearly all participants (91%, 41 out of 45) responded that it is important to be familiar with medical terms. However, only 47% (21 out of 45) believe that a basic understanding of medical terms is important. Additionally, 69% (31 out of 45) were interested in receiving additional information to understand and learn medical terminology (77%, 35 out of 45).

Furthermore, Wikipedia emerged as their top choice for learning medical terminology(Figure 6.28), which aligns with our proposed informal medical EL model that uses Wikipedia as a source for explaining specialized medical terms. (Section 1.1) as an overview. This finding is consistent with an earlier study by Polepalli et al. [PRHB+13], which highlighted that Wikipedia significantly improves the readability of Electronic Health Records (EHR) notes, making it suitable for laypeople to learn medical terminology.

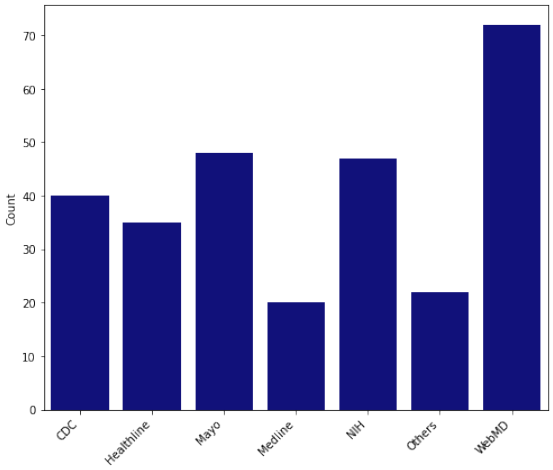


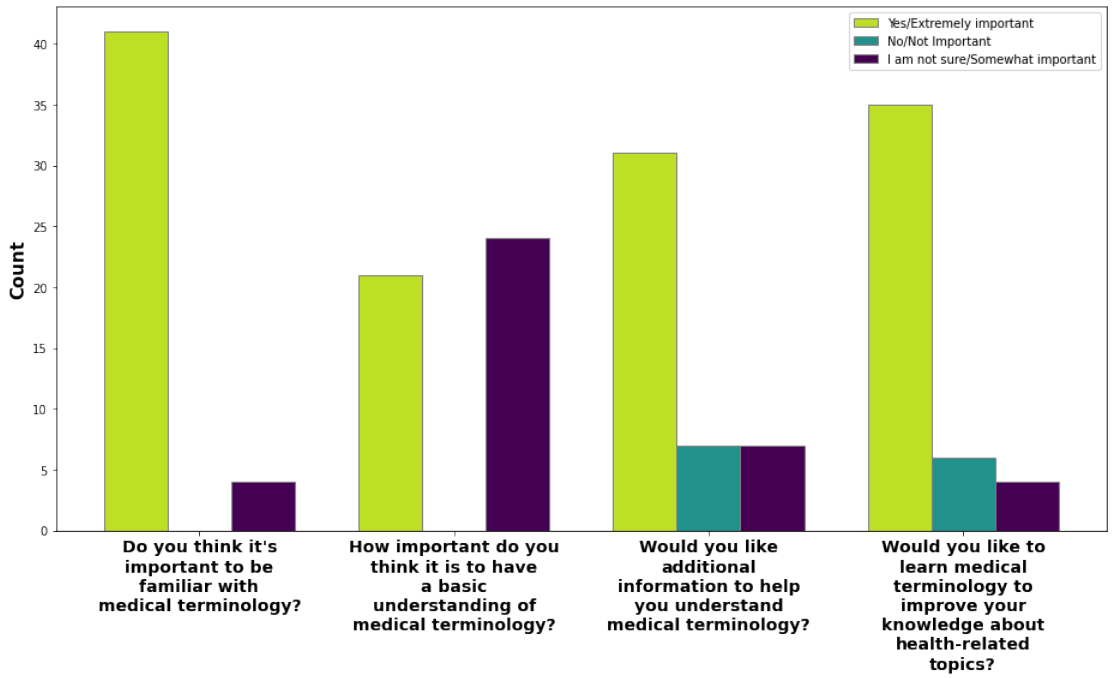
Figure 6.26: Feedback Survey **Section B** - Distribution of responses to the question 'Which additional sources of information would you like to have that helps you learn medical terminology from social media?' from the *Intervention* Group

We then collected the same questions for feedback survey **Section B** with the *intervention* group (Figure 6.27b). For the *non-intervention* group, we demonstrated the possible use of the proposed informal medical EL to aid their learning of medical terminology in a social media setting, although it was not utilized during their evaluation tasks. The feedback from the *non-intervention* group for **Section B** showed similarities to that of the *intervention group*. Most participants in the *non-intervention* group also agreed on the benefits of having basic knowledge about medical terminology (42/45 participants) and considered the feature of specialized medical terms (31/45 participants) and explanation necessary (38/45 participants). Wikipedia's explanations were reported as easy to follow and understand by the *non-intervention* group participants as well (43/45 participants). Additionally, many participants expressed interest in learning more about *symptoms, medical conditions, and treatments*, as shown in Figure 6.29. Websites like WebMD, NIH, CDC, and Healthline were identified as additional resources that could enhance their medical terminology knowledge (6.30).

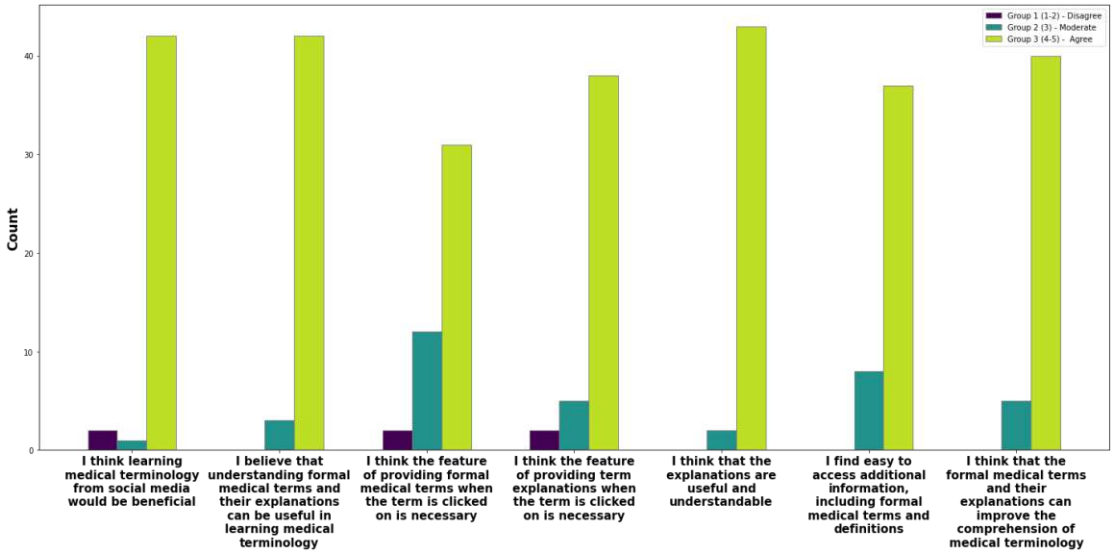
The feedback from both *intervention* and *non-intervention* groups highlighted the importance of understanding medical terminology. Both groups acknowledged the potential of the proposed informal medical EL model and additional features to enhance their medical terminology knowledge. Notably, Wikipedia is highlighted as a preferred learning source by the participants in the *non-intervention* group, showing its helpfulness in aiding medical terminology knowledge. Moreover, a majority of participants from both groups expressed a keen interest in learning about medical concepts such as *symptoms, medical conditions, and treatments*. In addition, resources like WebMD, NIH, CDC, and Healthline were identified as additional sources for enhancing their knowledge of medical terminology.

6.3.4 Findings

We demonstrated the effect of linking informal medical terms to formal medical terms to enhance functional literacy. This specifically improves medical terminology knowledge. We assessed this



(a) Feedback Survey **Section A**



(b) Feedback Survey **Section B**

Figure 6.27: Distribution of Responses from Feedback Survey from the *Non-Intervention* Group

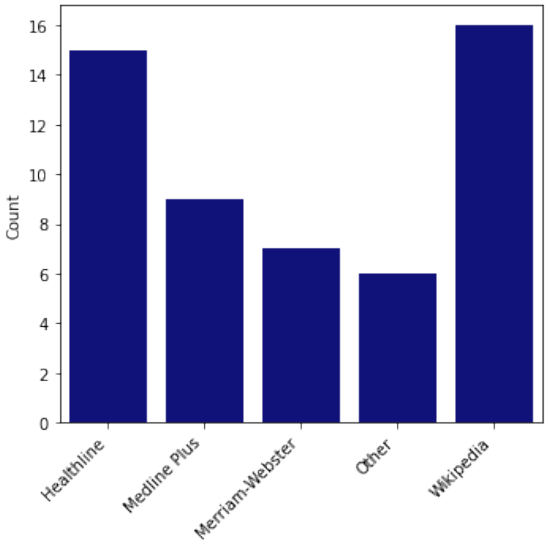


Figure 6.28: Feedback Survey **Section A** - Source of Health Information: Distribution of Responses from the *Non-Intervention* Group

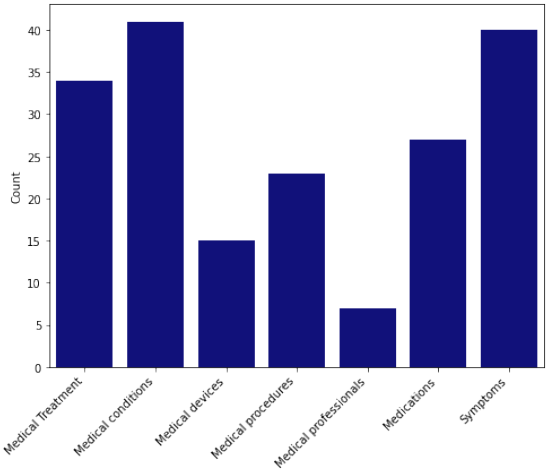


Figure 6.29: Feedback Survey **Section B** - Distribution of responses to the question 'Which types of entities would you be interested in learning more about to enhance your understanding of medical terminology' from the *Non-Intervention* Group

by measuring participants' ability to identify the word-form of specialized medical terms based on popularized medical terms and identify the meaning of the specialized medical terms.

The experiment results have shown that participants in the *intervention* group, with the support of our informal medical EL, may have gained knowledge in medical terminology, as evidenced by improved scores in both *surface-level* and *concept-level scores*, compared to those in the *non-intervention* group. This implies that the proposed model has the potential to enhance knowledge of medical terminology, although challenges remain, particularly for certain participants.

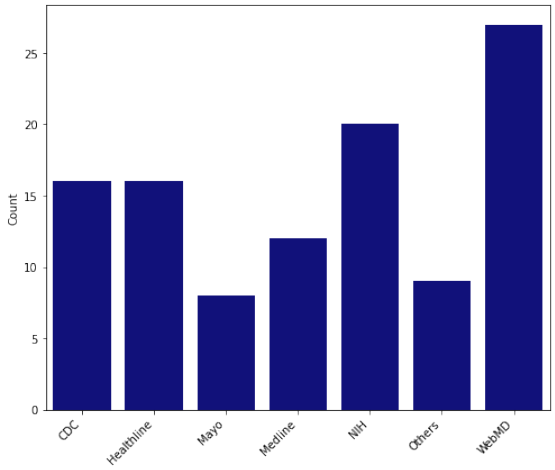


Figure 6.30: Feedback Survey **Section B** - Distribution of responses to the question 'Which additional sources of information would you like to have that helps you learn medical terminology from social media?' from the *Non-Intervention* Group

These findings align with a study by [LWY19], where participants had access to NoteAid [PRHB+13], an NLP system that provides simplified definitions for medical terms in electronic health records (EHR) notes. According to their research, NoteAid improved participants' ability to comprehend EHR notes. We propose that exposing laypeople to popularized medical phrases on social media and then providing them with the corresponding specialized medical terms and their explanations could potentially enhance their ability to comprehend medical terminology.

Furthermore, this improvement may not be limited to the simplification of medical terminology, as found in [LWY19]. Even when faced with more challenging tasks that involve progressing from simple to complex terms, the effectiveness remains consistent. This is supported by our experimental results, where we used popularized medical phrases from social media to introduce corresponding specialized medical terminology with explanations, aiming to help laypeople learn medical terminology.

The recommendation from the previous research is that medical professionals are advised to communicate with patients by avoiding medical jargon or medical terminology [GPH+22, ALL+20, LWY19]. However, with the increasing availability of health-related information on the internet and improved access to online resources, we cannot disregard the possibility that laypeople may have some background or knowledge of medical terminology. As mentioned in [FBNJ16a], social media plays an important educational platform in patient education. By leveraging social media platforms, we can provide detailed information and guidance to laypeople. This includes introducing and explaining specialized medical terms, building upon the popularized medical phrases they may already know or understand from social media. This approach could potentially contribute to improving their comprehension of medical terminology and increasing their functional literacy.

In addition, most of participants also agree that having basic knowledge of medical terms, could improve their ability to comprehend the medical information, as well as their navigational skills on searching the information. We gathered feedback about the future idea of incorporating the informal medical EL model as a feature in social media, and many participants responded

positively. Some comments included:

> **User A**: *I think the system is very helpful and clever. The more you can tear down the barriers of entry to knowledge - in this case being the inconvenience of having to go off-site and do a separate search - the better for literally everyone.*
>
> **User B**: *This was very interesting to learn different words and medical terms and meanings. Sometimes it is hard to understand medical terminology and apply into general symptoms, so having a bit of a better understanding helps ease the mind and help to understand and cope with symptoms. This is a very useful system and makes medical terminology much better to understand.*
>
> **User C**: *I think this is a wonderful idea to help people learn as they use social media and the likes. Over time people will surely learn more about the medical profession and terms used to better protect themselves and get correct diagnosis more promptly. Excellent system & I would find it greatly helpful. I learned many ailment & various symptoms from listening to my dad over the years and picked it up somewhat. For those who don't hear these terms will struggle to describe their health issues, therefore it would be great to see this system deployed.*

While many participants appreciated the idea, some expressed concerns about potential problems such as misinformation, misdiagnosis or the misuse of medical terms, which can lead to self-diagnosis and false conclusions. Here are their comments:

> **User D**: *My concern with this approach is that the social media user may be misdiagnosing or misusing the terms, leading to the reader making the wrong conclusions. The definitions may reinforce false conclusions*
>
> **User E**: *The system above (definition of highlighted word) would be very useful but has the potential for misunderstanding and self-diagnosing.*
>
> **User F**: *I love this system. but I think in a different way it could be bad, because people on social media could exaggerate symptoms. ..*

Additionally, some participants raised concerns about the use of Wikipedia and suggested considering more reliable sources for explaining specialized medical terms:

> **User G**: *I think using Wikipedia as a source of reference might not be the most reliable source as anyone can edit those pages. An official medical website might be a better choice as source for information.*
>
> **User H**: *I think it is very convenient to have access to an explanation of medical terms. But I wouldn't be using Wikipedia as the source because that can be edited by anyone, the source should be from trusted source(s).*
>
> **User I**: *Wiki gives a nice simple overview in layman's terms, it is perfect for basic knowledge but nothing in depth, for that i'd go to a more medical based source*
>
> **User J**: *Addition of information from the NHS website would be useful as many people use and rely on this as a source of information. ...*

6.3.5 Limitations

We recognize the limitations of this study. First, we acknowledge that our choice of evaluation design and metrics may not fully capture the effectiveness of the model in aiding laypeople. Since we aimed to assess whether the model could aid laypeople in learning medical terminology, our evaluation tasks might have been too simplistic, focusing solely on surface-level and concept-level tasks related to medical terms. Additionally, we are aware that the evaluation metrics we used may not adequately measure the participants' learning progress. Our evaluation design was inspired by spaced-repetition techniques commonly used for learning new languages, particularly vocabulary. However, due to technical constraints, we were unable to implement longitudinal studies, which might have provided a more comprehensive understanding of learning progress. Therefore, our evaluation metrics may not effectively capture this progress.

While our hypothesis testing revealed that participants in the intervention group scored higher on surface and concept tasks compared to the non-intervention group, we could improve the evaluation by introducing more advanced tasks. For instance, we could assess participants' ability to comprehend health-related information in connection with the medical concepts they learned during the experiment. Alternatively, we could incorporate established evaluation instruments like ComprehENotes [LWC+18].

Second, utilizing crowd workers means we cannot supervise the participants in our tasks, making it impossible to guarantee they do not use additional information or external resources to complete them. Additionally, we recognize that our time estimate for completing our survey is longer than previous research [LWY19]. This could potentially lead to a lack of motivation among participants over time.

Another limitation of this study is that our participants who completed the survey do not represent laypeople with low health literacy [HMCRT+18], based on the demographic data we collected. However, based on our objectives, this was not a significant concern. We aimed to assess the effectiveness of our proposed informal medical EL model in aiding laypeople with medical terminology knowledge. When comparing the demographics between two groups of participants, they were similarly distributed in terms of socio-demographics. As a result, we can conclude that the informal medical EL is significantly effective in improving participants' medical terminology knowledge, although challenges remain for certain participants.

Additionally, the majority of our participants are native English speakers, as we intentionally chose only from English-speaking countries to minimize language bias. However, we observed that certain specialized medical terms, like *Sleep deprivation*, closely resemble common English vocabulary. This made it easier for participants to identify the relationship between popularized phrases and specialized medical terms. Hence, in the future, we should focus on evaluating the effectiveness of the informal medical EL model with a larger population, such as participants with low English proficiency, to assess if the informal medical EL model could be equally effective in enhancing their medical terminology knowledge.

Currently, the intervention group is twice the size of the non-intervention group. Although both groups share similar characteristics (age, gender, education level, and country), future work should employ balanced randomization techniques to ensure more equal group sizes and maintain similar characteristics across groups.

6.3.6 Threats to Validity

This study faces three potential validity threats. The first threat relates to the subjective nature of selecting medical concepts in Section 6.1.4. This process could potentially limit the diversity of popularized medical phrases used by laypeople. The manual selection of medical concepts and popularized medical phrases introduces subjectivity, which can affect the consistency and reproducibility of the selection process.

In this case, different researchers might select different terms. To counter this, we specifically selected medical concepts with significant differences between their specialized and popularized terms to better address the communication gap between healthcare professionals and laypeople. Moreover, the diverse range of medical concepts demonstrates the effectiveness of the model across various discussions from different medications. For future work, implementing a more standardized process could help reduce subjectivity. One approach could be to analyze the frequency of medical concept usage in social media content. This method would provide a more empirical basis for selecting terms, potentially enhancing the reliability and generalizability of the research findings.

The second threat relates to the self-assessment of reading comprehension and familiarity with medical terminology, as reported in a demographic survey. This self-assessment could introduce bias, as participants might overestimate their abilities due to their confidence levels. Future research could address this limitation by employing standardized tests, such as the TOFHLA [PBWN95] for reading comprehension or the REALM [DLJ+93] for medical terminology comprehension. These tools would allow for a more objective measurement of participants' health literacy.

The third threat arises from the structure of the feedback survey, which consists of two sections designed to evaluate different aspects of the participants' experiences and perceptions. We used different questionnaires for the intervention and non-intervention groups, tailored to their specific experiences. However, responses could be biased, particularly in **Section B** of the feedback survey. Participants' answers might be influenced by their earlier responses in **Section A** or, for the intervention group, by tasks involving surface-level familiarity with terms. Future research could improve this by developing standardized questionnaires with clear, objective criteria to gather feedback on implementing the model in real-world social media settings.

6.4 Summary

Our primary research goal is to develop an informal medical EL model, which aims to assist laypeople in learning medical terminology. We achieve this by leveraging platforms that laypeople are already familiar with, such as social media, and gradually expanding their knowledge of medical terminology based on the popularized medical phrases they are already acquainted with. This approach can further empower them to improve their practical health literacy skills.

To evaluate the effectiveness of this approach, we conducted user experiments, dividing participants into *intervention* and *non-intervention* groups. The *intervention* group used the EL model for assistance, while the *non-intervention* group completed tasks without it.

The evaluation focused on two tasks of medical terminology knowledge :

- *Surface-level* term familiarity: recognize the word-form of formal medical terms that correspond to popularized medical phrases found in social media

- *Concept-level* term familiarity: identify the definition or meaning of a medical term.

Results for *surface-level* term familiarity showed some improvement in the *intervention* group compared to the *non-intervention* group. Statistical significance was observed in the $familiarityScore$, with the *intervention* group achieving higher scores. $Recall@2$ and $LRE@2$ scores were also significantly higher for the intervention group, suggesting improved ability to recall correct word-form of medical concepts.

Additionally, we found that the number of attempts significantly influenced scores within the *intervention* group, with three attempts yielding higher scores than two attempts. However, these results may have potential biases and reliability issues due to the different mastery level scales used in the experiments among the groups. It's worth noting that some participants in the *intervention* group still received lower than average scores, warranting further investigation.

Results for *concept-level* term familiarity task showed that the model was able to support participants in the *intervention* group in memorizing and grasping the meaning of the specialized medical terms while administering the *surface-level* term familiarity task, compared to the participants in the non-intervention group.

Qualitative feedback from participants was positive, highlighting the model's potential to support laypeople in enhancing their medical terminology knowledge. Nevertheless, concerns were raised regarding the reliability of Wikipedia as a source and the potential for misinformation.

There are several directions for future work. Further investigation is necessary to assess the integration of the model with social media platforms, ensuring that the general public can genuinely benefit and feel empowered by its use. The evaluation metrics employed in this study may not have adequately captured participants' progress in medical terminology knowledge. Furthermore, the selected evaluation scenario may have introduced limitations. One potential approach to address these issues is to conduct a single-cohort study utilizing more appropriate evaluation metrics, which could provide a more robust assessment of the model's effectiveness. Moreover, our study participants may not represent the broader population, especially those with limited health literacy or non-English speakers. As a result, we need to expand our user experiment to see if results remain consistent across diverse groups.

CHAPTER 7

Conclusion

This thesis proposes a novel approach to enhancing the functional health literacy of laypeople. It focuses on supporting laypeople in enhancing specialized medical terminology knowledge in social media settings. This goal is achieved by leveraging the benefits of the Medical Entity Linking task. The challenge of the availability of human-labeled informal medical phrases makes this task non-trivial. In this concluding chapter, we revisit the key contributions and research questions addressed by the thesis. Furthermore, we acknowledge the limitations encountered in our research and outline future research directions.

7.1 Research Questions and Contributions

We revisit the research question initially presented in Chapter 1 and provide summaries of the research conducted to address the questions. We developed a method to assist laypeople in enhancing their specialized medical terminology knowledge by first introducing terms they are familiar with from social media, and then presenting the corresponding specialized medical terminology and explanations using Wikipedia articles. This process expands upon the traditional Medical Entity Linking task. Our enhanced method consists of three key phases: (1) Identification of text spans that indicate popularized medical phrases within user-generated content, (2) normalization of these phrases into specialized medical terminology within a knowledge base, for which we employ SNOMED-CT, and (3) linking of these specialized medical terminology to relevant Wikipedia articles.

Our objective is to construct a model that encompasses a wide range of medical terminology, with the aim of introducing these concepts to laypeople effectively. However, we encounter a challenge in the second task, known as the Medical Concept Normalization (MCN) task, due to the scarcity of data. To overcome this challenge, we have proposed two approaches and ask the following question in this context:

RQ 2: How effective are data augmentation and distant supervision in overcoming the problem of data scarcity in Medical Concept Normalization (MCN) task?

We answer this question in two parts. First, we demonstrate the effectiveness of data augmentation in expanding the range of popularized medical phrases used by laypeople to describe specific specialized medical concepts. This is particularly important as specialized many medical concepts have limited examples of popularized medical phrases. We answered this question in detail in Chapter 3.

To address this issue, we proposed several data augmentation techniques. These techniques were designed to mimic the writing style of laypeople.

We applied these augmentation techniques to enrich existing medical concept normalization (MCN) datasets. Additionally, we extended our approach to the task of identifying popularized medical phrases, treating it as a Named Entity Recognition (NER) task. This involved augmenting the training data for NER datasets. The effectiveness of these augmentation methods was evaluated by comparing the performance of MCN and NER models trained on the augmented datasets with models trained on the original training data.

Based on our experiments, we found that for the NER tasks, data augmentation techniques did not improve model performance compared to the baseline model, which was trained only on the original training data. We also attempted to train the NER model with the combined data both of CADEC and MedRed, to evaluate whether it could lead to an increase in the model performance. However, the evaluation results indicated that the NER model did not show improved performance. We analyzed the impact of various augmentation levels (context-level, mention-level, and combined) on our proposed model and a state-of-the-art model. Context-level augmentations clustered near original texts, yielding slight improvements, while entity-level augmentations showed varied similarity but maintained robustness across datasets. Combined augmentations, especially character-based techniques, resulted in wider data spread from original points, potentially explaining the observed performance decrease in both models. These findings underscore the ongoing challenges in using data augmentation to improve NER task performance. In contrast, the data augmentation techniques in MCN tasks can significantly improve the MCN models' performance trained on the augmented datasets compared to those trained on the original data only. Models trained on data augmented by paraphrasing or a combination of augmentation techniques achieved the best performance.

In the second part, we focus on how the distant supervision approach effectively extends the range of medical concepts in public medical concept normalization (MCN) datasets such as CADEC, PsyTAR, and COMETA. This method automatically labels data for MCN tasks using Wikipedia articles as a source of informal medical phrases. These phrases are assigned Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT) concept labels by leveraging the properties in Wikidata. More details are provided in Chapter 4.

Our method involved selecting Wikipedia articles specifically related to medical topics, filtered through Wikidata based on medical properties like SNOMED-CT, International Classification of Diseases, tenth edition (ICD-10), and Unified Medical Language System (UMLS). We then mapped each article to a corresponding SNOMED-CT term. Our approach for extracting popularized medical phrases from Wikipedia focused on the first sentence of the article's summary. We employed techniques to extend noun phrases and to identify abbreviations. Additionally, we utilized Wikipedia's redirect pages and Wikilinks to find synonyms and frequently used phrases associated with Wikipedia titles.

Through this distant supervision method, we extracted 12,101 Wikipedia articles and identified 9,759 unique concepts, with 1,301 also appearing in the public datasets. The effectiveness of this approach was evaluated by comparing the performance of MCN models trained on:(a) a combination of distant supervision data and original training data, (b) original data only, and (c) distant supervision data only.

The results demonstrate that models trained on the combined dataset consistently outperform others. However, the models trained only on distant supervision data have the opposite results, especially in CADEC and PsyTAR datasets. This suggests a language gap between the original test data from these datasets and the popularized medical phrases extracted by our method. Meanwhile, the COMETA dataset appears to be more consistent with what is written on Wikipedia.

Additionally, we unified all MCN datasets (CADEC, PsyTAR, COMETA) with distant supervision data and synonyms from SNOMED-CT. For medical concepts with limited examples, we applied augmentation methods to increase the dataset. The MCN model trained on these unified datasets particularly excelled in the COMETA dataset, demonstrating its effectiveness in processing popularized medical phrases that are not overly simplified. However, our efforts to broaden concept coverage revealed a significant challenge. As we expanded the range of concepts through distant supervision and datasets unification. The proposed MCN model encountered challenges with granularity levels. The model tended to map terms to broader concepts rather than more specific ones. This highlights the complexity of the MCN task, particularly when approached as a multi-classification problem. The challenges lies in managing concept granularity issues, especially when dealing with an expanded range of medical terminology.

Recall that the primary objective of this thesis is to support laypeople in understanding medical terminology as a means to enhance their functional health literacy through the use of the Medical Entity Linking task. In Chapter 5, we developed a model for informal medical entity linking, consisting of three phases: (1) the Named Entity Recognition (NER) phase, which identifies informal medical phrases in the text; (2) the Medical Concept Normalization (MCN) phase, which normalizes each informal medical phrase to its corresponding formal medical terminology found in SNOMED-CT; and finally, (3) the Entity Disambiguation (ED) phase, which retrieves the most suitable Wikipedia article to serve as the source of explanation for the formal medical terminology. To determine the effectiveness of this informal medical entity linking model in supporting laypeople and improving their medical terminology knowledge, we posed the following question:

> **RQ 1**: How effective is medical entity linking, which maps popularized medical phrases to specialized medical terms with explanation, to increase digital health literacy among laypeople?

To answer this question, we conducted user experiments to assess the effectiveness of the model. Chapter 6 details user experiments designed to assess how the informal medical entity linking model enhances digital health literacy, focusing on functional health literacy, specifically in medical terminology knowledge.

This experiment was designed to evaluate how well the model helps laypeople learn specialized medical terminology. For this user experiment, we divided the participants into two groups: the *intervention* group and the *non-intervention* group. Participants in the *intervention* group used the informal medical entity linking model to complete the assessment tasks, while the *non-intervention* group completed the tasks without any support from the model. The evaluation

tasks are limited to two tasks: (1) *surface-level* term familiarity, and (2) *concept-level* term familiarity.

Demographic surveys indicated similar characteristics among participants in both groups. The effectiveness of the model was evaluated by comparing improvements in scores for each task between the *intervention* and *non-intervention* groups. The results showed that the *intervention* group demonstrated improved scores in both tasks compared to the *non-intervention* group. The effect of the model is particularly evident in the *concept-level* term familiarity task. This finding suggests that the informal medical entity linking model significantly aids in enhancing laypeople's medical terminology knowledge.

Qualitative feedback from participants in both groups was positive, highlighting the model's potential to support laypeople in enhancing their medical terminology knowledge. Nevertheless, concerns were raised regarding the reliability of Wikipedia as a source and the potential for misinformation.

However, we acknowledge that our evaluation tasks and metrics may not effectively capture the medical terminology knowledge of the participants in both groups, especially in *intervention* group. We could improved the evaluation tasks by assessing the participants' ability to comprehend health-related information in connection with the medical concepts they learned during the experiment, and incorporate established evaluation instruments, such as ComprehENotes used to evaluate the effectiveness of our proposed informal medical entity linking model.

To summarized our research contribution for this thesis, Functional health literacy (FHL) is one of the important basic skills for laypeople, either as patients or not to take an active role in their healthcare. The importance of medical terminology knowledge for better FHL cannot be overstated.

Previous research focused on addressing this problem by proposing systems to assist laypeople in understanding medical terminology found in medical documents by simplifying the medical terminology by substituting it with a synonym or translating it to lay definition.

Our research takes a novel approach by doing the reverse technique of the traditional method of simplifying formal medical concepts into easy-to-understand explanations for laypeople. Instead, we aim to enhance health literacy by introducing laypeople to formal medical terms by linking these terms to the informal medical phrases they already use. The proposed informal medical entity linking model, offers a reverse directions to bridge the gap between the popularized and specialized medical terminology for laypeople.

The effectiveness of the our proposed model evidenced by the improved performance of the participants in *intervention* group in ability to recognize word-form (*surface-level* task) and identify meaning (*concept-level* task) of the specialized medical terminology. While acknowledging our limitations in proposed evaluation approach and instruments, this thesis opens a new research avenue in supporting laypeople in enhancing their medical terminology knowledge, specifically to contribution to better FHL.

7.2 Reflections on the Research Done and the Achieved Results

The goal of this Section is to reflect upon and share practical experiences concerning the entire research. This reflection mainly focuses on those intangibles that do not necessarily belong to

the scope of the scholarly conclusion. While Section 7.1 focused on the scientific reasoning and contributions, and Section 7.3 highlights the open issues and recommends directions for further studies, the experiences and opinions are formulated below in a subjective way.

7.2.1 Reflection 1: The Importance of Domain Experts

As a first reflection, we want to highlight the importance of the domain experts during conducting the research. As the main goal of the research is to build a model that can help laypeople enhance medical terminology knowledge, domain experts play an important role in each of the phases of the model development. Their specialized knowledge and insights were invaluable in guiding the development process, from the initial conceptualization of the model to the fine-tuning of algorithms.

In our research, we involved domain experts at multiple stages to evaluate the model's output as a means of understanding the quality of the proposed model. However, the experts who participated had diverse medical specialties, which introduced variability into our evaluation results. The differing viewpoints sometimes made it difficult to reach a consensus on certain aspects of the model, such as shown in Chapter 5, when evaluating the accuracy of the MCN model's output. This highlights the complexity of integrating cross-disciplinary insights.

The experience underscored the difficulty of finding domain experts who are willing and able to dedicate their time voluntarily. Engaging such experts is often not only challenging but also costly, both in terms of time and financial resources. Despite these challenges, the iterative involvement of domain experts provides a continuous validation mechanism, which is vital for refining and improving the model. Their feedback helps to identify gaps and areas for improvement, enabling a more robust and effective model.

7.2.2 Reflection 2: The Medical Entity Linking Task Remains Challenging

We aimed to explore the potential benefits of the medical entity linking task for laypeople, particularly in enhancing their knowledge of medical terminology within social media settings. Our approach involved leveraging popularized medical terms to introduce specialized medical terminology and its definitions. While we successfully addressed one of the research challenges — specifically, the data scarcity problem — through the incorporation of data augmentation and distant supervision approaches (as discussed in Section 7.1), several challenges persisted.

One of the challenges we encountered was the variability in how laypeople express medical concepts. Laypeople often describe medical conditions using colloquial language, such as "no longer able to enjoy the occasional glass of wine or champagne b/c it makes me too drowsy," to more dictionary-defined medical terms like "alcohol intolerant.".

Upon reflection, we recognize that our current data augmentation and distant supervision techniques, while promising in addressing data scarcity, sometimes exacerbated the challenges in accurately identifying and normalizing these popularized medical phrases. This experience prompted a critical reflection on our methodologies, highlighting the need for more nuanced approaches that can better accommodate the intricacies of layperson language in medical contexts.

The task of medical concept normalization presented another set of challenges, primarily centered around ambiguity and granularity. The hierarchical structure of SNOMED-CT, which ranges from broad, general concepts to highly specific ones, contributed to this complexity. Our approach

of treating medical concept normalization as a multi-class classification task proved inadequate in capturing these nuances. We observed that the popularized medical phrases we identified extracted from the distant supervision approach often corresponded to multiple SNOMED-CT concepts, further complicating the normalization process. This experience has led us to consider alternative approaches, such as hierarchical classification, which might better address the granularity issues inherent in medical terminology.

In addition, our reliance on an existing entity disambiguation model posed challenges. While the model effectively disambiguates appropriate Wikipedia articles based on the input, we found instances where the top-ranked article was not necessarily the most accurate definition source for specialized medical terms. To mitigate this, we integrated a traditional machine learning model to re-rank the outputs. However, as we reflected on this process, we recognized the limitations of the data available to train the traditional machine learning model, which continues to pose challenges in the entity disambiguation phase.

7.2.3 Reflection 3: Challenge in Expert Evaluation

In reflecting on our approach to evaluating the medical entity linking model, we decided to use expert evaluation before delivering the model to the end-user of this research. This decision was influenced by the limitations of the benchmark data available for conducting automatic evaluations. We believed that expert evaluation would provide a more thorough assessment, particularly when testing with different users' posts found in the available data.

Our evaluation process was structured around two key tasks: (1) assessing the correctness of the specialized medical terms predicted by the MCN model; and (2) evaluating the appropriateness of the Wikipedia articles retrieved by the ED module, which were associated with both the popularized medical phrases and specialized terms.

While the results of this evaluation were promising, we now recognize several limitations in our approach. The scope of evaluation tasks was narrow, potentially missing nuances in the model's performance. Our binary (yes/no) approach to assessing term correctness may have oversimplified a complex normalization process. Additionally, the method of selecting appropriate Wikipedia articles based solely on rank position might have been too restrictive.

We can enhance our evaluation methodology in several ways. For specialized term prediction, we could explore the incorporation of a finer granularity of answers in the annotation process. This would allow for a more nuanced assessment of the specialized predicted terms, even if they are not the most precise but still relate to the popularized ones. For Wikipedia article retrieval, we could develop methods for annotators to indicate relevant articles not retrieved or ranked by our model. These improvements would allow for a more comprehensive assessment of our model's performance and better generalization of results.

7.2.4 Reflection 4: Challenge in User Experiment

As mentioned in Chapter 2, previous research has focused on bridging the language gap between lay medical expressions and specialized medical terminology by proposing systems that simplify medical terms found in Electronic Health Record (EHR) documents for laypeople. In contrast, our method reverses this process by leveraging social media to introduce specialized medical terms using popularized phrases commonly used by laypeople.

To assess the effectiveness of our model, we conducted user experiments designed to evaluate how well the model helps laypeople learn specialized medical terminology. The results of these

experiments indicated a positive effect for participants who received support from our proposed model.

We approached the user experiment with a micro-learning and spaced repetition strategy to simulate real-world scenarios. Participants were exposed multiple times to specific specialized medical terminology through popularized terms found on social media. This approach was adapted from spaced repetition techniques to ensure repeated exposure. However, due to time and cost constraints, we were unable to evaluate the memory retention of participants over an extended period, which ideally would take days or weeks to fully assess the effectiveness of our model. Instead, we opted for a controlled evaluation setting to provide a feasible yet informative assessment of our model's effectiveness.

We acknowledge that this decision may limit the generalizability of our evaluation results. Additionally, the selection of evaluation tasks was restricted to recognizing word forms and identifying the meanings of medical terms. This focus does not capture the full depth of our model's effectiveness in enhancing laypeople's medical terminology knowledge. Ideally, participants should not only recognize and understand medical terms but also be able to produce and use them correctly in context [Nat01]. Assessing such depth of knowledge presents significant challenges, and our study was constrained by these practical limitations.

Moreover, the difference in mastery levels (i.e., number of attempts) in one of our evaluation tasks (*surface-level* term familiarity) caused reliability issues when comparing the *surface-level score* ($familarityScore$), where we incorporated the mastery level as the source of the score calculation (Section 6.1.5). To address this challenge, we introduced a new method of score calculation. This approach allowed us to derive more reliable results and better measure improvements in the *surface-level score*. Although we were able to mitigate this challenge, we acknowledge that our evaluation could be further improved. Incorporating established health literacy assessment tools, such as ComprehENotes or the Test of Functional Health Literacy (TOFHLA), would likely provide a more accurate measure of our informal medical entity linking model's effectiveness.

We are also aware that the selection of user posts and medical terminology used to evaluate the model performance may introduce subjectivity into the process of the user experiments. Additionally, evaluating the effectiveness of the informal medical EL model with a larger population, such as participants with low English proficiency, would be necessary to assess if the informal medical EL model could be equally effective in enhancing their medical terminology knowledge.

Despite these limitations and challenges, we argue that our proposed idea of introducing laypeople to specialized medical terminology in social media settings can serve as an alternative or complementary source to empower individuals and enhance their functional health literacy. As laypeople become more familiar with medical terms, they can engage more effectively with healthcare providers and navigate health information online with greater confidence.

However, we acknowledge that this process is a long-term process, and comprehensive analysis in the future needs to be conducted to prove whether the proposed approach can be beneficial in real-world scenarios. Our thesis scratches the surface of this potential, opening avenues for further research and refinement of the model and its applications.

7.3 Recommendation for Future Directions

Our work has opened up the following possibilities for further development. In this section, we present the potential future research based on our findings in the context of the thesis.

7.3.1 Expert Participation and Human-in-the-Loop Systems

We recognize the necessity of involving experts in the evaluation process to refine the model where experts provide feedback to the EL model, guiding its learning process and ensuring the accuracy of its outputs. This collaboration between the model and human experts is crucial in the public health domain to ensure the correct information. Moreover, to improve the quality of the proposed entity-linking model, we emphasize the need for the expert's participation to evaluate the model prediction. In this case, the informal medical entity linking model generates predictions, and experts review these predictions, correcting any inaccuracies and identifying elements that may have been overlooked by the model. The feedback provided by experts is then utilized to improve the model.

7.3.2 A More Robust Empirical Analysis

In Chapter 6, the user experiment was conducted with participants with specific criteria, such as being fluent in English and located in an English-speaking country. Moreover, the selection evaluation tasks were limited to the memorizing and retention the medical terminology, which might be too simplistic to evaluate the effectiveness of the model in enhancing medical terminology knowledge among laypeople. We also employed A/B testing, comparing the usability of the model among participants with and without access to it, which might be able to evaluate the learning progress of the laypeople who received support from our proposed model. Future research will address these limitations.

Evaluating the Model's Effectiveness Among Different Demographic Groups: Our research findings show that the proposed informal medical entity linking model can increase their familiarity and understanding of medical terminology. However, the demographic characteristics of the participants in our user experiment may not reflect the characteristics of laypeople with limited health literacy. People with limited health literacy are often older, have less education, have lower income, have chronic health conditions, and may not be native English speakers [HMCRT+18]. A possible future research direction involves evaluating the system with a more diverse population that reflects different levels of health literacy. This will help us to better understand the effectiveness of the informal medical entity linking model in different demographic groups and ensure that it can be a valuable tool for a wider audience.

Expanding Evaluation to Include Reading Literacy and Real-world Impact: Our current evaluation primarily focuses on assessing the comprehension of medical terminology, specifically regarding familiarity and understanding the meanings of these terms. However, in order to gain a more comprehensive understanding of the influence of our proposed informal medical entity linking model, it may be beneficial to broaden our evaluation criteria to include reading proficiency tasks. For instance, one potential approach could involve integrating evaluation tasks with Electronic Health Records (EHR) notes. This approach would help us determine whether the informal medical entity linking model not only assists laypeople in comprehending medical terminology but also enhances their ability to navigate healthcare documents more effectively.

Longitudinal Study to Evaluate the Effectiveness of the Model: In our evaluation design, we attempted to apply the spaced repetition technique to assess the memory retention of specialized medical terminology in social media settings. However, due to time constraints, we could not effectively implement this technique. Spaced repetition requires a longer study duration to measure the retention of terminology over time. Conducting a longitudinal study to evaluate our proposed approach would help determine the model's benefits in supporting laypeople in learning medical terminology. Additionally, a cohort study could complement this longitudinal

study by measuring the gradual improvement of participants who received support from our model over time.

7.3.3 Incorporating Large Language Models (LLMs) for Informal Medical Entity Linking

In Chapter 3 and Chapter 4, the proposed data augmentation and distant supervision techniques are able to increase the number of variations of the popularized term given a specific specialized term and the broaden the concept coverage of the medical concepts. However, the proposed data augmentation approach introduced a simple substitution approach which may increase the noise from the original datasets. Furthermore, the distant supervision technique introduced granularity-level issues. Building on this limitation, there is a potential improvement of our method as a future direction.

In the deluge of exploration on LLMs, the exploration of utilizing LLM, such as for paraphrasing popularized medical phrases can be beneficial to complementing and enhancing traditional methods. By leveraging their generative capabilities, LLMs can produce diverse, context-rich examples that closely mimic real-world medical discussions by laypeople. This approach can significantly expand the variety and volume of training data.

Another promising research direction is the integration of LLMs, which has seen growing prominence during the last period of our research. For example, we can investigate the use of LLMs, such as the Medical Pre-trained Attention-based Language Model (Med-Palm), an advanced deep learning model specifically designed for processing medical data [SAT+23]. The Med-Palm model is capable of recognizing and extracting specific medical entities, such as diseases, symptoms, drugs, and procedures, from unstructured text. This offers an opportunity to enhance our proposed model, particularly the Named Entity Recognition (NER), by integrating it with Med-Palm.

Additionally, the potential benefits of using LLMs as concept mapping agents, which normalize popular medical phrases into specialized medical terms, can be explored. This line of research may address the ambiguity issues present in the current model. Furthermore, the capability of LLMs to resolve granularity issues within medical data can be investigated.

Furthermore, Med-Palm can be employed to simplify the explanation of complex medical terminology, making it more accessible to laypeople. To summarize, the investigation of LLMs for addressing the medical entity linking task, with a focus on empowering laypeople to increase functional health literacy, represents a promising avenue for future research.

List of Figures

List of Tables

List of Algorithms

Bibliography

[AAAA15] Samy A Azer, Nourah M AlSwaidan, Lama A Alshwairikh, and Jumana M AlShammari. Accuracy and readability of cardiovascular entries on wikipedia: are they reliable learning resources for medical students? *BMJ Open*, 5(10), 2015.

[ABB+19] Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59, 2019.

[ABV18] Alan Akbik, Duncan Blythe, and Roland Vollgraf. Contextual string embeddings for sequence labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, pages 1638–1649, 2018.

[ALL+20] M. Alfano, B. Lenzitti, G. Lo Bosco, C. Muriana, T. Piazza, and G. Vizzini. Design, development and validation of a system for automatic help to medical text understanding. *International Journal of Medical Informatics*, 138:104109, 2020.

[AMU17] Muhammad Abdul-Mageed and Lyle Ungar. EmoNet: Fine-grained emotion detection with gated recurrent neural networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 718–728, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[Ang71] William H Angoff. *Educational measurement*. Washington: American Council on Education, 1971.

[AO20] Noor Aljassim and Remo Ostini. Health literacy in rural and urban populations: A systematic review. *Patient Education and Counseling*, 103(10):2142–2154, 2020.

[APW18] Jacopo Amidei, Paul Piwek, and Alistair Willis. Rethinking the agreement in human evaluation tasks. In Emily M. Bender, Leon Derczynski, and Pierre Isabelle, editors, *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3318–3329, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.

[Aro01] Alan R. Aronson. Effective mapping of biomedical text to the umls metathesaurus: The metamap program. 2001.

[Bak06] David W. Baker. The meaning and the measure of health literacy. *Journal of General Internal Medicine*, 21(8):878–883, 2006.

[Bar08] Arunodaya Barman. Standard setting in student assessment: is a defensible method yet to come? *Annals of the Academy of Medicine, Singapore*, 37 11:957–63, 2008.

[BGdLG18] Mats Byrkjeland, Frederik Gørvell de Lichtenberg, and Björn Gambäck. Ternary Twitter sentiment classification with distant supervision and sentiment-specific word embeddings. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 97–106, Brussels, Belgium, October 2018. Association for Computational Linguistics.

[Blo68] Benjamin S Bloom. Learning for mastery. instruction and curriculum. regional education laboratory for the carolinas and virginia, topical papers and reprints, number 1. *Evaluation comment*, 1(2):n2, 1968.

[BLSC20] M. Basaldella, F. Liu, E. Shareghi, and N. Collier. COMETA: A corpus for medical entity linking in the social media. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3122–3137, Online, November 2020. ACL.

[BMH17] Adrian Benton, Margaret Mitchell, and Dirk Hovy. Multi-task learning for mental health using social media text. *CoRR*, abs/1712.03538, 2017.

[BRB+07] Steven H. Brown, S. Trent Rosenbloom, Brent A. Bauer, Dietlind Wahner-Roedler, David A. Froehling, Kent R. Bailey, Michael J. Lincoln, Diane Montella, Elliot M. Fielstein, and Peter L. Elkin. Direct comparison of medcin and snomed ct for representation of a general medical evaluation template. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, pages 75–79, 2007. Copyright: MEDLINE® is the source for the citation and abstract of this record.

[BWP+99] David W Baker, Mark V Williams, Ruth M Parker, Julie A Gazmararian, and Joanne Nurss. Development of a brief test to measure functional health literacy. *Patient Education and Counseling*, 38(1):33–42, 1999.

[CDPR+18] Jinying Chen, Emily Druhl, Balaji Polepalli Ramesh, Thomas K. Houston, Cynthia A. Brandt, Donna M. Zulman, Varsha G. Vimalananda, Samir Malkani, and Hong Yu. A natural language processing system that links medical terms in electronic health record notes to lay definitions: System development using physician reviews, 2018.

[CFZ22] Yiling Cao, Lu Fang, and Zhongguang Zheng. Enriching pre-trained language model with multi-task learning and context for medical concept normalization. *Proceedings of the 2022 International Conference on Intelligent Medicine and Health*, 2022.

[CG20] Jean-Philippe Corbeil and Hadi Abdi Ghadivel. BET: A Backtranslation Approach for Easy Data Augmentation in Transformer-based Paraphrase Identification Context. pages 1–12, 2020.

[CHC+19] Yixin Cao, Zikun Hu, Tat-seng Chua, Zhiyuan Liu, and Heng Ji. Low-resource name tagging learned with weakly labeled data. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 261–270, Hong Kong, China, November 2019. Association for Computational Linguistics.

[Coh60] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37 – 46, 1960.

[CS93] Nancy Chinchor and Beth Sundheim. Muc-5 evaluation metrics. In *Proceedings of the 5th Conference on Message Understanding*, MUC5 '93, page 69–78, USA, 1993. Association for Computational Linguistics.

[CvMG+14] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, October 2014. Association for Computational Linguistics.

[DA20] Xiang Dai and Heike Adel. An analysis of simple data augmentation for named entity recognition. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3861–3867, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[DA22] Pradip Dhal and Chandrashekhar Azad. A comprehensive survey on feature selection in the various fields of machine learning. *Applied Intelligence*, pages 1–39, 2022.

[Dan13] V Dang. The lemur project-wiki-ranklib. *Lemur Project*, 2013.

[DIRP21] Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. Autoregressive entity retrieval. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[DLB+20] Bosheng Ding, Linlin Liu, Lidong Bing, Canasai Kruengkrai, Thien Hai Nguyen, Shafiq Joty, Luo Si, and Chunyan Miao. DAGA: Data augmentation with a generation approach for low-resource tagging tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6045–6057, Online, November 2020. Association for Computational Linguistics.

[DLJ+93] TC Davis, SW Long, RH Jackson, EJ Mayeaux, RB George, PW Murphy, and MA Crouch. Rapid estimate of adult literacy in medicine: a shortened screening instrument. *Family medicine*, 25(6):391—395, June 1993.

[Don06] Kevin Donnelly. Snomed-ct: The advanced terminology and coding system for ehealth. *Studies in health technology and informatics*, 121:279—290, 2006.

[DWK19] Julia Dembowski, Michael Wiegand, and Dietrich Klakow. Language independent named entity recognition using distant supervision. Human Language Technologies as a Challenge for Computer Science and Linguistics. Proceedings of the 8th Language & Technology Conference, November 17-19, 2017, Poznań, Poland, pages 68 – 72, Poznań, 2019. Fundacja Uniwersytetu im. Adama Mickiewicza.

[EAZG18] Mostafa Janebi Enayat, Seyed Mohammad Reza Amirian, Gholamreza Zareian, and Saeed Ghaniabadi. Reliable measure of written receptive vocabulary size: Using the l2 depth of vocabulary knowledge as a yardstick. *Sage Open*, 8(1):2158244017752221, 2018.

[EENA+20] Alaa El-Ebshihy, Annisa Maulida Ningtyas, Linda Andersson, Florina Piroi, and Andreas Rauber. Artu/tu wien and artificial researcher@ longsumm 20. In *Proceedings of the First Workshop on Scholarly Document Processing*, pages 310–317, 2020.

[EENA+22] Alaa El-Ebshihy, Annisa Maulida Ningtyas, Linda Andersson, Florina Piroi, and Andreas Rauber. A platform for argumentative zoning annotation and scientific summarization. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 4843–4847, 2022.

[EENP+23] Alaa El-Ebshihy, Annisa Maulida Ningtyas, Florina Piroi, Andreas Rauber, Ade Romadhony, Said Al Faraby, and Mira Kania Sabariah. Using semi-automatic annotation platform to create corpus for argumentative zoning. In *International Conference on Theory and Practice of Digital Libraries*, pages 132–145. Springer, 2023.

[ESNG18] Ehsan Emadzadeh, Abeed Sarker, Azadeh Nikfarjam, and Graciela Gonzalez. Hybrid semantic analysis for mapping adverse drug reaction mentions in tweets to medical terminology. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2017:679–688, 04 2018.

[FBJ13] A. M. Fage-Butler and M. N. Jensen. The interpersonal dimension of online patient forums: How patients manage informational and relational aspects in response to posted questions. *HERMES-Journal of Language and Communication in Business*, (51):21–38, 2013.

[FBM17] Marzieh Fadaee, Arianna Bisazza, and Christof Monz. Data augmentation for low-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 567–573, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[FBNJ16a] A. M. Fage-Butler and M. Nisbeth Jensen. Medical terminology in online patient–patient communication: evidence of high health literacy? *Health Expectations*, 19(3):643–653, 2016.

[FBNJ16b] Antoinette M Fage-Butler and Matilde Nisbeth Jensen. Medical terminology in online patient–patient communication: evidence of high health literacy? *Health Expectations*, 19(3):643–653, 2016.

[FC14] Joseph D Feuerstein and Adam S Cheifetz. Ulcerative colitis: epidemiology, diagnosis, and management. In *Mayo Clinic Proceedings*, volume 89, pages 1553–1563. Elsevier, 2014.

[FDBA22] Matús Falis, Hang Dong, Alexandra Birch, and Beatrice Alex. Horses to zebras: Ontology-guided data augmentation and synthesis for ICD-9 coding. In *Workshop on Biomedical Natural Language Processing*, 2022.

[FLN20] Jinlan Fu, Pengfei Liu, and Graham Neubig. Interpretable multi-dataset evaluation for named entity recognition. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6058–6069, Online, November 2020. Association for Computational Linguistics.

[FM23] Evan French and Bridget T. McInnes. An overview of biomedical entity linking throughout the years. *Journal of Biomedical Informatics*, 137:104252, 2023.

[GB08] Suzanne Graham and John Brookey. Do patients understand? *The Permanente Journal*, 12(3):67–69, 2008.

[Gil06] Jim Giles. Internet encyclopaedias go head to head. *Nature*, 438:900–1, 01 2006.

[GK14] Omid Ghiasvand and Rohit Kate. UWM: Disorder mention extraction from clinical text using CRFs and normalization using learned edit distance patterns. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 828–832, Dublin, Ireland, August 2014. Association for Computational Linguistics.

[GPH+22] Rachael Gotlieb, Corinne Praska, Marissa A. Hendrickson, Jordan Marmet, Victoria Charpentier, Emily Hause, Katherine A. Allen, Scott Lunos, and Michael B. Pitt. Accuracy in Patient Understanding of Common Medical Phrases. *JAMA Network Open*, 5(11):e2242972–e2242972, 11 2022.

[GR20] Siddhant Garg and Goutham Ramakrishnan. Bae: Bert-based adversarial examples for text classification. *arXiv preprint arXiv:2004.01970*, 2020.

[GTC+20] Yu Gu, Robert Tinn, Hao Cheng, Michael R. Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3:1 – 23, 2020.

[HK18] Michael A. Hedderich and Dietrich Klakow. Training a neural network in a low-resource setting on automatically annotated noisy data. In *Proceedings of the Workshop on Deep Learning Approaches for Low-Resource NLP*, pages 12–18, Melbourne, July 2018. Association for Computational Linguistics.

[HL18] Xinzhi Han and Sen Lei. Feature selection and model comparison on microsoft learning-to-rank data sets. *CoRR*, abs/1803.05127, 2018.

[HLA+20] Michael A Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. A survey on recent approaches for natural language processing in low-resource scenarios. *arXiv preprint arXiv:2010.12309*, 2020.

[HMCRT+18] Kathleen Hickey, Ruth Masterson Creber, Meghan Reading Turchioe, Robert Sciacca, Teresa Riga, Ashton Frulla, and Jesus (Jessie) Casida. Low health literacy: Implications for managing cardiac patients in practice. *The Nurse practitioner*, 43:49–55, 08 2018.

[HMMW06] R. Harrison, A. MacFarlane, E. Murray, and P. Wallace. Patients' perceptions of joint teleconsultations: a qualitative evaluation. *Health Expectations*, 9(1), 2006.

[Hug12] Theo Hug. *Encyclopedia of the Sciences of Learning: Microlearning*, pages 2268–2271. Springer US, Boston, MA, 2012.

[HW15] James M Heilman and Andrew G West. Wikipedia and medicine: Quantifying readership, editors, and the significance of natural language. *J Med Internet Res*, 17(3):e62, Mar 2015.

[HXY15] Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional LSTM-CRF models for sequence tagging. *ArXiv*, abs/1508.01991, 2015.

[Kan94] Michael Kane. Validating the performance standards associated with passing scores. *Review of Educational Research*, 64(3):425–461, 1994.

[Kat15] Rohit Kate. Normalizing clinical terms using learned edit distance patterns. *Journal of the American Medical Informatics Association : JAMIA*, 23, 07 2015.

[KCZT10] Sasikiran Kandula, Dorothy Curtis, and Qing Zeng-Treitler. A semantic and syntactic text simplification tool for health content. In *AMIA annual symposium proceedings*, volume 2010, page 366. American Medical Informatics Association, 2010.

[KGH18] Nikolaos Kolitsas, Octavian-Eugen Ganea, and Thomas Hofmann. End-to-end neural entity linking. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 519–529, Brussels, Belgium, October 2018. Association for Computational Linguistics.

[KMJKW15] Sarvnaz Karimi, Alejandro Metke-Jimenez, Madonna Kemp, and Chen Wang. Cadec: A corpus of adverse drug event annotations. *Journal of Biomedical Informatics*, 55:73 – 81, 2015.

[Kob18] Sosuke Kobayashi. Contextual augmentation: Data augmentation by words with paradigmatic relations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 452–457, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[KPT+20] Tian Kang, Adler Perotte, Youlan Tang, Casey Ta, and Chunhua Weng. UMLS-based data augmentation for natural language processing of clinical research literature. *Journal of the American Medical Informatics Association*, 28(4):812–823, 12 2020.

[KRS16] Daniel Kershaw, Matthew Rowe, and Patrick Stacey. Towards modelling language innovation acceptance in online social networks. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, WSDM '16, page 553–562, New York, NY, USA, 2016. Association for Computing Machinery.

[KYJ+22] Sunjae Kwon, Zonghai Yao, Harmon Jordan, David Levy, Brian Corner, and Hong Yu. MedJEx: A medical jargon extraction model with Wiki's hyperlink span and contextualized masked language model score. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11733–11751, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[LAS19] Lukas Lange, Heike Adel, and Jannik Strötgen. NLNDE: Enhancing neural sequence taggers with attention and noisy channel for robust pharmacological entity detection. In *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, pages 26–32, Hong Kong, China, November 2019. Association for Computational Linguistics.

[LBHT20] Pierre Lison, Jeremy Barnes, Aliaksandr Hubin, and Samia Touileb. Named entity recognition without labelled data: A weak supervision approach. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1518–1533, Online, July 2020. Association for Computational Linguistics.

[LC15] Nut Limsopatham and Nigel Collier. Adapting phrase-based machine translation to normalise medical terms in social media messages. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1675–1680, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[LC16] Nut Limsopatham and Nigel Collier. Normalising medical concepts in social media texts by learning semantic representation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1014–1023, Berlin, Germany, August 2016. Association for Computational Linguistics.

[LCW+19] A.R. Links, W. Callon, C. Wasserman, J. Walsh, M.C. Beach, and E.F. Boss. Surgeon use of medical jargon with parents in the outpatient setting. *Patient Education and Counseling*, 102(6):1111–1118, 2019.

[LDl13] Robert Leaman, Rezarta Dogan, and Zhiyong lu. Dnorm: Disease name normalization with pairwise learning to rank. *Bioinformatics (Oxford, England)*, 29, 08 2013.

[LHF+17] K. Lee, S. A. Hasan, O. Farri, A. Choudhary, and A. Agrawal. Medical concept normalization for online user-generated texts. In *2017 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 462–469, 2017.

[Lip00] Carolyn E Lipscomb. Medical subject headings (mesh). *Bulletin of the Medical Library Association*, 88(3):265, 2000.

[LOG+19] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.

[LTMB17] Chris Lu, Destinee Tormey, Lynn McCreedy, and Allen Browne. Enhanced lexsynonym acquisition for effective umls concept mapping. *Studies in health technology and informatics*, 245:501–505, 01 2017.

[LWC+18] John P Lalor, Hao Wu, Li Chen, Kathleen M Mazor, and Hong Yu. Comprehenotes, an instrument to assess patient reading comprehension of electronic health record notes: Development and validation. *J Med Internet Res*, 20(4):e139, Apr 2018.

[LWJ+23] Xu Li, Chengkun Wei, Zhuoren Jiang, Wenlong Meng, Fan Ouyang, Zihui Zhang, and Wenzhi Chen. Eduner: a chinese named entity recognition dataset for education research. *Neural Computing and Applications*, 35(24):17717–17731, 2023.

[LWY19] John P Lalor, Beverly Woolf, and Hong Yu. Improving electronic health record note comprehension with noteaid: Randomized trial of electronic health record note comprehension interventions with crowdsourced workers. *J Med Internet Res*, 21(1):e10793, Jan 2019.

[Ma19] Edward Ma. [nlp] augmentation. https://github.com/makcedward/nlpaug, 2019.

[MBS24] Frank Mangold Marko Bachl, Elena Link and Sebastian Stier. Search engine use for health-related purposes: Behavioral data on online health information-seeking in germany. *Health Communication*, 0(0):1–14, 2024. PMID: 38326714.

[MBSJ09] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011, 2009.

[MC07] Rada Mihalcea and Andras Csomai. Wikify! linking documents to encyclopedic knowledge. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, CIKM '07, page 233–242, New York, NY, USA, 2007. Association for Computing Machinery.

[mes] Medical subject headings. `https://www.nlm.nih.gov/mesh/introduction.html`. Accessed: 23 April 2024.

[MFdlC+20] Louis Martin, Angela Fan, Éric de la Clergerie, Antoine Bordes, and Benoît Sagot. Multilingual unsupervised sentence simplification. *CoRR*, abs/2005.00352, 2020.

[MGL+21] Michal Monselise, Jane Greenberg, Ou Stella Liang, Sonia Pascua, Heejun Kim, Mat Kelly, Joan P Boone, and Christopher C Yang. An automatic approach to extending the consumer health vocabulary. *Journal of Data and Information Science*, 6(1):35–49, 2021.

[MKA13] Maznah Mat Kasim and Siti Abdullah. Simple weighted average as an alternative method in aggregating students' academic achievements. *Malaysian Journal of Learning and Instruction*, 10:119–132, 12 2013.

[MSS+23] Tomasz Miksa, Marek Suchánek, Jan Slifka, Vojtech Knaisl, Fajar J Ekaputra, Filip Kovacevic, Annisa Maulida Ningtyas, Alaa El-Ebshihy, and Robert Pergl. Towards a toolbox for automated assessment of machine-actionable data management plans. *Data Science Journal*, 22:28, 2023.

[MT19] Z. Miftahutdinov and E. Tutubalina. Deep neural models for medical concept normalization in user-generated texts. In *Proc. of the 57th Annual Meeting of the ACL: Student Research Workshop*, pages 393–399, Florence, Italy, July 2019. ACL.

[MTT17] Zulfat Miftahutdinov, Elena Tutubalina, and Alexander Tropsha. Identifying disease-related expressions in reviews using conditional random fields. In *Proceedings of International Conference Dialog*, volume 1, pages 155–167, 2017.

[MWM00] Kyongho Min, William H. Wilson, and Yoo-Jin Moon. Typographical and orthographical spelling error correction. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, Athens, Greece, May 2000. European Language Resources Association (ELRA).

[MYS+19] Sepideh Mesbah, Jie Yang, Robert-Jan Sips, Manuel Valle Torre, Christoph Lofi, Alessandro Bozzon, and Geert-Jan Houben. Training data augmentation for detecting adverse drug reactions in user-generated content. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2349–2359, Hong Kong, China, November 2019. Association for Computational Linguistics.

[Nat01] I. S. P. Nation. *Learning Vocabulary in Another Language*. Cambridge Applied Linguistics. Cambridge University Press, 2001.

[NEEH+22] A. M. Ningtyas, A. El-Ebshihy, G. B. Herwanto, F. Piroi, and A. Hanbury. Leveraging wikipedia knowledge for distant supervision in medical concept normalization. In Alberto Barrón-Cedeño, G. Da San Martino, M. Degli Esposti, F. Sebastiani, C. Macdonald, G. Pasi, A. Hanbury, M. Potthast, G. Faggioli, and N. Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 33–47, Cham, 2022. Springer International Publishing.

[NEEP+20] Annisa Maulida Ningtyas, Alaa El-Ebshihy, Florina Piroi, Allan Hanbury, and Linda Andersson. Tuw-ifs at trec news 2020: Wikification task. In *TREC*, 2020.

[NFL+20] Ana Luisa Neves, Lisa Freise, Liliana Laranjo, Alexander W Carter, Ara Darzi, and Erik Mayer. Impact of providing patients access to electronic health records on quality and safety of care: a systematic review and meta-analysis. *BMJ Quality & Safety*, 29(12):1019–1032, 2020.

[NHPA21] Annisa Maulida Ningtyas, Allan Hanbury, Florina Piroi, and Linda Andersson. Data augmentation for layperson's medical entity linking task. In *Proceedings of the 13th Annual Meeting of the Forum for Information Retrieval Evaluation*, pages 99–106, 2021.

[Nie20] Dillon Niederhut. niacin: A Python package for text data enrichment. *The Journal of Open Source Software*, 5(50):2136, June 2020.

[Nin22] Annisa Maulida Ningtyas. Medical entity linking in laypersons' language. In *European Conference on Information Retrieval*, pages 513–519. Springer, 2022.

[NTK+19] Duy Hoa Ngo, Donna Truran, Madonna Kemp, Michael Lawley, and Alejandro Metke-Jimenez. Can wikipedia be used to derive an open clinical terminology? In *Digital Health: Changing the Way Healthcare is Conceptualised and Delivered: Selected Papers from the 27th Australian National Health Informatics Conference (HIC 2019)*, volume 266, page 136. IOS Press, 2019.

[Nut00] Don Nutbeam. Health literacy as a public health goal: a challenge for contemporary health education and communication strategies into the 21st century. *Health Promotion International*, 15(3):259–267, 09 2000.

[NYZ+19] Jinghao Niu, Yehui Yang, Siheng Zhang, Zhengya Sun, and Wensheng Zhang. Multi-task character-level attentional networks for medical concept normalization. *Neural Process. Lett.*, 49(3):1239–1256, June 2019.

[OHLD+13] R.L. O'Connell, S.K. Hartridge-Lambert, N. Din, E.R. St John, C. Hitchins, and T. Johnson. Patients' understanding of medical terminology used in the breast clinic. *The Breast*, 22(5):836–838, 2013.

[oM04] Institute of Medicine. *Health Literacy: A Prescription to End Confusion*. The National Academies Press, Washington, DC, 2004.

[PAP+20] N. Pattisapu, V. Anand, S. Patil, G. Palshikar, and V. Varma. Distant supervision for medical concept normalization. *Journal of Biomedical Informatics*, 109:103522, 2020.

[PBWN95] Ruth M Parker, David W Baker, Mark V Williams, and Joanne R Nurss. The test of functional health literacy in adults: a new instrument for measuring patients' literacy skills. *Journal of general internal medicine*, 10:537–541, 1995.

[Pew11] Pew Research Center. Pew Research Center: Health topics. Washington, D.C., 2011.

[PMFN15] Ira Puspitasari, Koichi Moriyama, Ken-ichi Fukui, and Masayuki Numao. Effects of individual health topic familiarity on activity patterns during health information searches. *JMIR Med Inform*, 3(1):e16, Mar 2015.

[PPPV20] Nikhil Pattisapu, Sangameshwar Patil, Girish Palshikar, and Vasudeva Varma. Medical Concept Normalization by Encoding Target Knowledge. volume 116 of *Proceedings of Machine Learning Research*, pages 246–259. PMLR, 13 Dec 2020.

[PRGD22] Eyal Peer, David Rothschild, Andrew Gordon, and Ekaterina Damer. Data quality of platforms and panels for online behavioral research. *Behavior research methods*, 54(5):2618—2620, October 2022.

[PRHB+13] B. Polepalli Ramesh, T. Houston, C. Brandt, H. Fang, and H. Yu. Improving patients' electronic health record comprehension with noteaid. In *MEDINFO 2013*, pages 714–718. IOS Press, 2013.

[PSG08] Martin Potthast, Benno Stein, and Robert Gerling. Automatic vandalism detection in wikipedia. In *European Conference on Information Retrieval*, 2008.

[PSM14] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[PYL19] Yifan Peng, Shankai Yan, and Zhiyong Lu. Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. In Dina Demner-Fushman, Kevin Bretonnel Cohen, Sophia Ananiadou, and Junichi Tsujii, editors, *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65, Florence, Italy, August 2019. Association for Computational Linguistics.

[RDD22] François Remy, Kris Demuynck, and Thomas Demeester. Biolord: Learning ontological representations from definitions (for biomedical concepts and their textual descriptions). *ArXiv*, abs/2210.11892, 2022.

[RG19] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019.

[RSP23] François Remy, Simone Scaboro, and Beatrice Portelli. Boosting adverse drug event normalization on social media: General-purpose model initialization and biomedical semantic text similarity benefit zero-shot linking in informal contexts. *ArXiv*, abs/2308.00157, 2023.

[RXA+22] Pâmela M Rezende, Joicymara S Xavier, David B Ascher, Gabriel R Fernandes, and Douglas E V Pires. Evaluating hierarchical machine learning approaches to classify biological databases. *Briefings in Bioinformatics*, 23(4):bbac216, 06 2022.

[RZNC20] Shruti Rijhwani, Shuyan Zhou, Graham Neubig, and Jaime Carbonell. Soft gazetteers for low-resource named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8118–8123, Online, July 2020. Association for Computational Linguistics.

[RZT04] Stephen Robertson, Hugo Zaragoza, and Michael Taylor. Simple bm25 extension to multiple weighted fields. In *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management*, CIKM '04, page 42–49, New York, NY, USA, 2004. Association for Computing Machinery.

[SAT+23] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

[SBMHZ13] Isabel Segura-Bedmar, Paloma Martínez, and María Herrero-Zazo. SemEval-2013 task 9 : Extraction of drug-drug interactions from biomedical texts (DDIExtraction 2013). In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 341–350, Atlanta, Georgia, USA, June 2013. Association for Computational Linguistics.

[Sch14] Norbert Schmitt. Size and depth of vocabulary knowledge: What the research shows. *Language Learning*, 64(4):913–951, 2014.

[SHB15] Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. *CoRR*, abs/1511.06709, 2015.

[SHE+17] Zoya Surani, Rahim Hirani, Anita Elias, Lauren Quisenberry, Joseph Varon, Sara Surani, and Salim Surani. Social media usage among health care providers. *BMC research notes*, 10(1):1–5, 2017.

[Sim] Simple English Wikipedia. Simple english wikipedia. `https://simple.wikipedia.org/wiki/Simple_English_Wikipedia`. Accessed: June 9, 2024.

[SKW+17] Michael A Scaffidi, Rishad Khan, Christopher Wang, Daniela Keren, Cindy Tsui, Ankit Garg, Simarjeet Brar, Kamesha Valoo, Michael Bonert, Jacob F de Wolff, James Heilman, and Samir C Grover. Comparison of the impact of wikipedia, uptodate, and a digital textbook on short-term knowledge acquisition among medical students: Randomized controlled trial of three web-based resources. *JMIR Med Educ*, 3(2):e20, Oct 2017.

[SM16] Burr Settles and Brendan Meeder. A trainable spaced repetition model for language learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1848–1858, Berlin, Germany, August 2016. Association for Computational Linguistics.

[Smi20] Denise Smith. Situating wikipedia as a health information resource in various contexts: A scoping review. *PLOS ONE*, 15:e0228786, 02 2020.

[SMK+17] Thomas Shafee, Gwinyai Masukume, Lisa Kipersztok, Diptanshu Das, Mikael Häggström, and James Heilman. Evolution of wikipedia's medical content: past, present and future. *J Epidemiol Community Health*, 71(11):1122–1129, 2017.

[SMLQB20] S. Scepanovic, E. Martin-Lopez, D. Quercia, and K. Baykaner. Extracting medical entities from social media. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, CHIL '20, page 170–181, New York, NY, USA, 2020. Association for Computing Machinery.

[SMM+20] Laura Seiffe, Oliver Marten, Michael Mikhailov, Sven Schmeier, Sebastian Möller, and Roland Roller. From witch's shot to music making bones - resources for medical laymen to technical language and vice versa. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6185–6192, Marseille, France, May 2020. European Language Resources Association.

[sno] Snomed ct starter guide. `https://confluence.ihtsdotools.org/display/DOCSTART/SNOMED+CT+Starter+Guide?preview=/28742871/47677485/doc_StarterGuide_Current-en-US_INT_20170728.pdf`. Accessed: 22 April 2024.

[SPC+22] Simone Scaboro, Beatrice Portelli, Emmanuele Chersoni, Enrico Santus, and Giuseppe Serra. Increasing adverse drug events extraction robustness on social media: Case study on negation and speculation. *Experimental Biology and Medicine*, 247:2003 – 2014, 2022.

[SPR+15] Kristine Sørensen, Jürgen M. Pelikan, Florian Röthlin, Kristin Ganahl, Zofia Slonska, Gerardine Doyle, James Fullam, Barbara Kondilis, Demosthenes Agrafiotis, Ellen Uiters, Maria Falcon, Monika Mensing, Kancho Tchamov, Stephan van den Broucke, and on behalf of the HLS-EU Consortium Brand, Helmut. Health literacy in Europe: comparative results of the European health literacy survey (HLS-EU). *European Journal of Public Health*, 25(6):1053–1058, 04 2015.

[SS20] Kalyan Katikapalli Subramanyam and Sangeetha. Deep contextualized medical concept normalization in social media text. *Procedia Computer Science*, 171:1353 – 1362, 2020. Third International Conference on Computing and Network Communications (CoCoNet'19).

[STT17] Miftahutdinov Z Sh, EV Tutubalina, and AE Tropsha. Identifying disease-related expressions in reviews using conditional random fields. *Computational Linguistics and Intellectual Technologies*, 1(16):155–166, 2017.

[SVF+12] Kristine Sørensen, Stephan Van den Broucke, James Fullam, Gerardine Doyle, Jürgen Pelikan, Zofia Slonska, Helmut Brand, (HLS-EU) Consortium Health Literacy Project European, and A.J. Schuit. Health literacy and public health: A systematic review and integration of definitions and models. *BMC Public Health*, 12, January 2012.

[TCL+16] Salvatore Trani, Diego Ceccarelli, Claudio Lucchese, Salvatore Orlando, and Raffaele Perego. Sel: A unified algorithm for entity linking and saliency detection. In *Proceedings of the 2016 ACM Symposium on Document Engineering*, DocEng '16, page 85–94, New York, NY, USA, 2016. Association for Computing Machinery.

[TKSDM03] Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147, 2003.

[TLSD20] Mikhail Tikhomirov, N. Loukachevitch, Anastasiia Sirotina, and Boris Dobrov. Using bert and augmentation in named entity recognition for cybersecurity domain. In Elisabeth Métais, Farid Meziane, Helmut Horacek, and Philipp Cimiano, editors, *Natural Language Processing and Information Systems*, pages 16–24, Cham, 2020. Springer International Publishing.

[TMNM18] E. Tutubalina, Z. Miftahutdinov, S. Nikolenko, and V. Malykh. Medical concept normalization in social media posts with recurrent neural networks. *Journal of Biomedical Informatics*, 84:93 – 102, 2018.

[TN17] Elena Tutubalina and Sergey Nikolenko. Combination of deep recurrent neural networks and conditional random fields for extracting adverse drug reactions from user reviews. *Journal of healthcare engineering*, 2017:9451342, 2017.

[TPK+21] Archana Tapuria, Talya Porat, Dipak Kalra, Glen Dsouza, Sun Xiaohui, and Vasa Curcin. Impact of patient access to their electronic health record: systematic review. *Informatics for Health and Social Care*, 46(2):194–206, 2021. PMID: 33840342.

[TSH+19] Houcemeddine Turki, Thomas Shafee, Mohamed Ali Hadj Taieb, Mohamed Ben Aouicha, Denny Vrandečić, Diptanshu Das, and Helmi Hamdi. Wikidata: A large-scale collaborative ontological medical database. *Journal of Biomedical Informatics*, 99:103292, 2019.

[VJN+20] Shikhar Vashishth, Rishabh Joshi, Denis Newman-Griffis, Ritam Dutt, and Carolyn Rose. MedType: Improving Medical Entity Linking with Semantic Type Prediction. *arXiv e-prints*, page arXiv:2005.00460, May 2020.

[VKSL19] Clara Vania, Yova Kementchedjhieva, Anders Søgaard, and Adam Lopez. A systematic comparison of methods for low-resource dependency parsing on genuinely low-resource languages. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1105–1116, Hong Kong, China, November 2019. Association for Computational Linguistics.

[vMdKNC18] Hugo van Mens, Nicolette de Keizer, Remko Nienhuis, and Ronald Cornet. Clarifying diagnoses to laymen by employing the snomed ct hierarchy. *Studies in health technology and informatics*, 247:900–904, 01 2018.

[vMDN+20] Hugo J. T. van Mens, Ruben D. Duijm, Remko Nienhuis, Nicolette F. de Keizer, and Ronald Cornet. Towards an adoption framework for patient access to electronic health records: Systematic literature mapping study. *JMIR Medical Informatics*, 8, 2020.

[VMHZ14] V.G.Vinod Vydiswaran, Qiaozhu Mei, D. Hanauer, and Kai Zheng. Mining consumer health vocabulary from community-generated text. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2014:1150–9, 2014.

[VVP23] Riikka Vuokko, Anne Vakkuri, and Sari Palojoki. Systematized nomenclature of medicine-clinical terminology (snomed ct) clinical use cases in the context of electronic health record systems: Systematic literature review. *JMIR medical informatics*, 11, 2023.

[WFG+18] D. Wishart, Y. D. Feunang, Anchi Guo, Elvis J. Lo, A. Marcu, J. Grant, Tanvir Sajed, D. Johnson, Carin Li, Zinat Sayeeda, Nazanin Assempour, Ithayavani Iynkkaran, Yifeng Liu, A. Maciejewski, Nicola Gale, Alex Wilson, Lucy Chin, Ryan Cummings, Diana Le, Allison Pon, Craig Knox, and M. Wilson. Drugbank 5.0: a major update to the drugbank database for 2018. *Nucleic Acids Research*, 46:D1074 – D1082, 2018.

[Wik] Wikidata. Wikidata:Introduction - Wikidata — wikidata.org. `https://www.wikidata.org/wiki/Wikidata:Introduction`. [Accessed 03-05-2024].

[WJW+20] Qiong Wang, Zongcheng Ji, Jingqi Wang, Stephen Wu, Weiyan Lin, Wenzhen Li, Li Ke, Guohong Xiao, Qing Jiang, Hua Xu, and Yi Zhou. A study of entity-linking methods for normalizing chinese diagnosis and procedure terms to icd codes. *Journal of Biomedical Informatics*, 105:103418, 2020.

[Wor17] World Health Organization (WHO). The mandate for health literacy. `https://www.who.int/teams/health-promotion/enhanced-wellbeing/ninth-global-conference/health-literacy`, 2017.

[WSM+20] Leon Weber, Mario Sänger, Jannes Münchmeyer, Maryam Habibi, Ulf Leser, and Alan Akbik. Hunflair: An easy-to-use tool for state-of-the-art biomedical named entity recognition. *arXiv preprint arXiv:2008.07347*, 2020.

[WZ19] Jason Wei and Kai Zou. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China, November 2019. Association for Computational Linguistics.

[YZTB19] Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. PAWS-X: A Cross-lingual Adversarial Dataset for Paraphrase Identification. In *Proc. of EMNLP*, 2019.

[ZBH19] Yuan Zhang, Jason Baldridge, and Luheng He. PAWS: Paraphrase Adversaries from Word Scrambling. In *Proc. of NAACL*, 2019.

[ZFP+19] M. Zolnoori, K. W. Fung, T. B. Patrick, P. Fontelo, H. Kharrazi, A. Faiola, N. D. Shah, Y. S. Shirley Wu, C. E. Eldredge, J. Luo, M. Conway, J. Zhu, Soo K. Park, K. Xu, and H. Moayyed. The psytar dataset: From patients generated narratives to a corpus of adverse drug events and effectiveness of psychiatric medications. *Data in Brief*, 24:103838, 2019.

[ZHZ+15] Jin G Zheng, Daniel Howsmon, Boliang Zhang, Juergen Hahn, Deborah Mcguinness, and James Hendler. Entity linking for biomedical literature. *BMC Medical Informatics and Decision Making*, 15(Suppl 1):1–9, 2015.

[ZLH+16] Bin Zhou, Yuan Lu, Kaveh Hajifathalian, James Bentham, Mariachiara Di Cesare, Goodarz Danaei, Honor Bixby, Melanie J Cowan, Mohammed K Ali, Cristina Taddei, et al. Worldwide trends in diabetes since 1980: a pooled analysis of 751 population-based studies with 4 · 4 million participants. *The lancet*, 387(10027):1513–1530, 2016.

[ZT06] Qing T. Zeng and Tony Tse. Exploring and developing consumer health vocabularies. *Journal of the American Medical Informatics Association*, 13(1):24–29, 2006.

[ZTGK+07] Qing Zeng-Treitler, Sergey Goryachev, Hyeoneui Kim, Alla Keselman, and Douglas Rosendale. Making texts in electronic health records comprehensible to consumers: a prototype translator. In *AMIA Annual Symposium Proceedings*, volume 2007, page 846. American Medical Informatics Association, 2007.

[ZTGW+06] Qing Zeng-Treitler, Sergey Goryachev, Scott Weiss, Margarita Sordo, Shawn Murphy, and Ross Lazarus. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: Evaluation of a natural language processing system. *BMC medical informatics and decision making*, 6:30, 02 2006.

Appendix A
Data Augmentation Examples

This appendix provides the various examples of the applied augmentation levels and technique conducted in the experiment in Chapter 3 Section 3.3. The example of the augmented texts using different augmentation levels and techniques presented in the table below.

Table 1: Original Text 1 and Its Augmentations

Original Text: Extremely severe pain in right shoulder as if from extreme workout or injury (none which apply). Stopped taking lipitor seven days ago and still experiencing pain in shoulder and tingling and numbness down right arm radiating into fingers. No more medications. will attempt holistic approach. vitamins C and/or niacin.		
Aug. Level	Tech.	Augmented Text
Context	Char	Extremely 8eveke pain in right shoulder as if fkum extkeme workout uk injury (nune which apply). Stopped taking lipitor 8even days ago and still experiencing pain in shoulder and tingling and numbness down right arm kadiatin9 into fingers. No more medications. wi11 attempt holistic approach. vitamins C and/or niacin. Extremely severe pain in right shoulder as if fomr extreme workout or injury (none which apply). Stopped taking lipitor seven days ago and still experiencing pain in shoulder and tingling and numbness down right arm radiating into fingers. No more medications. will attempt holistic approach. vitamins C and/or niacin. Extremely sfvere pain in right shoulder as if ff;m dxtre,e workout od injury (glne which apply). Stopped taking lipitor weven days ago ane still experiencing pain in shoulder and tingling and numbness down right arm rasiqting into fingers. No more medications. sill attwhoh holistic approach. vitamins C and/or niacin.
	Word	Extremely severe pain in right shoulder as if from extreme workout or injury (none which apply) . Stopped taking lipitor seven twenty-four hours ago and still experiencing pain in shoulder and tingling and numbness down right build up radiating into fingers . No more medications . will attempt holistic approach . vitamins C and/or niacin .

Table 1 (Continued)

Aug. Level	Tech.	Augmented Text
		Extremely severe renal colic in right shoulder as if from extreme workout or injury (none which apply) . Stopped taking lipitor seven days ago and abreact experiencing pain in shoulder and tingling and numbness down right rearm radiating into fingers . No more medications . will attempt holistic approach . vitamins C and/or niacin . Extremely severe symptom in right shoulder as if from extreme workout or injury (none which apply) . Stopped taking lipitor seven days ago and still experiencing pain in shoulder and tingling and numbness down right limb radiating into fingers . No more medications . will attempt holistic approach . vitamins C and/or niacin .
Entity	Char	Extremely severe pain in right shoulder as if from extreme workout or injury (none which apply). Stopped taking lipit0k seven days ago and still experiencing pain in shoulder and tin9lin9 and nombne8s down right arm radiating into fingers. No more medications. will attempt holistic approach. vitamins c and/or niacin. Extremely severe pain in right shoulder as if from extreme workout or injury (none which apply). Stopped taking lipitor seven days ago and still experiencing pain in shoulder and tingling and num8ness down right arm radiating into fingers. No more medications. will attempt holistic approach. vitamin c and/or niacin. Extremely severe pain in right shoulder as if from extreme workout or injury (none which apply). Stopped taking lilitor seven days ago and still experiencing pain ln shoulder and timvling and numbness down right arm radiating into fingers. No more medications. will attempt holistic approach. vitamins f and/or niacin.
	Word	Extremely severe pain in right shoulder as if from extreme workout or injury (none which apply). Stopped taking lipitor seven days ago and still experiencing shoulder painful and pins and needles and sensation loss down right arm radiating into fingers. No more medications. will attempt holistic approach. vitamins c and/or β-pyridinecarboxylic acid. Extremely severe pain in right shoulder as if from extreme workout or injury (none which apply). Stopped taking lipitor seven days ago and still experiencing pain in shoulder and tingling and numbness down right arm radiating into fingers. No more medications. will attempt holistic approach. vitamin c and/or niacin.

Table 1 (Continued)

Aug. Level	Tech.	Augmented Text
		Extremely severe pain in right shoulder as if from extreme workout or injury (none which apply). Stopped taking lipitor seven days ago and still experiencing pain in shoulder and tingling and numbness down right arm radiating into fingers. No more medications. will attempt holistic approach. vitamin c and/or niacin.
Combine	Char	Extremely 8eveke pain in right shoulder as if fkum extkeme workout uk injury (nune which apply). Stopped taking 1ipituk 8even days ago and still experiencing pain in shoulder and tin91in9 and nom6nes8 down right arm kadiatin9 into fingers. No more medications. wi11 attempt holistic approach. vitamin8 c and/or niacin. Extremely severe pain in right shoulder as if fomr extreme workout or injury (none which apply). Stopped taking 1ipit0k seven days ago and still experiencing pain in shoulder and tin91in9 and nombne88 down right arm radiating into fingers. No more medications. will attempt holistic approach. vitamins c and/or niacin. Extremely sfvere pain in right shoulder as if ff;m dxtre,e workout od injury (glne which apply). Stopped taking 1ipit0k weven days ago ane still experiencing pain in shuoldek and tin91in9 and nom6ne8s down right arm rasiqting into fingers. No more medications. sill attwhoh holistic approach. vitamins c and/or niacin.
	Word	Extremely severe pain in right shoulder as if from extreme workout or injury (none which apply). Stopped taking lipitor seven twenty-four hours ago and still experiencing shoulder painful and pins and needles and numbness down right build up radiating into fingers. No more medications. will attempt holistic approach. vitamins c and/or 3-Pyridylcarboxylic acid. Extremely severe renal colic in right shoulder as if from extreme workout or injury (none which apply). Stopped taking lipitor seven days ago and abreact experiencing shoulder sense of pain and pins and needles and loss of sensation down right rearm radiating into fingers. No more medications. will attempt holistic approach. vitamins c and/or β-pyridinecarboxylic acid. Extremely severe symptom in right shoulder as if from extreme workout or injury (none which apply). Stopped taking lipitor seven days ago and still experiencing shoulder pain and pins and needles and sensation loss down right limb radiating into fingers. No more medications. will attempt holistic approach. vitamins c and/or β-pyridinecarboxylic acid.

Table 2: Original Text 2 and Its Augmentations

Original Text: Pain in hip, lower back, knees & elbow. Stiffness in lower back and legs when getting up in the morning. Excercise intolerance and general muscle weakness. I ask my doctor about these pains months ago and he said lipitor would not cause my problems. However, after reading the common side effects with others on this site, I will stop this medication immediately, it not worth feeling like an old man at age 46!		
Aug. Level	Tech.	Augmented Text
Context	Char	Pain in hip, lower back, knees & e16ow. Stiffness in lower back and legs when getting up in the muknin9. Excercise intolerance and general muscle weakness. I ask my doctor about the8e pains months ago and he said lipitor would not cause my problems. However, after reading the common side effect8 with others on thi8 site, I will 8t0p this medication immediately, it not worth fee1in9 like an u1d man at age 46! Pain in hip, lower back, knees & elbow. Stiffness in lower back and legs when getting up in tghe morning. Excercise intolerance and general muscle weakness. I ask my doctor about theese pains months ago and he said lipitor would not cause my problems. However, after reading the common side effects with others on thsi site, I will stop this medication immediately, it not worth feeling like an old man at age 46! Pain in hip, lower back, knees & e;biw. Stiffness in lower back and legs when getting up in rhe morjigg. Excercise intolerance and general muscle weakness. I ask my doctor about yhese pains months ago and he said lipitor would not cause my problems. However, after reading the common side rfcfcys wifh others on ghls site, I will stip this medication immediately, it not worth reelinv like an pld man at age 46!
	Word	Pain in hip, lower back, knees & elbow. Stiffness in lower back and legs when getting up in the morning. Excercise intolerance and general muscle weakness. I ask my doc about these pains months ago and he said lipitor would not cause my job. However, after reading the common side effects with others on this site, I will stop this medication immediately, it non worth feeling like an old man at age 46! Pain in hip, lower back, knees & elbow. Stiffness in lower back and legs when getting up in the morning. Excercise intolerance and general muscle weakness. I ask my doctor about these pains months ago and he said lipitor would not cause my problems. However, after reading the common side effects with others on this site, I will stop this medication immediately, it not worth agitation like an old man at age 46!

Table 2 (Continued)

Aug. Level	Tech.	Augmented Text
		Pain in hip, lower back, knees & elbow. Stiffness in lower back and legs when getting up in the morning. Excercise intolerance and general muscle weakness. I ask my doctor about these pains months ago and he said lipitor would not cause my problems. However, after reading the common side effects with others on this site, I will stop this medication immediately, it not worth state like an old man at age 46!
Entity	Char	Pain in hip, lower back, knees & elbow. stiffness in 1ower back and legs when getting up in the morning. excercise intu1ekance and general muscle wearnes8. I ask my doctor about these pain8 months ago and he said 1ipit0r would not cause my problems. However, after reading the common side effects with others on this site, I will stop this medication immediately, it not worth feeling like an old man at age 46! Pain in hip, lower back, knees & elbow. stiffness in lower bakc and legs when getting up in the morning. excercise intolerance and general muscle weakness. I ask my doctor about these pains months ago and he said lipitor would not cause my problems. However, after reading the common side effects with others on this site, I will stop this medication immediately, it not worth feeling like an old man at age 46! Pain in hip, lower back, knees & elbow. stiffness in lower bacl and legs when getting up in the morning. eacegcisd intolerqnce and generql musc,e deakness. I ask my doctor about these pains months ago and he said .upitod would not cause my problems. However, after reading the common side effects with others on this site, I will stop this medication immediately, it not worth feeling like an old man at age 46!
	Word	Pain in hip, lower back, knees & elbow. stiffness in lower back structure and legs when getting up in the morning. excercise Hypersensitivity and generalized muscular weakness. I ask my doctor about these pain months ago and he said lipitor would not cause my problems. However, after reading the common side effects with others on this site, I will stop this medication immediately, it not worth feeling like an old man at age 46! Pain in hip, lower back, knees & elbow. stiffness in incline back and legs when getting up in the morning. excercise intolerance and Blucher sphincter weakness. I ask my doctor about these pain months ago and he said lipitor would not cause my problems. However, after reading the common side effects with others on this site, I will stop this medication immediately, it not worth feeling like an old man at age 46!

Table 2 (Continued)

Aug. Level	Tech.	Augmented Text
		Pain in hip, lower back, knees & elbow. stiffness in move back and legs when getting up in the morning. excercise intolerance and general muscle weakness. I ask my doctor about these pain months ago and he said lipitor would not cause my problems. However, after reading the common side effects with others on this site, I will stop this medication immediately, it not worth feeling like an old man at age 46!
Combine	Char	Pain in hip, lower back, knees & e16ow. stiffness in 10wer back and legs when getting up in the muknin9. excekci8e intolerance and general mo8cle weakness. I ask my doctor about the8e pain8 months ago and he said 1ipit0k would not cause my problems. However, after reading the common side effect8 with others on thi8 site, I will 8t0p this medication immediately, it not worth fee1in9 like an u1d man at age 46! Pain in hip, lower back, knees & elbow. stiffness in 10wer back and legs when getting up in tghe morning. excercise intu1ekance and general mosc1e weakness. I ask my doctor about theese pain8 months ago and he said 1ipit0k would not cause my problems. However, after reading the common side effects with others on thsi site, I will stop this medication immediately, it not worth feeling like an old man at age 46! Pain in hip, lower back, knees & e;biw. 8tiffne88 in lower back and legs when getting up in rhe morjigg. excekci8e intolerance and general muscle wearne88. I ask my doctor about yhese pain8 months ago and he said 1ipit0k would not cause my problems. However, after reading the common side rfcfcys wifh others on ghls site, I will stip this medication immediately, it not worth reelinv like an pld man at age 46!
	Word	Pain in hip, lower back, knees & elbow. stiffness in lower back structure and legs when getting up in the morning. excercise Hypersensitivity and generalized muscle weaknesses. I ask my doc about these pain months ago and he said lipitor would not cause my job. However, after reading the common side effects with others on this site, I will stop this medication immediately, it non worth feeling like an old man at age 46! Pain in hip, lower back, knees & elbow. stiffness in lower back structure and legs when getting up in the morning. excercise intolerance, function and generalized paresis. I ask my doctor about these pain months ago and he said lipitor would not cause my problems. However, after reading the common side effects with others on this site, I will stop this medication immediately, it not worth agitation like an old man at age 46!

Table 2 (Continued)

Aug. Level	Tech.	Augmented Text
		Pain in hip, lower back, knees & elbow. stiffness in lower back structure and legs when getting up in the morning. excercise intolerance, function and generalized weakness muscles. I ask my doctor about these pain months ago and he said lipitor would not cause my problems. However, after reading the common side effects with others on this site, I will stop this medication immediately, it not worth state like an old man at age 46!

Appendix B
Annotation Guidelines

This appendix provides the annotation guideline utilized for the expert evaluation conducted in Chapter 5.

Introduction

The annotation guidelines for evaluating the performance of our Medical Entity Linking (MEL) system are described in this document. Our MEL system was designed to assist laypeople in becoming familiar with medical terminology. The MEL system accepts text input from social media and will retrieve highlighted terms indicating informal medical phrases from the input text. It contains information about the Systematized Nomenclature of Medicine - Clinical Terms (SNOMED-CT) code, the SNOMED-CT description as a formal term, and a predicted Wikipedia article associated with the term as a source of the medical term explanation for each highlighted term. SNOMED-CT is a standardized, multilingual clinical terminology vocabulary used by physicians and other health care providers for electronic clinical health information exchange.

We intend to evaluate our MEL system's performance before delivering it to laypeople. Therefore, we conduct this annotation process to assess the correctness of the predicted SNOMED-CT descriptions and correctly select the Wikipedia articles associated with the highlighted terms.

Workflow

1. The annotators will be given a link and access to the annotation tool. Before you can begin the annotation process, you must click the 'Annotation' button and then log in to the system by entering your name and token.

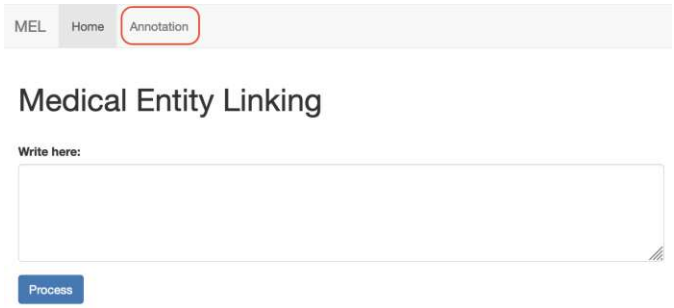


Figure 1: Landing page

Figure 2: Login

2. After logging in, annotators are able to access their account information. The list of documents to be annotated will be discovered by the annotators. For each file, the system will display the annotation progress.

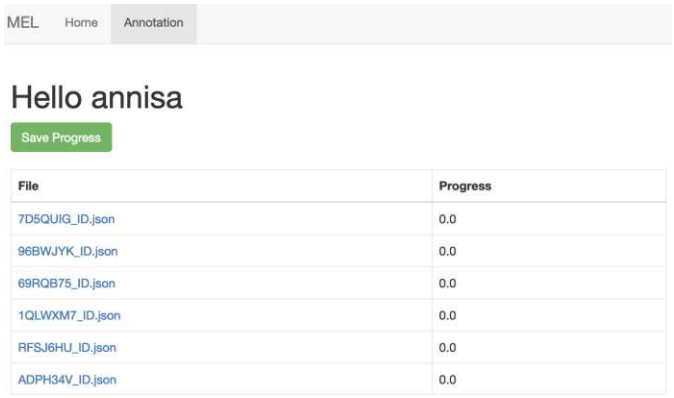


Figure 3: List of Annotation Documents

3. For each selected document, the system will display the social media post containing the highlighted terms. The highlighted term is a predicted informal medical phrase produced by our model. The highlighted term may relate to symptoms, disease, disorder, finding, drug, or substance.

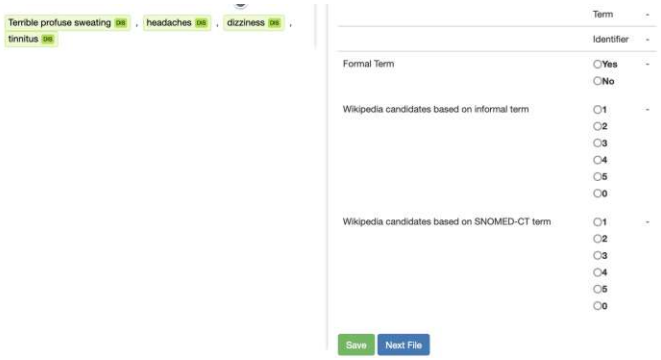


Figure 4: Selected Document

4. To begin the annotation process, the annotator must select each of the highlighted terms.

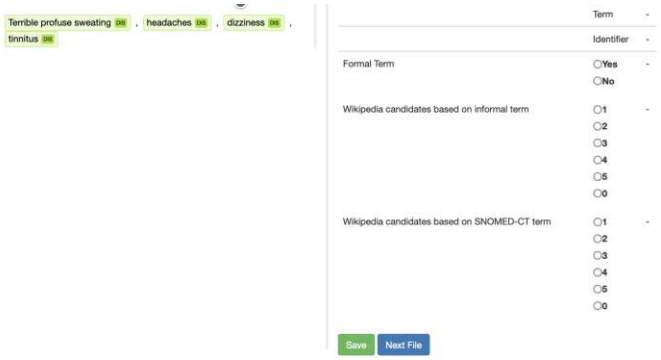


Figure 5: Term Selection

Once selected, annotators must indicate whether the highlighted term is correctly assigned to a formal medical term or not. The annotators must then choose the appropriate Wikipedia article position for the highlighted and formal terms.

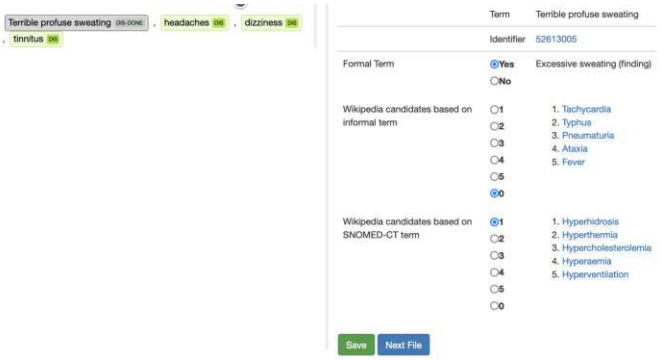


Figure 6: Annotated Term

Finally, click 'Save' and then click another term to continue the annotation process.

5. Once all of the terms have been annotated, the annotator can move on to the next file by clicking the 'Next File' button.

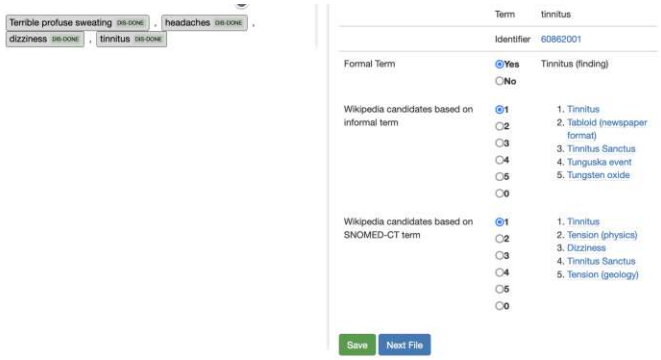


Figure 7: All Annotated Terms - Next File

The annotation guidelines are described in detail in the General guidelines section.

Definitions

Each field in the annotation box is defined below:

1. **Term**: highlighted term, which indicates an informal medical term.
2. **Identifier**: SNOMED-CT code for the formal term related to the **Term**.
3. **Formal Term**: SNOMED-CT descriptions of **Identifier**.
4. **Wikipedia candidates based on informal term**: list of Wikipedia articles that were retrieved using **Term**.
5. **Wikipedia candidates based on SNOMED-CT term**: list of Wikipedia articles that were retrieved using a **Formal term**.

General guidelines

The goal of the first task is to annotate the **correctness** of predicted SNOMED-CT description in relation to the highlighted term, which is considered an informal medical term. The goal of the second task is **finding the correct rank** of Wikipedia articles with respect to the input term.

General rules for correctness of SNOMED-CT Mapping

- In relation to the highlighted term, the Formal term must be annotated with 'Yes' or 'No'.
- 'Yes' indicates that the **Formal Term** is correct for the highlighted term, while 'No' indicates that the **Formal Term** is incorrect for the highlighted term.
- If there is doubt about the predicted formal term, annotators could use the SNOMED-CT browser to double-check the correct codes. Link: `https://bit.ly/3d6e9v7`.

General rules for selecting appropriate Wikipedia articles

There are two types of the input term, namely the **highlighted term** and **Formal Term**. Each input has a list of Wikipedia articles. The field labeled with **Wikipedia candidates based on informal term** means the list of the Wikipedia candidates is generated from the highlighted term as the input. Meanwhile, for **Wikipedia candidates based on SNOMED-CT term**, it means the list of the candidates is generated from the SNOMED-CT description.

Annotated for Wikipedia candidates based on informal term

- The annotator is required to select the appropriate Wikipedia article for the highlighted term.
- It is annotated with '1', '2', '3', '4', or '5' based on its position within the 'Wikipedia candidates based on informal term'.
- Annotators must annotate with '0' if no proper Wikipedia article is found in the lists.
- Since Wikipedia articles may not cover all medical terms, the annotator may choose the article that comes closest to covering the highlighted term.

Annotated for Wikipedia candidates based on SNOMED-CT term

- The annotator must choose the correct Wikipedia article for the **correct** Formal term (annotated with 'Yes') that corresponds to the highlighted term.
- It is annotated with '1', '2', '3', '4', or '5' based on its position within the 'Wikipedia candidates based on SNOMED-CT term'.
- Annotators must annotate with '0' if no proper Wikipedia article is found in the lists.
- Since Wikipedia articles may not cover all medical terms, the annotator may choose the article that comes closest to covering the highlighted term.
- For the **incorrect** Formal term associated with the highlighted term (annotated with 'No'), the annotator must mark it with '0'.

Appendix C
User Experiment Questionnaire

This appendix provides the questionnaire that was administered to the participants who took part in the user experiments conducted in Chapter 6.

Demographic Survey

Survey Questions

1. Your age (years):
    - 18 - 21 years old
    - 22 - 34 years old
    - 35 - 44 years old
    - 45 - 54 years old
    - 55 - 64 years old
    - 65 and over
2. Your gender:
    - Male
    - Female
    - Other
    - Prefer not to answer
3. What is your highest level of education?
    - Elementary school
    - Middle school/Junior high school
    - High school/Senior high school
    - Diploma/Vocational/Technical school
    - Bachelor degree
    - Master degree

- Doctoral/PhD
- Other

4. In what country do you currently reside?
    - List of countries truncated for brevity
5. Which area do you primarily work in (regardless of your actual position)
    - Information Technology (IT)
    - Healthcare and Medical Services
    - Finance and Accounting
    - Sales and Marketing
    - Education and Teaching
    - Engineering
    - Business Management and Administration
    - Media and Entertainment
    - Other
6. Your net annual income (in US Dollars)
    - Under $15,000
    - Between $15,000 and $29,999
    - Between $30,000 and $49,999
    - Between $50,000 and $74,999
    - Between $75,000 and $99,999
    - Over $100,00
    - I do not know
    - Prefer not to answer
7. Do you use Internet for work/studies?
    - Yes
    - No
    - I do not work/study at the moment, I use Internet only for my personal inquiries
8. What is your level of English?
    - No knowledge of English
    - Basic knowledge of English
    - Average knowledge of English
    - Good knowledge of English
    - Fluent English
    - Native English

9. How often do you search online?
    - Every day
    - A few times a week
    - About once a week
    - A few times a month
    - Once a month
    - Less than once a month
10. How often do you search online in your own language?
    - Every day
    - A few times a week
    - About once a week
    - A few times a month
    - Once a month
    - Less than once a month
11. How often do you search for or read any information on the Internet in English?
    - Every day
    - A few times a week
    - About once a week
    - A few times a month
    - Once a month
    - Less than once a month
12. How often do you search for online health information regarding your/your family's/friends' health?
    - Every day
    - A few times a week
    - About once a week
    - A few times a month
    - Once a month
    - Less than once a month
13. What types of online health information do you look for?
    - Specific disease or medical problem
    - Certain medical treatment or procedure
    - How to lose weight or how to control your weight
    - Food safety or recalls
    - Drug safety or recalls / a drug you saw advertised

- Medical tests results
- Caring for an aging relative or friends
- Pregnancy and childbirth
- Other

Text 1: Ulcerative colitis (UC) is a type of inflammatory bowel disease (IBD). IBD comprises a group of diseases that affect the gastrointestinal (GI) tract. UC occurs when the lining of your large intestine (also called the colon), rectum, or both become inflamed. This inflammation produces tiny sores called ulcers on the lining of your colon. Inflammation usually begins in the rectum and spreads upward. It can involve your entire colon. The inflammation causes your bowel to move its contents rapidly and empty frequently. As cells on the surface of the lining of your bowel die, ulcers form. The ulcers may cause bleeding and discharge of mucus and pus. While this condition affects people of all ages, most people develop UC between ages 15 and 30 years old, according to the American Gastroenterological Association. After 50 years old, there's another small increase in diagnosis of IBD, usually in men.

14. How well do you think you understand what the Text 1 is about?

Not at all 1 2 3 4 Very well

15. How well do you understand medical terminology in the Text 1?

Not at all 1 2 3 4 Very well

Text 2: Ulcerative colitis is a chronic idiopathic inflammatory bowel disease characterized by continuous mucosal inflammation that starts in the rectum and extends proximally. Typical presenting symptoms include bloody diarrhea, abdominal pain, urgency, and tenesmus. In some cases, extraintestinal manifestations may be present as well. In the right clinical setting, the diagnosis of ulcerative colitis is based primarily on endoscopy, which typically reveals evidence of continuous colonic inflammation, with confirmatory biopsy specimens having signs of chronic colitis. The goals of therapy are to induce and maintain remission, decrease the risk of complications, and improve quality of life. Treatment is determined on the basis of the severity of symptoms and is classically a step-up approach. 5-Aminosalycilates are the mainstay of treatment for mild to moderate disease. Patients with failed 5-aminosalycilate therapy or who present with more moderate to severe disease are typically treated with corticosteroids followed by transition to a steroid-sparing agent with a thiopurine, anti-tumor necrosis factor agent, or adhesion molecule inhibitor. Despite medical therapies, approximately 15% of patients still require proctocolectomy. In addition, given the potential risks of complications from the disease itself and the medications used to treat the disease, primary care physicians play a key role in optimizing the preventive care to reduce the risk of complications.

16. How well do you think you understand what the Text 2 is about?

Not at all 1 2 3 4 Very well

17. How well do you understand medical terminology in the Text 2?

Not at all 1 2 3 4 Very well

Feedback Survey for *Intervention Group*

Section A: Learning system Feedback

1. I found the formal medical terminology were easy to remember

   Strongly disagree 1 2 3 4 Strongly agree

2. I found that the Wikipedia explanation was easy to understand

   Strongly disagree 1 2 3 4 Strongly agree

3. I think I gained knowledge of medical terms after using the training system

   Strongly disagree 1 2 3 4 Strongly agree

4. I believe that having knowledge and understanding of formal medical terms will help enhance my ability to comprehend and navigate medical information

   Strongly disagree 1 2 3 4 Strongly agree

5. I believe that having knowledge of formal medical terms will improve my ability to communicate with doctors and other medical professionals

   Strongly disagree 1 2 3 4 Strongly agree

6. All the medical phrases in learning phase were new to me

   Strongly disagree 1 2 3 4 Strongly agree

Section B: Integrating the learning of medical terms into the social media

Below you can take a look at the terms highlighted in the green box. They are health-related terms that you sometimes see on social media. If you click on them, you'll get a definition. After trying it out, we'd like to know what you think about using this feature on social media sites like AskAPatient.com to better understand medical terms in everyday situations.

1. I think learning medical terminology from social media would be beneficial

   Strongly disagree 1 2 3 4 Strongly agree

2. I believe that understanding formal medical terms and their explanations can be useful in learning medical terminology

   Strongly disagree 1 2 3 4 Strongly agree

3. I think the feature of providing formal medical terms when the term is clicked on is necessary

   Strongly disagree 1 2 3 4 Strongly agree

4. I think the feature of providing term explanations when the term is clicked on is necessary

   Strongly disagree 1 2 3 4 Strongly agree

5. I think that the explanations are useful and understandable

   Strongly disagree 1 2 3 4 Strongly agree

6. I find easy to access additional information, including formal medical terms and definitions

   Strongly disagree 1 2 3 4 Strongly agree

7. I think that the formal medical terms and their explanations can improve the comprehension of medical terminology

   Strongly disagree 1 2 3 4 Strongly agree

8. Which types of entities would you be interested in learning more about to enhance your understanding of medical terminology?

   - Medical conditions
   - Symptoms
   - Medical Treatment
   - Medications
   - Medical procedures
   - Medical professionals
   - Medical devices

9. Which additional sources of information would you like to have that helps you learn medical terminology from social media?

   - WebMD
   - Mayo Clinic
   - MedlinePlus
   - Centers for Disease Control and Prevention (CDC)
   - National Institutes of Health (NIH)
   - Healthline
   - Others

Feedback Survey for *Non-Intervention Group*

Section A: Practice system Feedback

1. Do you think it's important to be familiar with medical terminology?
    - Yes
    - No
    - I'm not sure
2. How important do you think it is to have a basic understanding of medical terminology?
    - Extremely important
    - Somewhat important
    - Not very important
    - Not at all important
3. Would you like additional information to help you understand medical terminology?
    - Yes
    - No
    - I'm not sure
4. Which type of source would you like to use?
    - Wikipedia
    - Healthline
    - Medline Plus
    - Merriam-Webster
    - Other
5. Would you like to learn medical terminology to improve your knowledge about health-related topics?
    - Yes
    - No
    - I'm not sure
6. If you were interested in learning about medical terminology, what type of system would you prefer to use?

Section B: Integrating the learning of medical terms into the social media

Similar to the one in Section 7.3.3.