ОДЕСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ імені І.І.МЕЧНИКОВА

Факультет математики, фізики та інформаційних технологій
(повне найменування вищого навчального закладу)

Кафедра математичного забезпечення комп'ютерних систем
(повна назва кафедри (предметної, циклової комісії))

Кваліфікаційна робота

на здобуття рівня вищої освіти «магістр
(рівень вищої освіти)

Research on Feature Extraction Technology of Medical Record Data Based on Big Data and NLP Natural Language Recognition

Дослідження технології виділення функцій із даних медичних записів на основі великих даних і розпізнавання природної мови NLP

Виконав: студент денної форми навчання
спеціальності 126 – Інформаційні системи та технології .
(шифр і назва напряму підготовки, спеціальності)
Освітня програма _« Інформаційні системи та технології»_
(назва освітньої програми)
ВАН Чуньмей
(прізвище, ім'я, по-батькові)

Керівник к.ф.-м.н., доц. Петрушина Т.І.
(науковий ступінь, вчене звання, прізвище та ініціали, підпис)

Рецензент к.ф.-м.н., доц. Антоненко О. С.
(науковий ступінь, вчене звання, прізвище та ініціали)

Рецензент Рецензент д.т.н., проф. Арсірій О. О. _____
(науковий ступінь, вчене звання, прізвище та ініціали)

Рекомендовано до захисту:
Протокол засідання кафедри
№_1_ від «_6_»_12__ 2024 р.

Завідувач кафедри
_____ _ Євгеній МАЛАХОВ_
(підпис) (ім'я, прізвище)

Захищено на засіданні ЕК №. __
протокол №__ від «___» ____ 2024 р.
Оцінка_____/____/_____
(за національною шкалою, шкалою ECTS, бали)

Голова ЕК
_____ _Володимир ВИЧУЖАНІН_
(підпис) (ім'я, прізвище)

Одеса - 2024

MINISTRY OF EDUCATION AND SCIENCE OF UKRAINE

ODESA I.I. Mechnikov NATIONAL UNIVERSITY

(full name of higher education institution)

Faculty of mathematics, physics and information technologies

(full name of the institute, name of the faculty)

Department of mathematical support of computer systems

(full name of the department)

Qualification work

for obtaining the "master" higher education level

(education level)

on the topic "Research on Feature Extraction Technology of Medical Record Data Based on Big Data and NLP Natural Language Recognition"

Performed by: a full-time student of the
specialty 126 – Information Systems & Technologies

(specialty code and name)

educational program «Information Systems & Technologies»

(educational program name)

Wang Chunmei

(Full Name)

Scientific supervisor PhD (Math.), Assoc. Prof. Tatiana Petrushina

(academic degree, academic title, surname and initials, signature)

Reviewer PhD (Math.), Assoc. Prof. Oleksandr Antonenko

(academic degree, academic title, surname and initials)

Reviewer PhD (Tech.), Prof., ScD. (Tech.) Elena Arsiriy

(academic degree, academic title, surname and initials)

Recommended for defense:
Minutes of the department meeting
No.___ of "___"_____ 2024

Head of Department
_____ _Eugene MALAKHOV_
(signature) (surname, initials)

Defended at the EC No. ___ meeting
Minutes No. __ of "___" _____ 2024
Score _____/____/_____

(by national scale, ECTS scale, points)

Chairman of the EC
_____ _Volodymyr VYCHUZHANIN_
(signature) (surname, initials)

Odesa - 2024

ABSTRACT

The topic of "Research on Feature Extraction Technology of Medical Record Data Based on Big Data and NLP Natural Language Recognition " is considered in the thesis.

This article investigates a medical record data feature extraction technique based on big data and NLP natural language recognition. We propose a scheme to use Scispacy's English medical model to extract keywords from the original text medical records for the first time through NLP natural language recognition technology, and then use HPO standard words and OMIM disease words to extract keywords from the medical records for the second round through NLP natural language recognition technology. This article provides a detailed introduction to the processes of data preprocessing, model design, training, and validation, and discusses the results and future research directions. This experiment selected some long text electronic medical records and HPO medical records from the company's historical orders, and validated and compared the keywords selected from the long text with the HPO medical records. The experimental results showed that the algorithm can have good accuracy in data feature extraction and can assist medical tasks in feature extraction and classification of large text medical records.

АННОТАЦІЯ

У магістерській роботі розглядається тема «Дослідження технології вилучення ознак із даних медичних записів на основі великих даних та розпізнавання природної мови NLP».

У роботі досліджується метод вилучення ознак із даних медичних записів на основі великих даних та розпізнавання природної мови NLP. Пропонується схема використання англійської медичної моделі Scispacy для отримання ключових слів з вихідних текстових медичних записів за допомогою технології розпізнавання природної мови NLP, а потім використання стандартних слів HPO та слів захворювань OMIM для вилучення ключових слів з медичних записів для другого раунду за допомогою технології розпізнавання природної мови NLP. У цій роботі дається докладне введення в процеси попередньої обробки даних, проектування моделі, навчання та перевірки, а також обговорюються результати та майбутні напрямки досліджень. Для проведення тестування було вибрано деякі довгі текстові електронні медичні записи та медичні записи HPO з історичних замовлень компанії, а також перевірено та порівняно ключові слова, вибрані з довгого тексту, з медичними записами HPO. Експериментальні результати показали, що алгоритм може мати хорошу точність при добуванні ознак даних і може допомогти в медичних завданнях вилучення ознак і класифікації великих текстових медичних записів.

CONTENT

KEY WORDS 8
1. INTRODUCTION 9
1.1 The Importance of Medical Record Data Analysis 11
1.1.1 Improving diagnostic accuracy 11
1.1.2 Personalized treatment 12
1.1.3 Reduce costs 12
1.1.4 Supporting medical research 13
1.1.5 Public health monitoring 13
1.2 Challenges Faced by Medical Record Data Analysis 13
1.2.1 Data volume and complexity 15
1.2.2 Data privacy and security 15
1.2.3 Data standardization and interoperability 15
1.2.4 Data quality 15
1.2.5 Analytical techniques 15
1.2.6 Interpretability and interpretability 16
1.2.7 Legal and ethical issues 16
1.2.8 Lack of professionals 16
1.3 Challenges Faced by Medical Record Data Analysis 16
1.3.1 Improving diagnostic accuracy and efficiency 17
1.3.2 Improving healthcare services 17
1.3.3 Expanding the coverage of medical services 17
1.3.4 Promoting medical research and personalized treatment 18
1.3.5 Coping with the challenges of big data 18
1.3.6 Data augmentation and transfer learning 18
1.4 Related work 19
1.4.1 Integration of Electronic Medical Records and Medical Imaging Data . 19
1.4.2 Opportunities and challenges of big data analysis in healthcare 20
1.4.3 Application of Convolutional Neural Networks in Medical Image Analysis 20
1.4.4 Development trend of intelligent technology application in the medical field 20
1.4.5 Current Status Analysis and Prospects of Medical Data Visualization Research 21

2 RESEARCH METHOD 21
2.1 The preparation of the basic database 21
2.1.1 Preparation of HPO standard medical record phenotype word database 22
2.1.2 Preparation of OMIM disease database 23
2.2 The preparation of the research datas 24
2.2.1 Data collection 25
2.2.2 Data cleaning 25
2.2.3 Data standardization 27
2.3 Methodology 28
2.3.1 Use SciSpacey for Named Entity Recognition 28
2.3.2 Compare with HPO and OMIM libraries using fuzzywuzzy software .. 28
2.3.3 Build a negative word library 29
2.4 Result 30
3 EXPERIMENT 30
3.1 NLP Experimental steps 30
3.2 BigData Experimental steps 36
4 RESULT 54
4.1 Implications of Accurate Extraction and Labeling 54
4.2 Enhancing Clinical Decision Support Systems 55
4.3 Advancing Medical Research 55
4.4 Potential Applications in Healthcare 56
4.5 Future Directions 57
5 FUTURE WORK 57
5.1 Model Optimization 58
5.1 Multimodal Data Fusion 58
5.2 Cross-Disease Generalization Ability 59
5.3 Clinical Validation 59
5.4 Data Privacy and Security 60
5.5 Model Interpretability 60
5.6 Cross-Cultural and Cross-Population Model Adaptability 60
6 DISCUSS 61
6.1 Integration of Scispacy and Fuzzywuzzy 61
6.2 Enhancing Recognition Accuracy 62
6.3 Improving Data Analysis Quality and Efficience 62

6.4 Addressing Variability in Medical Records 62
6.5 Potential for Automation in Data Cleaning 62
6.6 Applications in Clinical Decision Support Systems 63
6.7 Enhancing Research and Drug Discovery 63
6.8 Challenges and Limitations 63
6.9 Ethical Considerations 63
6.10 Future Developments 64
7 CONCLUSION 64
REFERENCES 67

KEY WORDS

Big data, NLP natural language, medical record data, feature extraction

INTRODUCTION

The clinical phenotype of patients is crucial for the precise diagnosis of genetic diseases. Accurate interpretation of genetic variations requires the assistance of precise clinical phenotype data. Currently, there are multiple international methods for extracting patient phenotype data, such as openAI. However, due to the confidentiality requirements of medical data, the extraction of key terms from patient phenotypes and their further optimization to HPO (Human Phenotype Ontology) standard terms often need to be completed within an internal local area network. It is not allowed for patient medical record-related data to flow into the internet's third-party interfaces. Therefore, each medical institution needs its own clinical keyword extraction technology. Currently, due to some limitations, most institutions use manual methods for extracting patient phenotype keywords, which not only consumes a lot of medical staff's energy but also, due to varying levels of expertise and fatigue, can easily lead to the omission of keywords. This paper presents a technology for extracting keywords through NLP (Natural Language Processing) and big data technology, which efficiently and high-quality extracts clinical keywords.

The research object is the technology of extracting medical keywords through NLP natural language processing.

The main theme of the research is to use NLP and big data research methods to extract medical keywords.

To achieve this goal, the following tasks must be completed:

- Research existing Python modules for extracting noun entities from medical records.
- Construct a big data HPO standard term library and an OMIM term library. Use NLP natural language recognition methods to match and obtain medical noun keywords from noun entities. Construct a negation term library (negation terms are key information in medical records)

- Analyze the effectiveness of keyword extraction by big data of our history datas. By comparing the NLP matching degree between historical HPO words and selected keywords from medical records, and analyzing the accuracy of each NLP matching score through sampling, the accuracy of medical record keyword selection is calculated
- Construct simulated data of big data historical clinical medical records and corresponding HPO standardized medical record term data.
- Analyze the probability of matching between the extracted medical keywords from clinical records and HPO standard medical record term data through NLP, and construct a term frequency correspondence library between natural medical record language keywords and HPO standard medical record terms.
- Obtain the medical record key word library, noise word library, and the need to expand the recognition word library, and evaluate the accuracy of keyword recognition again with a large amount of data

1. BACKGROUND INTRODUCTION

1.1 The Importance of Medical Record Data Analysis

Medical record data is an essential component of the healthcare system, as it records a patient's medical history and treatment process. With the popularity of electronic health records (EHR), the volume and complexity of medical record data continue to increase, posing challenges for data management and analysis. Medical record data analysis plays a crucial role in the healthcare field, involving the extraction, analysis, and interpretation of information from medical records to support clinical decision-making, improve patient care, optimize resource allocation, and drive medical research. NLP and big data technologies provide potential solutions for processing this data, making it easier to access and analyze while protecting patient privacy. The following are several key importance of medical record data analysis.

1.1.1 Improving diagnostic accuracy

By analyzing medical record data, doctors can obtain more comprehensive patient information, including medical history, symptoms, laboratory test results, etc., thereby improving diagnostic accuracy.

Medical record data analysis is crucial for improving the accuracy of medical diagnosis. By conducting in-depth analysis of patients' medical records, including detailed medical history, symptom descriptions, laboratory test results, and past treatment plans, doctors can obtain a comprehensive view of patient information. This comprehensive information helps to reveal potential disease patterns and associations, thereby assisting doctors in identifying complex or rare diseases. In addition, the application of NLP technology can further extract key information from unstructured text, making the diagnostic process more efficient and accurate. This detailed analysis of medical record data can not only reduce the

risk of misdiagnosis and missed diagnosis, but also provide patients with more personalized treatment plans, ultimately improving the overall quality of medical services.

1.1.2 Personalized treatment

Medical record data analysis can help doctors develop personalized treatment plans based on the specific situation of patients, improving treatment effectiveness.

Medical record data analysis can not only help doctors develop personalized treatment plans based on the specific situation of patients, improve treatment effectiveness, but also provide support for precision medicine through in-depth analysis of multidimensional data such as patients' genetic information, lifestyle habits, and physiological characteristics. For example, in tumor treatment, by conducting in-depth analysis of the patient's genome and pathological characteristics, doctors can accurately select suitable targeted drugs and immunotherapy regimens. This personalized treatment method can significantly improve treatment effectiveness and patient survival rates, while reducing unnecessary drug side effects. In addition, environmental sensing technology and lifestyle analysis can also help doctors develop personalized treatment plans based on patients' environmental and lifestyle habits. Through the analysis of these comprehensive data, personalized treatment can more accurately meet the health needs of patients, improve treatment effectiveness, and also enhance the quality of life of patients.

1.1.3 Reduce costs

Data analysis can help identify unnecessary medical procedures and resource waste, thereby reducing healthcare costs.

Data analysis can help identify unnecessary medical procedures and resource waste, thereby reducing healthcare costs. In medical practice, reading and analyzing raw medical records by medical personnel is a time-consuming and complex task. Medical records contain a large amount of unstructured text, from which medical personnel need to extract key information such as medical history, symptoms, laboratory test results, etc. This not only consumes a lot of energy, but also increases the burden on human resources. By applying NLP and big data technology, key information can be automatically extracted and analyzed from medical records, reducing the time investment of medical staff in medical record review and allowing them to focus more on patient care and treatment. In addition, data analysis can reveal the usage patterns and efficiency of medical resources, identify inefficient or unnecessary medical procedures, thereby avoiding resource waste and further reducing medical costs. The application of this technology not only improves the efficiency of medical record analysis, but also helps optimize the allocation of medical resources and achieve cost control

1.1.4 Supporting medical research

Medical record data is an important resource for medical research, which can be used for new drug development, disease mechanism research, and clinical trials.

1.1.5 Public health monitoring

Medical record data analysis is crucial for monitoring and preventing public health events such as epidemics and epidemics

1.2 Challenges Faced by Medical Record Data Analysis

Despite the importance of medical record data analysis mentioned above, it also faces a series of challenges in practice.

Medical record data analysis is a multidisciplinary field that requires the joint efforts of experts from multiple fields such as medicine, computer science, data science, and law to overcome the aforementioned challenges and fully unleash its potential in healthcar.

1.2.1 Data volume and complexity

With the popularity of electronic health records (EHRs), the volume of medical record data has increased sharply, and the complexity of data has also increased, posing challenges to storage, processing, and analysis. This growth in data volume requires more efficient data processing techniques, larger storage capacity, and more complex data analysis methods to extract useful information.

1.2.2 Data privacy and security

Medical record data contains sensitive personal health information, and protecting patient privacy and data security is a major challenge. Both internal security risks and external threats can lead to data breaches, therefore it is necessary to strengthen technology research and talent development, and improve the level of data security and privacy protection.

1.2.3 Data standardization and interoperability

Different medical institutions may use different recording systems and formats, leading to issues with data standardization and interoperability. The total amount of physical data aggregation in China is growing rapidly, but data resources cannot be effectively integrated and analyzed due to inconsistent collection standards, diverse data structures, and semantic information loss or ambiguity, greatly limiting the exploration of potential value of data.

1.2.4 Data quality

The accuracy and completeness of medical record data are crucial for analyzing results, but data entry errors, missing values, and inconsistencies are common issues. Improving data quality requires standardization of medical data to reduce issues such as data inconsistency, incompleteness, and redundancy.

1.2.5 Analytical techniques

Advanced data analysis techniques such as machine learning and artificial intelligence are required to process complex medical record data and extract valuable information. The technical features of multimodal data fusion and cross domain transfer learning enable the processing of various modal data, including text, images, speech, etc.

1.2.6 Interpretability and interpretability

Complex models such as deep learning may provide high-precision predictions, but the decision-making process of the model is often a black box that is difficult to explain and verify. This is particularly important for medical decision-making, as doctors and patients need to understand the predictive basis of the model.

1.2.7 Legal and ethical issues

Medical record data analysis needs to comply with relevant laws and regulations, such as HIPAA (Health Insurance Portability and Accountability Act), and consider ethical issues. With the development of Internet medical business, a large amount of medical information has been accumulated. The collection and use of medical information need to respect the privacy and informed consent of patients.

1.2.8 Lack of professionals

The lack of interdisciplinary professionals who understand both medical knowledge and data analysis limits the depth and breadth of medical record data analysis. This interdisciplinary professional talent is crucial for promoting the sharing and application of medical big data.

1.3 Challenges Faced by Medical Record Data Analysis

This study aims to explore the potential and necessity of using NLP natural language recognition and big data technology to process medical record data. The following are several key motivations driving this study.

1.3.1 Improving diagnostic accuracy and efficiency

In the medical field, doctors' interpretation of medical records is crucial, but even experienced experts find it difficult to avoid misdiagnosis. NLP natural language recognition algorithms have been proven to exceed the diagnostic accuracy of human doctors in certain tasks. In addition, NLP natural language recognition models can assist doctors in reducing cognitive biases and improving diagnostic efficiency. By automating the extraction of clinical information from medical records, NLP technology can significantly reduce the reading burden on doctors, enabling them to make diagnostic decisions faster and thus improve the overall efficiency of medical services..

1.3.2 Improving healthcare services

The development of deep learning assisted tools, such as CheXNet, not only improves diagnostic accuracy, but also helps reduce missed diagnoses and accelerate radiologists' workflow by generating heat maps, thereby improving healthcare services (Irvin et al., 2019). These tools provide valuable auxiliary information to doctors by automatically identifying and labeling key medical imaging features, making medical services more efficient and accurate.

1.3.3 Expanding the coverage of medical services

In resource scarce areas, deep learning models such as CheXNet can provide auxiliary diagnosis to help patients who have difficulty accessing professional radiologists and expand the coverage of medical services (Rajpurkar et al., 2017). The application of this technology can greatly alleviate the problem of

uneven distribution of medical resources, enabling patients in remote areas to enjoy high-quality medical services.

1.3.4 Promoting medical research and personalized treatment

Deep learning methods have been used to establish disease risk prediction models that predict the probability of disease occurrence in patients based on past time-series data from electronic health records, thereby facilitating the development of personalized treatment plans (Cheng et al., 2016). By analyzing patients' medical records, researchers can better understand the development trends and influencing factors of the disease, and provide more personalized treatment plans for patients.

1.3.5 Coping with the challenges of big data

With the popularization of electronic health records, the healthcare industry has generated a large amount of data. The application of big data technology, such as creating and using electronic medical records, predictive analytics, medical imaging, etc., can help reduce healthcare costs, predict epidemic outbreaks, and avoid preventable diseases (Topol, 2019). These technologies enable medical institutions to more effectively manage and analyze large amounts of medical data, thereby improving the quality and efficiency of medical services.

1.3.6 Data augmentation and transfer learning

In situations where the amount of data is insufficient, data augmentation and transfer learning techniques can be used to train CNN models with less data, reducing reliance on large amounts of data. This is particularly important for disease research where data is scarce (Yosinski et al., 2014). The application of these technologies can improve the generalization ability of models, enabling

effective models to be trained even with limited data, providing new possibilities for medical research.

Through the above motivation, we can see that the research on using NLP and big data technology to process medical record data has important practical significance and application prospects. This study aims to further explore the application of these technologies in medical record data analysis and evaluate their potential in improving diagnostic accuracy, optimizing patient care, and promoting medical research. Through the application of these technologies, we are expected to achieve personalized, precise, and efficient medical services, providing better medical services for patients.

1.4 Related work

Significant progress has been made in the application of big data and NLP natural language recognition in the field of medical record data analysis. Here are some key research developments and existing issues.

1.4.1 Integration of Electronic Medical Records and Medical Imaging Data

The fusion of electronic medical records and medical imaging data is one of the most concerned research directions in medical multimodal fusion. Research has shown that combining pathological tissue sections, ultrasound, endoscopy, chest X-ray and other imaging data with clinical data such as basic patient information and biochemical examination results can improve the accuracy of diagnosis. For example, Qiu et al. developed a deep learning fusion model that accurately distinguishes dementia populations with different levels of cognitive function by combining MRI imaging data and electronic medical record data. This hybrid fusion method not only improves the accuracy of diagnosis, but also provides richer information for clinical decision-making.

1.4.2 Opportunities and challenges of big data analysis in healthcare

The application of big data analysis in the field of healthcare is increasing, including precision medicine, patient health management, and more. However, this field also faces challenges such as data privacy and security, data standardization, and interoperability. Research has pointed out that although big data analysis has enormous potential in the medical field, ensuring data security and privacy, as well as achieving effective sharing of data between different medical institutions, are still urgent issues that need to be addressed.

1.4.3 Application of Convolutional Neural Networks in Medical Image Analysis

The application research of Convolutional Neural Networks (CNN) in medical image analysis has made significant progress. Research has shown that CNN can effectively extract features from medical images for disease diagnosis and classification. For example, Shen et al. achieved good results in classifying lung nodules using multi-scale CNN. In addition, transfer learning techniques have also been used for medical image analysis, by applying CNN models pre trained on large-scale datasets to medical images to solve the problem of insufficient medical image data.

1.4.4 Development trend of intelligent technology application in the medical field

The application of intelligent technologies, including the Internet of Things, cloud computing, artificial intelligence, 5G, etc., in the medical field is showing a steady growth trend. These technologies have shown great potential in drug development, clinical applications, chronic disease management, and other areas. However, with the development of technology, issues such as security and privacy protection, technology regulation, human resource development, and technology

energy consumption are becoming increasingly prominent, requiring more attention in future research.

1.4.5 Current Status Analysis and Prospects of Medical Data Visualization Research

Medical data visualization is an important means to improve the efficiency of medical data resource utilization. Research in this field has shown a growing trend both domestically and internationally, but compared to foreign research, domestic research has problems such as a late start, insufficient innovation vitality, large and imprecise research topic granularity, limited data sources and application scope. Future research needs to improve in these areas to promote the effective utilization of medical data resources.

In summary, although some progress has been made in the field of medical record data analysis, there are still many challenges, including data privacy and security, data standardization and interoperability, technical regulation, and other issues. These challenges require joint efforts from academia, industry, and policymakers to drive the development and application of technology.

2 RESEARCH METHOD

2.1 The preparation of the basic database

In the research of natural language processing (NLP) and big data technology applied to feature extraction of medical record data, the preparation of the basic database is the cornerstone of the entire research process. The following are the detailed steps and methods for preparing the HPO standard medical record phenotype word database and OMIM disease database.

2.1.1 Preparation of HPO standard medical record phenotype word database

Human Phenotype Ontology (HPO) is a widely used standardized vocabulary resource used to describe human phenotypic abnormalities. It plays an important role in clinical genetics research and diagnosis. The following are the detailed steps for preparing the HPO database:

- Step 1: Gain a deeper understanding of the structure and terminology of HPO. The structure of the HPO database is hierarchical, with each phenotypic feature having a unique ID and description.Researchers need to have a deep understanding of the hierarchical structure and terminology of HPO in order to accurately map phenotype descriptions in medical record data. This includes understanding how to access and query terminology through the HPO browser or API.

- Step 2: Use the HPO official website and tools for data retrieval. The HPO official website provides a user-friendly interface that allows researchersto browse and search for phenotype terms.In addition, HPO provides APIs and other tools to facilitate researchers in directly extracting and matching phenotype terms from medical record data. These tools can help researchers quickly locate relevant phenotypic features and match them with descriptions in medical record data.

- Step 3: Combining genomic data for phenotype gene association analysis. The
  combination of HPO and genomic data can help researchers understand the relationship between specific gene variations and phenotypes, which is particularly important for the study of genetic diseases. By associating HPO phenotypes with genetic variations, researchers can better understand the genetic basis of diseases and provide support for clinical diagnosis.

- Step 4: Using third-party tools and databases to assist in phenotype recognition such as Phenomizer and Exomiser can improve the accuracy of

gene variation screening by utilizing HPO phenotype information, which is of great significance for the study of rare diseases. These tools help identify potential disease genes by comparing patient phenotypes with known disease phenotypes.

- Step 5: Regular updates and community participation to maintain the timeliness of the HPO database. Researchers need to regularly check the update logs to ensure that the terminology and information used are up-to-date. Meanwhile, HPO encourages community participation, and researchers can contribute their findings through community feedback mechanisms to continuously enrich and improve the HPO database.
- Step 6: Learn and refer to the documentation and tutorials provided by HPO. HPO provides detailed documentation and tutorials to help researchers gain a deeper understanding of its functionality and applications.These resources are particularly important for novice researchers as they can help them quickly get started and effectively utilize the HPO database.

2.1.2 Preparation of OMIM disease database

The OMIM (Online Mendelian Inheritance in Man) database is an important resource for genetic disease information, providing association information between disease phenotypes and their pathogenic genes. The following are the detailed steps for preparing the OMIM database:

- Step 1: Fully utilize the search and browsing functions of OMIM. One of the core features of OMIM database is its powerful search and browsing capabilities.Researchers can quickly find relevant entries by searching for disease names, gene names, or OMIM numbers. This requires researchers to be familiar with the search syntax and browsing mechanism of OMIM in order to efficiently retrieve information.

- Step 2: Gain a deep understanding of the content of gene and disease entries. Each entry in OMIM contains detailed information, including gene location, function, related diseases, phenotypic characteristics, etc.Researchers need to carefully read these articles to fully understand their content. This includes understanding the genetic patterns, disease descriptions, gene descriptions, and relevant literature in the entries.
- Step 3: Use screening and filtering tools to improve retrieval efficiency OMIM provides a variety of screening and filtering tools to help researchers quickly find the information they need.These tools can be used for screening based on disease types, genetic patterns, phenotypic characteristics, and other factors. Researchers need to be familiar with the usage of these tools in order to quickly locate information of interest in massive amounts of data.
- Step 4: Download and export data for subsequent analysis

  OMIM allows users to download and export data for ease of analysis and research.Researchers can use this data for more in-depth genetic and phenotypic studies. This requires researchers to understand how to correctly download and export data, as well as how to manage and analyze this data in a local environment.

Through the above steps, HPO and OMIM, two important medical databases, can be effectively prepared and used, providing a solid foundation for feature extraction and analysis of medical record data. The preparation of these databases not only involves technical challenges, but also requires interdisciplinary collaboration, including the joint efforts of medical experts, data scientists, and software engineers. Through these databases, we can more accurately extract features from medical record data, providing support for clinical decision support and medical research

2.2 The preparation of the research datas

In this study, we employed a series of data preprocessing steps to ensure the quality and consistency of the data, laying a solid foundation for subsequent feature extraction and model training

2.2.1 Data collection

Data collection is the starting point of the entire research process, involving extracting medical record data from electronic health record (EHR) systems in healthcare institutions. These data typically include patients' personal information, medical history, diagnostic results, treatment plans, and follow-up records. Due to the unstructured nature of medical record data, specific technical means are required for collection and organization.

Template based input: By selecting specific medical record templates, such as outpatient initial visit template, outpatient follow-up visit template, admission record template, and discharge summary template, the efficiency and quality of data collection can be improved. These templates contain structured modules, but the internal content of the modules is still in text format and requires further processing

2.2.2 Data cleaning

Data cleaning is the first step in preprocessing, aimed at removing noise and inconsistencies from medical record data to improve data quality. The following are the detailed steps for data cleaning:

1. Remove irrelevant

characters from medical record text that may contain non-medical related characters, such as headers, footers, doctor signatures, annotation marks, etc. These characters have no value for data analysis and may even interfere with the analysis process. Therefore, it is necessary to develop algorithms or use existing text processing tools to recognize and remove these irrelevant characters.

2. Correcting

spelling errors in medical records may affect the accuracy and consistency of data. By using medical terminology databases and natural language processing tools such as spell checkers and context sensitive error correction algorithms, these errors can be corrected. These tools are typically based on large medical vocabulary databases and are capable of identifying common medical terms and phrases.

3. unify

Different doctors may use different terminology to describe the same medical conditions or symptoms. In order to reduce the ambiguity of data, it is necessary to unify these synonyms or synonyms into standard medical terms. This typically involves mapping to an authoritative medical terminology set, such as UMLS (Unified Medical Language System) or SNOMED CT (Systematically Named Medical Terminology).

4. Missing values

Missing values may exist in medical record data, which may be caused by incomplete records or data input errors. For missing key information, various interpolation methods can be used, such as mean, median, or mode padding. For categorical data, mode or model-based prediction can be used to fill in missing values. For continuous data, mean, median, or regression model-based predictions can be used to fill in missing values.

5. Outlier detection and processing

Outlier detection and processing of medical record data may contain outliers, which may be caused by measurement errors, data input errors, or genuine medical anomalies. Algorithms need to be developed to identify and handle these outliers in order to avoid their adverse effects on the analysis results.

6. Standardized

Standardized medical records may contain text in various formats, such as uppercase letters, abbreviations, plural forms, etc. These texts need to be standardized for subsequent text analysis and processing.

7. Entity recognition and disambiguation

Entity recognition and disambiguation identify entities (such as diseases, drugs, laboratory test results) in medical record text and solve entity disambiguation problems to ensure the correctness and consistency of each entity.

2.2.3 Data standardization

Data standardization is a key step in ensuring data consistency, which involves the following aspects:

1. Timestamps in medical record data may appear in different formats, such as date, time, or timestamp. To ensure the consistency of time data, it is necessary to convert all timestamps into a unified format, such as the ISO 8601 standard. This is helpful for subsequent time series analysis and event correlation analysis.

2. The measurement units in medical records may be inconsistent, such as blood pressure being recorded in millimeters of mercury (mmHg) or kilopascals (kPa), and blood glucose being recorded in milligrams per deciliter (mg/dL) or millimoles per liter (mmol/L). All measurement units need to be converted to the International System of Units (SI) or other standardized units for comparison and analysis.

3. The text in medical record data may be stored in different encoding formats, such as ASCII, ISO-8859-1, or UTF-8. All text data needs to be converted to a unified encoding format, such as UTF-8, to avoid encoding errors and character set incompatibility issues.

4. Unified data format: Medical record data may be stored in different data formats, such as CSV, XML, or JSON. All data needs to be converted into a unified data format for storage, processing, and analysis.

5. Different terms and code sets may be used in mapping medical record data to describe the same medical conditions or procedures. These terms and code sets need to be mapped to an authoritative standard, such as ICD-10 (International Classification of Diseases) or CPT (Current Procedural Terminology).

6. Unstructured text in structured medical record data needs to be converted into structured data for storage and analysis. This may involve using natural language processing techniques to extract key information and store it in relational databases or data warehouses.

Through the above steps of data cleaning and standardization, the quality and consistency of medical record data can be ensured, providing a solid foundation for subsequent data analysis and feature extraction. These steps require interdisciplinary collaboration, including the joint efforts of medical experts, data scientists, and software engineers, to ensure the accuracy and effectiveness of data preprocessing.

2.3 Methodology

2.3.1 Use SciSpacey for Named Entity Recognition

Scispacy is an NLP library designed specifically for scientific literature and clinical documents, which extends the functionality of spaCy to include recognition of biomedical entities. In this study, we used Scispacy to perform NER on medical record text to identify medical terms.

- Data preprocessing: Firstly, preprocess the medical record text, including word segmentation, removal of stop words, and non-medical related characters.
- Entity recognition: Using Scispacy's NER model to identify medical terms in medical records, such as diseases, drugs, treatments, and other entities. Scispace is set to The kb_dets attribute on spaCy Spans contains a list of knowledge base concept IDs and related scores to identify entities.

2.3.2 Compare with HPO and OMIM libraries using fuzzywuzzy software

Fuzzywuzzy is a Python library used to calculate the degree of difference between two strings and find the best match. In this study, we used fuzzywuzzy software to compare the medical terms identified by Scispacy with the terms in the HPO and OMIM databases to improve the accuracy of recognition.

- Building a comparison dictionary: Construct a dictionary of medical terms from the HPO and OMIM databases for comparison with the terms identified by Scispacy.
- Calculate similarity: Use fuzzywuzzy to calculate the similarity between medical terms in medical records and terms in dictionaries, and select the best matching item. Fuzzybuzz helps us find the closest match by calculating the degree of difference between two sequences, thereby improving the accuracy of matching.

2.3.3 Build a negative word library

To identify negative words in medical records, we constructed a negative word library and matched it using fuzzywuzzy software.

- Lexicon construction: Collect common medical negative words such as "none", "not", "deny", etc., and construct a negative word lexicon.
- Negative word recognition: Using fuzzywuzzy software to identify negative words in medical record text, in order to mark negative statements in medical records. The string matching ability of fuzzywuzzy enables us to accurately identify negative words in medical records, which is crucial for understanding the semantics of medical record text.

4. Mark medical terms and negative words in medical records

Finally, we will mark the medical terms identified by Scispacy and the negative words identified by fuzzywuzzy in the medical record text for subsequent analysis.

- Text tagging: Marking medical terms and negative words in medical record text to provide structured data for further clinical decision support and

research. In this way, we can convert unstructured medical record data into structured data for storage, retrieval, and analysis.

2.4 Result

Through the above architecture model, we are able to accurately extract and label medical terms and negative words from medical record data, providing valuable information for clinical decision support systems and medical research.

3 EXPERIMENT

3.1 NLP Experimental steps

1. Install and import necessary packages

Firstly, install scispace, spaCy, and specific NLP models within scispace. These models can be installed through their URLs

pip install scispacy

pip install spacy

pip install fuzzywuzzy

2. Import package

Import the installed scispacy and spaCy packages in Python.

```
import spacy
import scispacy
from fuzzywuzzy import fuzz,process
import re
from collections import Counter
```

3. Enter the revised original medical record

chief_complaint = """

Chief complaint: Difficulty in communication and developmental delay of 5 years. At the age of 2, the child will call Mom and Dad and seek medical attention at a childrens hospital due to developmental delay. They can now speak simple sentences, say I want to play with toys, I want to use the bathroom, execute 3-step instructions, have limited eye contact and joint attention, have facial features, do not understand size, can understand most adults words, can partially understand adults expressions, cannot comfort others, have a bad temper, are hyperactive, have many conflicts with family members, watch cartoons repeatedly, understand danger, cannot write, can write numbers, cannot count, do not know time, do not provide feedback on school situation, have difficulty learning, cannot concentrate in class, and have many small movements. Suspicion: Generalized Developmental Disorder, Attention Deficit Hyperactivity Disorder, Childhood Autism?.

"""

4.Using SciSpacey for Named Entity Recognition

```
#Load the scispacy model
nlp = spacy.load(en_core_sci_sm)
#Perform NER
doc = nlp(text)
#Load HPO and OMIM databases
def read_second_column(file_path):
    second_column_data = []
    with open(file_path, r, encoding=utf-8) as file:
        for line in file:
            second_column_data.append(line.lower())
return second_column_data

file1_path = hpo.txt
    file2_path = omim.txt
```

```
#Read the second column content of two files
   file1_data = read_second_column(file1_path)
   file2_data = read_second_column(file2_path)
Medical_terms=file1_data+file2data # Merge data from two files
```

5. Compare with HPO and OMIM libraries using fuzzywuzzy software to obtain medical terminology

```
keywords_filter = []
 for ent in doc.ents:
       result1 = fuzzy_match_medical_terms(ent.text.lower(), medical_terms)
       if(result1 != ""):
         keywords_filter.append(ent.text)
```

6. Building a negative lexicon

```
#List of negative words
   negative_words = [
   "not", "no", "never", "none", "nobody", "nothing", "nowhere",
   "neither", "nor", "no longer", "no more", "dont", "isnt",
   "arent", "wasnt", "werent", "cant", "couldnt", "shouldnt",
   "wouldnt", "might not", "must not", "without", "hardly",
   "scarcely", "barely", "refuse", "deny", "disallow", "prohibit",
   "forbid", "against", "un-", "non-", "de-", "anti-", "il-", "im-",
   "in-", "ir-", "-less", "-free"
   ]
```

7. Mark medical terminology words, mark negative words, and obtain medical records after marking medical terminology words

```
def highlight_words_terminal(sentence, words_to_highlight):
#ANSI escape code, used to set the text color to red
   red_start = "\033[31m"
   red_end = "\033[0m"
```

```
        #Find all the words that need to be highlighted using regular expressions and wrap them with ANSI escape codes
            for word in words_to_highlight:
                sentence = re.compile(word).sub(f"{red_start}{word}{red_end}", sentence)
            return sentence
        def highlight_negative_words_terminal(sentence, negative_words):
        #ANSI escape code, used to set the text color to green
            yellow_start = "\033[32m"
            yellow_end = "\033[0m"

        #Traverse the list of negative words and wrap each negative word with a green label
            for word in negative_words:
        #Use regular expressions to ensure that the entire word is matched, rather than partial matches
                pattern = r\b + re.escape(word) + r\b
                sentence = re.sub(pattern, f{yellow_start}{word}{yellow_end}, sentence, flags=re.IGNORECASE)
        return sentence

          highlighted_sentence_terminal = highlight_words_terminal(text, keywords_filter)
            final_sentence_terminal = highlight_negative_words_terminal(highlighted_sentence_terminal, negative_words)
         print(final_sentence_terminal)

        def ansi_to_html(ansi_text):
```

```
#Define the mapping from ANSI colors to HTML colors
    ansi_to_html_colors = {
        30: black, 31: red, 32: green, 33: yellow,
        34: blue, 35: magenta, 36: cyan, 37: white,
        40: black, 41: red, 42: green, 43: yellow,
        44: blue, 45: magenta, 46: cyan, 47: white
    }
#Replace ANSI escape sequences with HTML
    def replace_ansi(match):
        code = match.group(1)
        color = ansi_to_html_colors.get(code, black)
        return f<span style="color: {color};">
#Match ANSI color codes and replace them
    html_text = re.sub(r\033\[(\d{1,2})m, replace_ansi, ansi_text)
Html_text+=</span># Close the last span tag
    return html_text
#Convert to HTML
    html_output = ansi_to_html(final_sentence_terminal)

#Create HTML file and write content
    with open(output.html, w, encoding=utf-8) as f:
        f.write(<html><body>)
        f.write(html_output)
        f.write(</body></html>)

Print ("HTML file has been generated.")
```

8. Output medical records marked with medical terminology words

Chief complaint: Difficulty in communication and developmental delay of 5 years. At the age of 2, the child will call Mom and Dad and seek medical attention at

a childrens hospital due to developmental delay. They can now speak simple sentences, say I want to play with toys, I want to use the bathroom, execute 3-step instructions, have limited eye contact and joint attention, have facial features, do not understand size, can understand most adults words, can partially understand adults expressions, cannot comfort others, have a bad temper, are hyperactive, have many conflicts with family members, watch cartoons repeatedly, understand danger, cannot write, can write numbers, cannot count, do not know time, do not provide feedback on school situation, have difficulty learning, cannot concentrate in class, and have many small movements. Suspicion: Generalized Developmental Disorder, Attention Deficit Hyperactivity Disorder, Childhood Autism?

9. Select 966 clinical medical records with a character count greater than 200, totaling 7044 HPO standard words, and calculate the matching accuracy of the selected keywords in the medical records

1. After program processing, standard words, medical record words, and NLP matching data are obtained

2. Select 20 samples from each interval for review and calculate the matching and unmatched data. If there are less than 20 data samples, extract all data to obtain the accuracy of each NLP matching interval and clinical word selection.

Table 3.1 – Choose Sample to stastic the accuray

nlp match ratio	sample num	hit num	not hit num	accuray
0-20	18	2	16	0.111111
21-30	20	10	10	0.5
31-40	20	18	2	0.9
41-50	20	18	2	0.9
51-60	20	18	2	0.9
61-70	20	19	1	0.95
71-80	20	20	0	1
81-90	20	20	0	1
91-100	20	20	0	1

2. Assign the accuracy values of each NLP match based on the sampling structure, and the following table shows the process of assigning values to some data

Table 3.2 –Calculate the accuracy of nlp match ratio

hpo case	medical case	nlp match ratio	accuracy
epileptic seizures	apoplectic seizure	78	100
catatonic seizures	apoplectic seizure	65	95
hyperinsulinemia	insulin	61	95
hypertrophic cardiomyopathy	hypertrophic cardiomyopathy	98	100
cardiac hypertrophy	cardiac hypertrophy	97	100
mitral regurgitation	mitral regurgitation	98	100
patent foramen ovale	patent foramen ovale not closed	77	100
ventricular hypertrophy	ventricular hypertrophy	98	100
left ventricular outflow tract abnormalities	left ventricular outflow tract obstruction	80	100
hypoglycemia	hypoglycemia	96	100
inspiratory stridor	inspiratory trismus	77	100
neonatal sepsis	neonatal	67	95
stereotypic behavior	hyperactivity	36	90
short attention span	poor concentration	46	90
irritability	irritable	73	100
delayed speech and language development	developmental delay	37	90
hyperactivity	hyperactivity	96	100
mental retardation	poor concentration	54	90
global developmental delay	developmental delay	83	100
delayed motor development	developmental delay	49	90

3. Calculate the average accuracy for 7044 samples, and obtain an average accuracy of 92.90%

3.2 BigData Experimental steps

1. Prepare past medical record data, organize and revise medical records from multiple sources, and form revised medical records and standard medical records

信息 结果 1 剖析 状态

rsid	reviseDiseaseCase	phoFeature
DD17000099	浮肿，蛋白尿，大剂量激素依赖。诊断：肾病综合征 首次发病日期2(	蛋白尿,肾病综合征,水肿
DD17000100	听力损失	听力受损
DD17000101	患者双结膜充血，双眼无虹膜缺损，双晶体棕黑色，双眼白内障，双眼	结膜充血,视力障碍,白内障,无虹膜,眼球震颤,重度
DD17000103	患者，左眼视网膜脱离，左眼偏小，双眼球震颤，双眼牵牛花综合征	小眼,视网膜脱离,眼球震颤,脑基底异常血管网现
DD17000104	患者6个月首发病性发作，首次发作后发烧，近些天发作频繁。	癫痫发作,热性惊厥发作,发热
DD17000105	患儿2015年11月份1次抽搐，表现为意识丧失，呼之不应，双眼上翻,	癫痫发作,热性惊厥发作,意识丧失
DD17000106	首次发病日期2017.12。诊断：癫痫。	癫痫发作
DD17000107	癫痫 首次发病日期2010.10	癫痫发作
DD17000108	首次发病日期：2017-12月。患儿无明显诱因发热、伴寒战，左眼结膜	结膜充血,眼睑水肿,中性粒细胞减少,白细胞减少
DD17000110	患者查体 左右运动可，眼震（±）指鼻可走直线欠佳，闭眼（+），步	眼球震颤,反射消失,闭目难立征阳性,感觉障碍,步
DD17000111	首次发病日期：2017-12月。患者运动倒退，现竖颈都不稳，手不能抓	运动迟缓,发育倒退,心脏肥大,左心室肥大,左心室
DD17000112	患者脾大。父亲有脾大切除史。诊断：溶血性贫血。	溶血性贫血,脾肿大
DD17000113	患者夜盲，视力下降。诊断：视网膜色素变性。	视力障碍,视网膜色素变性,夜盲
DD17000114	患者肌酸激酶偏高11748U/L，轻度骨强度不足。诊断：DMD。	骨密度减少,血清肌酸磷酸激酶增高,肌营养不良
DD17000115	因"腹痛1次、外院B超发现双肾结石10余天"于2017-12-21入复旦大学	肾积水,肾结石,膀胱肥厚,肾盂扩张,视网膜色素沉

Figure 3.1 – Original data of chinese medical record

2. Translate and revise medical record data into English (due to confidentiality and ethical issues, we have selected several of them as examples of experimental steps)

Revise medical records:

D100001 One month ago, I discovered that I had lost consciousness, drooled and bitten my tongue, and had convulsions in my limbs after waking up at 7am. The symptoms lasted for one minute and improved. I also had occasional episodes of confusion, and my EEG showed 3HZ, slow waves, and no family history. Diagnosis: Epilepsy (Absence Seizure)

D100002 patient G2P2, full-term cesarean section, bw6Kg. The mother's prenatal monitoring showed elevated blood sugar levels, normal fetal heart rate and fetal movement, smooth delivery process, premature rupture of membranes for 8 hours, no abnormalities in placenta, and no asphyxia after birth. The Apgar scores at 1 minute, 5 minutes, and 10 minutes were all 9 points (skin color -1 point). The child gradually developed breathing difficulties after birth, manifested as moaning, blue purple skin color, and inspiratory three concave signs, without convulsions or fever. On December 18, 2017, the chest X-ray showed slight thickening of the lung markings on both sides, and the heart shadow showed

enlargement to both sides. The monitored blood sugar was less than 1.1 mmol/L. After admission, assisted respiratory therapy with a ventilator was given, and a complete cardiac ultrasound showed the presence of hypertrophic cardiomyopathy. The family history of this disease was denied, and treatment with propranolol was given. The heart function is gradually improving, but there is still myocardial hypertrophy. At admission, there was hypoglycemia, and insulin was found to be elevated. Glucose supplementation was given, and blood sugar is now normal. Currently, formula milk is fed at 70ml/3h, and blood sugar is normal. Postnatal monitoring showed an increase in infection indicators, indicating the presence of sepsis. Intravenous administration of meropenem, amoxicillin, and clavulanate potassium combined with anti infective treatment was given, and the monitored infection indicators gradually improved. Examination of echocardiography on December 20, 2017: ventricular hypertrophy, left ventricular outflow tract obstruction, patent foramen ovale, mitral regurgitation (mild), tricuspid regurgitation (mild), pulmonary hypertension (mild); Echocardiography on December 25, 2017: Ventricular hypertrophy, left ventricular outflow tract obstruction, patent foramen ovale, mild mitral regurgitation, mild tricuspid regurgitation, and mild pulmonary hypertension. Diagnosis: diabetes mother newborn, hypertrophic cardiomyopathy, neonatal hypoglycemia, neonatal sepsis.

D100003 has convulsions when experiencing high fever, overall developmental delay, intellectual disability, inability to speak, frequent "ah, ah" sounds, irritable personality, fear of loud noises, poor attention, hyperactivity, hyperactivity, stereotyped movements, poor dexterity of hands, easy dropping of fingers when holding things, inversion of the right foot, frequent tiptoeing, and can walk at the age of 2. Diagnosed with "pediatric cerebral palsy" on June 19, 2013.

D100004 Swelling, proteinuria, high-dose hormone dependence. Diagnosis: nephrotic syndrome first onset date November 2017

D100005 hearing loss

D100006 patient has conjunctival congestion, no iris defect in both eyes, brownish black bicrystals, cataracts in both eyes, nystagmus in both eyes,

visual acuity of 0.08 in the right eye and 0.01 in the left eye. Diagnosis: Congenital absence of iris.

D100007 patient, left eye retinal detachment, left eye small, double eye ball tremor, double eye morning glory syndrome. Diagnosis: Moyamoya disease?.Patient DD17000104 had their first seizure at 6 months, followed by fever and frequent seizures in recent days.In November 2015, the DD17000105 child had a convulsion, which was characterized by loss of consciousness, inability to respond to the call, eyes turning upward, closed teeth, limbs rigidity, foam vomit at the mouth, blisters, incontinence in urine (once), and lasted for one minute. On December 20, 2017, at 1:49, the seizure occurred once, the performance was the same as before, body temperature was 37 ℃, and the total number of seizures was 8. My father and grandmother had a history of convulsions during their childhood. Diagnosis: Heat related epilepsy.

Standard medical record:

D100001 epileptic seizures, absent-minded seizures

D100002 hyperinsulinemia, hypertrophic cardiomyopathy, cardiac hypertrophy, mitral regurgitation, patent foramen ovale, ventricular hypertrophy, abnormal left ventricular outflow tract, hypoglycemia, inspiratory wheezing, neonatal sepsis

D100003 Stereotypical behavior, short attention span, irritability, delayed speech and language development, hyperactivity, intellectual disability, generalized developmental delay, delayed motor development, speech loss, febrile seizures, toe walking, cerebral palsy, clubfoot

D100004 proteinuria, nephrotic syndrome, edema

D100005 hearing loss

D100006 Conjunctival congestion, visual impairment, cataracts, absence of iris, nystagmus, severe myopia

D100007 microphthalmia, retinal detachment, nystagmus, abnormal vascular network phenomenon at the base of the brain

Compare the corrected medical records with standard medical records, obtain a frequency correspondence table between natural language medical record words and standard medical record words through big data analysis, and complete the standardization of clinical medical records through deep learning and learning feedback, and further remove noisy words

3. Comparison between medical record data and standard data

```
import spacy
import scispacy
from fuzzywuzzy import fuzz,process
import re
from collections import Counter

# Load the English medical model of scispacy
# Make sure to have scispacy and the corresponding model installed. Use the following commands to install:
# pip install scispacy
# pip install https://s3-us-west-2.amazonaws.com/ai2-s2-scispacy/releases/v0.5.4/en_core_sci_md-0.5.4.tar.gz

nlp = spacy.load("en_core_sci_md")

# Initialize an empty dictionary to store the relationship between keywords and HPOs and their counts
keyword_matches = {}

# Use fuzzywuzzy for fuzzy matching of medical terms
def fuzzy_match_medical_terms(text, medical_terms, score_threshold=70):
    keywords = ""
    ratio = 0
    for line in medical_terms:
```

```
            columns = line.strip().split('\t')
            term = ""
            if len(columns) > 1:
                term = (columns[1].lower())  # Convert to lowercase and add to the
list
            matches = fuzz.ratio(text, term)
            if matches >= score_threshold and matches > ratio:  # If the highest
similarity is greater than or equal to the threshold
                ratio = matches  # Add the matched term and similarity score
                keywords = line
        re = keywords
        return re

    def read_line(file_path):
        line_data = []
        with open(file_path, 'r', encoding='utf-8') as file:
            for line in file:
                line_data.append(line.lower())
        return line_data

    def get_keywords(text, hpo_text, medical_terms):
        # Use scispacy for named entity recognition
        doc = nlp(text)
        keywords_filter = []
        # Filter out medical terms
        for ent in doc.ents:
            result1 = fuzzy_match_medical_terms(ent.text.lower(), medical_terms)
            if result1 != "":
                keywords_filter.append(ent.text)
        hpos = hpo_text.strip().split(',')
```

```
            for keyword in keywords_filter:
                hpo_choose = ""
                ratio = 0
                for hpo in hpos:
                    match_ratio = fuzz.ratio(keyword, hpo)
                    if match_ratio > ratio:
                        ratio = match_ratio
                        hpo_choose = hpo
                if (keyword, hpo_choose) in keyword_matches:
                    keyword_matches[(keyword, hpo_choose)] += 1
                else:
                    keyword_matches[(keyword, hpo_choose)] = 1


        def get_col2(line):
            columns = line.strip().split('\t')
            if len(columns) > 1:
                return columns[1].lower()
            return ""


        def main():
            file1_path = 'hpo.txt'  # Path to the HPO file
            file2_path = 'omim.txt'  # Path to the OMIM file
            # Read the second column content of both files
            file1_data = read_line(file1_path)
            file2_data = read_line(file2_path)
            medical_terms = file1_data + file2_data  # Merge the data from both files
            case_path = "case_data_t1.txt" # Corrected medical records
            case_hpo_path = "case_data_hpo_t2.txt" # HPO terms of the medical
records
            case_data = read_line(case_path)
```

```
        case_hpo_data = read_line(case_hpo_path)
        #print(case_data)

        for i in range(len(case_data)):
            case_line = case_data[i]
            case_hpo_line = case_hpo_data[i]
            # print(get_col2(case_line))
            get_keywords(get_col2(case_line), get_col2(case_hpo_line), medical_terms)

        # Sort the dictionary by count
        sorted_keyword_matches = sorted(keyword_matches.items(), key=lambda item: item[1], reverse=True)

       # Write the sorted results to a text file
        with open('keyword_matches.txt', 'w') as file:
            for (keyword, corresponding_word), count in sorted_keyword_matches:
                file.write(f"{keyword}, {corresponding_word},  {count}\n")

        # Print to the console for checking
        for (keyword, corresponding_word), count in sorted_keyword_matches:
            print(f"Keyword: {keyword}, Corresponding Word: {corresponding_word}, Count: {count}")

    if __name__ == '__main__':

        main()
```

4.Output the medical record terms and standard terms frequency library (this result contains only demonstration content).

mild, mitral regurgitation, 4

diagnosis, nystagmus, 3

monitoring, inspiratory wheezing, 2

sides, hyperinsulinemia, 2

presence, inspiratory wheezing, 2

hypertrophic cardiomyopathy, hypertrophic cardiomyopathy, 2

sepsis, neonatal sepsis, 2

ventricular hypertrophy, ventricular hypertrophy, 2

left ventricular outflow tract obstruction, abnormal left ventricular outflow tract, 2

patent foramen ovale, patent foramen ovale, 2

tricuspid regurgitation, mitral regurgitation, 2

frequent, delayed motor development, 2

hyperactivity, hyperactivity, 2

seizure, nystagmus, 2

seizures, nystagmus, 2

consciousness, epileptic seizures, 1

convulsions, epileptic seizures, 1

improved, epileptic seizures, 1

episodes, epileptic seizures, 1

confusion, epileptic seizures, 1

slow waves, absent-minded seizures, 1

family history, absent-minded seizures, 1

diagnosis, epileptic seizures, 1

absence seizure, absent-minded seizures, 1

full-term cesarean section, mitral regurgitation, 1

prenatal, neonatal sepsis, 1

elevated blood sugar levels, patent foramen ovale, 1

normal fetal heart rate, abnormal left ventricular outflow tract, 1

fetal movement, neonatal sepsis, 1

premature rupture of membranes, abnormal left ventricular outflow tract, 1

no abnormalities, neonatal sepsis, 1

placenta, hypoglycemia, 1

no asphyxia, neonatal sepsis, 1

apgar scores, neonatal sepsis, 1

skin color, hypertrophic cardiomyopathy, 1

child, hyperinsulinemia, 1

breathing difficulties, hyperinsulinemia, 1

inspiratory, inspiratory wheezing, 1

concave signs, neonatal sepsis, 1

convulsions, neonatal sepsis, 1

fever, patent foramen ovale, 1

thickening, inspiratory wheezing, 1

ventilator, ventricular hypertrophy, 1

cardiac ultrasound, cardiac hypertrophy, 1

family history, cardiac hypertrophy, 1

disease, hyperinsulinemia, 1

heart function, mitral regurgitation, 1

myocardial hypertrophy, cardiac hypertrophy, 1

hypoglycemia, hypoglycemia, 1

insulin, hyperinsulinemia, 1

increase, neonatal sepsis, 1

anti, neonatal sepsis, 1

infection, mitral regurgitation, 1

improved, inspiratory wheezing, 1

examination, mitral regurgitation, 1

mitral regurgitation, mitral regurgitation, 1

pulmonary hypertension, ventricular hypertrophy, 1

mild mitral regurgitation, mitral regurgitation, 1

mild pulmonary hypertension, ventricular hypertrophy, 1

diagnosis, neonatal sepsis, 1

neonatal hypoglycemia, hypoglycemia, 1

neonatal, neonatal sepsis, 1

convulsions, short attention span, 1

high fever, febrile seizures, 1

developmental delay, generalized developmental delay, 1

intellectual disability, intellectual disability, 1

poor attention, short attention span, 1

fingers, febrile seizures, 1

inversion, febrile seizures, 1

right foot, clubfoot, 1

pediatric, hyperactivity, 1

swelling, proteinuria, 1

proteinuria, proteinuria, 1

dependence, nephrotic syndrome, 1

diagnosis, edema, 1

nephrotic syndrome, nephrotic syndrome, 1

patient, visual impairment, 1

conjunctival congestion, conjunctival congestion, 1

iris defect, absence of iris, 1

cataracts, cataracts, 1

nystagmus, nystagmus, 1

visual acuity, visual impairment, 1

right eye, cataracts, 1

congenital absence, absence of iris, 1

iris, absence of iris, 1

patient, retinal detachment, 1

left eye retinal detachment, retinal detachment, 1

double eye ball tremor, abnormal vascular network phenomenon at the base of the brain, 1

fever, retinal detachment, 1

frequent, retinal detachment, 1

child, microphthalmia, 1

convulsion, microphthalmia, 1

loss of consciousness, retinal detachment, 1

closed teeth, microphthalmia, 1

limbs rigidity, retinal detachment, 1

incontinence, microphthalmia, 1

urine, retinal detachment, 1

history, nystagmus, 1

convulsions, microphthalmia, 1

childhood, microphthalmia, 1

2. By the experimental with history big data, We can mark words to three class. Some words we can use as keywords database, there are a part words in the next table but not all：

Table 3.3 - The medical case frequence distribution

entity	frequence
seizures	132
fever	122
pneumonia	93
developmental delay	83
anemia	81
mental retardation	70
motor development	58
cough	57
developmental backwardness	49
seizure	43
convulsions	41
vomiting	32
bilirubin	27
hypoglycemia	26
muscle tone	26
lower limbs	26
hypoproteinemia	25
diarrhea	25
alkaline phosphatase	25
malnutrition	24

liver function	24
hypertension	23
unsteady walking	23
proteinuria	23
abdominal pain	22
bilateral	22
lateral ventricles	22
slurred speech	21
poor appetite	21
episodes	21
head circumference	20
intellectual development	20
cranial mri	19
loss of consciousness	18
hypotonia	18
eeg abnormalities	18
respiratory failure	17
infantile spasms	17
blood routine	17
hypospadias	17
immunodeficiency	16
tendon reflexes	16
hyperbilirubinemia	16
jaundice	16
myelination	16
corpus callosum	16
leukocytes	16
bronchopneumonia	15
nystagmus	14
cesarean section	14
coughing	14
muscular dystrophy	14
hyponatremia	14
urine	14
hemoglobin	14
growth retardation	14
atrial septal defect	13
erythrocytes	13
global developmental delay	13
severe	13
nephrotic syndrome	13
hypokalemia	13
cyanosis	12
hematuria	12
knee tendon reflexes	12
myopathy	12
hypocalcemia	12
facial features	12

basal ganglia	12
tricuspid regurgitation	11
thrombocytopenia	11
lower extremities	11
cerebral white matter	11
short stature	11
microcephaly	11
headache	11
myotonic dystrophy	11
hypoxia	11
lactate dehydrogenase	11
c-reactive protein	11
parathyroid hormone	11
metabolic acidosis	10
ataxia	10
cryptorchidism	10
developmental regression	10
hyperlactatemia	10
squatting	10
cerebral palsy	10
developmental	10
hyperkalemia	10
cerebral hypoplasia	10
chronic kidney disease	10
neonatal	9
neutrophils	9
low muscle tone	9
viral encephalitis	9
bronchiectasis	9
pulmonary hypertension	9
recurrent infections	9
edema	9
hypoalbuminemia	9
sepsis	9
stiffness	9
sinusitis	9
knee reflexes	9
no abnormality	9
left ventricle	9
psychomotor retardation	9
nausea	9
hepatomegaly	9
hyperactivity	8
cardiac ultrasound	8
rehabilitation	8
upturned eyes	8
gastrocnemius muscle	8
palatal	8

upper respiratory tract infection	8
recurrent cough	8
convulsive seizures	8
dry skin	8
developmentally delayed	8
cerebral dysplasia	8
backward motor development	8
history	8
hemolytic anemia	8
myocardial damage	8
overall development	8
acute liver failure	8
chronic kidney disease stage	8
motor regression	8
petechiae	7
unconsciousness	7
disease	7
salivation	7
generalized developmental delay	7
epileptiform discharges	7
eyelids	7
renal function	7
strabismus	7
tremor	7
poor head control	7
enlarged	7
muscle atrophy	7
convulsion	7
pancreas	7
white matter	7
neutrophil	7
consciousness	7
autism	7
encephalitis	7
high lactate	7
ankle clonus	7
aplastic anemia	7
negative	7
peroneal muscular dystrophy	7
hypoplasia	7
long eyelashes	7
low ear position	7
arachnoid cyst	7
subarachnoid space	7
regression	6
cerebral atrophy	6
generalized seizures	6
tuberous sclerosis	6

developmental delays	6
pericardial effusion	6
pleural effusion	6
scoliosis	6
hypothyroidism	6
respiration	6
diabetes mellitus	6
ventricular septal defect	6
behavioral abnormalities	6
convulsive seizure	6
epileptic encephalopathy	6
epileptic seizures	6
liver	6
hearing	6
peripheral neuropathy	6
achilles tendon reflexes	6
microscopic hematuria	6
postural tremor	6
developmental lag	6
metabolic acid	6
brain development	6
eye contact	6
intermittent	6
mitochondrial encephalomyopathy	6
metabolic encephalopathy	6
history of jaundice	6
splenomegaly	6
spleen	6
low hairline	6
cerebral white matter disease	6
mild anemia	6
hyphae	6
aortic regurgitation	6
coagulation abnormalities	6
mitral regurgitation	5
cataracts	5
intracranial hemorrhage	5
nasal congestion	5
triggers	5
feeding difficulties	5
left lateral ventricle	5
hypertrophy	5
chicken chest	5
weight loss	5
unsteady gait	5
photophobia	5
renal cysts	5
short penis	5

intracranial infection	5
lymph nodes	5
prone position	5
renal insufficiency	5
hydronephrosis	5
coagulation	5
cystic fibrosis	5
limb muscle tone	5
urinary protein	5
metabolic disease	5
patent foramen ovale	5
arthralgia	5
dystonia	5
visual evoked potentials	5
pituitary gland	5
lethargy	5
granulocyte deficiency	5
clenching	5
sides	5
dull expression	5
recurrent eczema	5
lymphocyte	5
immune thrombocytopenic purpura	5
leukoencephalopathy	5
hyperactive	5
congenital cataract	5
cholestasis	5
acute upper respiratory tract infection	5
adrenocortical insufficiency	5
abdomen	5
bilateral ventricles	5
albumin	5
poor reaction	5
neonatal pneumonia	5
hyperuricemia	5
gastroesophageal reflux	5
chlorine	5
personal history	5
low bone density	5
elevated creatinine	5
gross motor development	5
small head circumference	5
motor developmental delay	5
cardiac mucinous tumor	5

In the next table， we picked Some words of noise words

Table 3.3 - The noise words marked

entity	frequence	noise words
diagnosis	83	1
examination	36	1
presentation	31	1
patient	21	1
no abnormalities	19	1
family history	16	1
speech	11	1
growth	10	1
medication	8	1
child	7	1
pain	6	1
healthy	5	1
size	4	1
frequent	4	1
face	4	1
child's	4	1
tests	4	1
active	4	1
patients	3	1
uncle	3	1
increase	3	1

Some words need to expend with behind word

Table 3.3 - The words need expend with behind word marked

entity	frequence	need extend words(mark as 2)
abnormal	108	2
abnormality	72	2
abnormalities	68	2
increased	42	2
decreased	23	2
weakness	22	2
difficulty	19	2

protein	8	2
expression	6	2
decrease	5	2
short	4	2
weight	3	2

Through the analysis of word frequencies, the correspondence between medical record terms and standard terms can be obtained. With the accumulation of medical records and repeated training, effective medical record classification and more accurate standard terms can be achieved. Additionally, for noise words, the word frequency table can be repeatedly trained and refined through positive user feedback to remove noise and obtain more accurate medical record keywords.

When we removed noisy words and performed NLP natural language matching again to obtain clinical keywords, we used the data from the previous experiment. After removing the noise, a total of 6750 pieces of data were analyzed and found that the accuracy increased from 92.90% to 96.14%.

4 RESULT

The results of the experiment have demonstrated a significant advancement in the field of medical data analysis through the accurate extraction and labeling of medical terms and negative words from medical record data. This achievement not only enhances the utility of electronic health records but also paves the way for more sophisticated clinical decision support systems (CDSS) and medical research initiatives. In this discussion, we will explore the implications of these results, the potential applications in various medical domains, and the future directions for research and development

4.1 Implications of Accurate Extraction and Labeling

The ability to accurately extract and label medical terms and negative words is crucial for transforming raw medical record data into actionable insights. This

process allows for the standardization of medical terminology, which is essential for consistency and comparability across different datasets. By providing a structured format for unstructured data, it becomes easier to identify patterns, trends, and anomalies that can inform clinical decisions and research outcomes.

4.2 Enhancing Clinical Decision Support Systems

Clinical decision support systems rely on accurate and timely data to assist healthcare providers in making informed decisions. The results of this experiment indicate that the integrated model can effectively preprocess and analyze medical records, which can be directly fed into CDSS. This integration can lead to several benefits:

- **Improved Diagnostic Accuracy**: By identifying key medical terms and excluding irrelevant information, the model can help in the early and accurate diagnosis of conditions.
- **Personalized Treatment Plans**: The detailed analysis of patient records can facilitate the creation of personalized treatment plans based on individual patient histories and current conditions.
- **Medication Management**: The model can flag potential drug interactions and contraindications, reducing the risk of adverse drug events.
- **Quality Improvement**: By monitoring the effectiveness of treatments and identifying areas for improvement, the model can contribute to the continuous enhancement of healthcare services.

4.3 Advancing Medical Research

The accurate extraction of medical terms and negative words also has profound implications for medical research. Researchers can utilize this capability to:

- **Analyze Large Datasets**: The model can process and analyze large volumes of medical records, identifying trends and correlations that might otherwise be missed.
- **Accelerate Clinical Trials**: By quickly identifying eligible participants and monitoring their progress, the model can help in the efficient execution of clinical trials.
- **Genomic Research**: In the context of genomic medicine, the model can assist in linking genetic variations to specific medical conditions, contributing to the understanding of genetic diseases.
- **Public Health Surveillance**: The model can be used to monitor the spread of diseases and track public health trends, enabling timely interventions during outbreaks.

4.4 Potential Applications in Healthcare

Beyond direct clinical applications, the accurate extraction and labeling of medical terms can also be applied in various other areas of healthcare:

- **Epidemiological Studies**: Understanding the distribution and determinants of diseases within populations.
- **Healthcare Administration**: Streamlining administrative tasks and improving the management of healthcare facilities.
- **Patient Education**: Developing tailored educational materials based on individual patient records.
- **Health Policy Development**: Informing policy decisions with data-driven

insights into healthcare system performance.

4.5 Future Directions

Looking forward, the following directions can be explored to build upon the current results:

- **Integration with Other Data Types**: Combining text data with other forms of medical data, such as imaging and genomics, to create a more comprehensive view of patient health.
- **Real-Time Analysis**: Developing the model's capability to provide real-time analysis of medical records during clinical encounters.
- **Adaptive Learning**: Enhancing the model's ability to learn from new data and adapt its algorithms accordingly, ensuring it remains current with the latest medical knowledge.

Global Health Applications: Expanding the model's application to global health initiatives, particularly in low-resource settings where advanced healthcare technologies can have a significant impact

In conclusion, the successful extraction and labeling of medical terms and negative words from medical record data is a significant milestone in the journey towards more intelligent healthcare systems. It has the potential to revolutionize clinical decision-making, enhance medical research, and improve patient outcomes. As we continue to refine and expand this technology, it is essential to do so with a keen awareness of the ethical, legal, and social implications, ensuring that the benefits of these advancements are realized responsibly and equitably.

5 FUTURE WORK

As artificial intelligence and machine learning technologies continue to penetrate the medical field, future research can be conducted in the following directions to promote progress in medical diagnosis and treatment, and ultimately improve patient treatment outcomes and quality of life.

5.1 Model Optimization

Explore more efficient network architectures and training strategies to improve model performance and computational efficiency.

In the field of deep learning, model optimization is an eternal topic. Future research can focus on designing more efficient neural network architectures, such as lightweight network structures that maintain accuracy while reducing model complexity and computational resource requirements. Additionally, researchers can explore new training strategies, such as transfer learning, federated learning, etc., which can train models using decentralized data while protecting privacy and improving model generalization capabilities. Research can also be conducted on adaptive learning rate adjustment algorithms to accelerate model convergence, reduce training time, and improve training efficiency.

5.1Multimodal Data Fusion

Combine text data with other types of medical data (such as images, sounds, etc.) to achieve a more comprehensive disease diagnosis and prediction.

Multimodal learning is a hot research direction in the field of artificial intelligence, which improves model performance by integrating different types of data. In the medical field, this means combining text data (such as electronic health records) with image data (such as X-rays, MRI scans) and sound data (such as heart murmurs) to obtain more comprehensive disease characteristics. This

integration can help the model capture more dimensional information, thereby improving diagnostic accuracy and predictive reliability. Future research can explore how to effectively integrate these different types of data and how to design algorithms to handle the complex relationships between these data.

5.2 Cross-Disease Generalization Ability

Explore the model's generalization ability across different diseases to apply it to a broader range of clinical scenarios.

An ideal medical model should have cross-disease generalization capabilities, meaning a model trained on one disease can also have some recognition and prediction capabilities for other diseases. This ability can be achieved through multi-task learning, zero-shot learning, and other techniques. Future research can explore how to build a unified framework that allows models to understand and adapt to the characteristics of different diseases, thereby improving the model's applicability and flexibility.

5.3 Clinical Validation

Collaborate with clinical partners to conduct large-scale clinical validation of the model to assess its effectiveness and feasibility in actual clinical environments.

No matter how well the model performs in a laboratory environment, its ultimate value must be verified in a clinical environment. Future research needs to collaborate with medical institutions to conduct large-scale clinical trials to evaluate the actual application effects of the model. This includes the model's accuracy, reliability, user acceptance, and impact on patient treatment outcomes. In addition, research is needed on how to integrate the model into existing medical workflows and how to address medical regulations and ethical issues.

5.4 Data Privacy and Security

With the increasing amount of medical data, how to protect patient privacy and data security has become an important issue.

Future research needs to explore new data encryption technologies and privacy protection algorithms to ensure that patient data is not leaked during model training and application. This may involve the application of differential privacy, homomorphic encryption, and other technologies, as well as privacy protection improvements to existing machine learning algorithms.

5.5 Model Interpretability

Improve the transparency of the model so that its decision-making process can be understood by doctors and patients.

Models in the medical field need to have a certain degree of interpretability so that doctors and patients can understand the basis of their decisions. Future research can focus on developing new interpretability tools and technologies, as well as improving the structure of existing models to make them more transparent and understandable.

5.6 Cross-Cultural and Cross-Population Model Adaptability

Cross-Cultural and Cross-Population Model Adaptability: Considering the differences between different regions and populations, the model needs to be adaptable to the characteristics of different cultures and populations.

Future research can explore how to improve the model's cross-cultural adaptability by increasing data diversity and adjusting model parameters, making it

applicable globally.

Through these future research directions, we hope to further enhance model performance, expand its application scope, and ultimately have a positive impact on clinical practice. This will not only improve the efficiency and quality of medical services but also provide patients with more personalized and precise treatment plans. With the continuous advancement of technology and in-depth research, we have reason to believe that artificial intelligence will play an increasingly important role in the medical field and make a greater contribution to the cause of human health.

6 DISCUSS

The architecture model proposed in this experiment combines the Named Entity Recognition (NER) capability of Scispacy and the string matching capability of fuzzywuzzy, effectively improving the recognition accuracy of medical terminology and negative words in medical record data. The application of this model is expected to improve the quality and efficiency of medical record data analysis. This discussion will delve into the implications, potential enhancements, and broader applications of this integrated approach.

6.1 Integration of Scispacy and Fuzzywuzzy

The integration of Scispacy, a scientific domain-specific NER model, with fuzzywuzzy, a library for fuzzy string matching, is a novel approach that leverages the strengths of both technologies. Scispacy's proficiency in identifying entities in scientific text provides a robust foundation for extracting medical terms, while fuzzywuzzy's ability to find close matches in strings addresses the variability and inconsistencies in how medical terms are recorded in unstructured data.

6.2 Enhancing Recognition Accuracy

The recognition accuracy of medical terminology is crucial for the quality of data analysis. Misidentified or missed terms can lead to incorrect diagnoses, treatment plans, and research outcomes. By using Scispacy, the model can accurately identify standard medical terms. Fuzzywuzzy further enhances this by matching terms that may be misspelled or phrased differently, thus capturing a wider range of medical terminology present in the data.

6.3 Improving Data Analysis Quality and Efficience

The improved accuracy in recognizing medical terms directly translates to better quality data for analysis. This means that insights derived from the data are more reliable, leading to more informed decision-making in clinical settings. Moreover, the efficiency of the analysis is improved because the model can process large volumes of data quickly, identifying relevant terms and filtering out noise with minimal human intervention.

6.4 Addressing Variability in Medical Records

Medical records are notorious for their variability, with different healthcare providers using different terminologies and formats. The model's ability to handle this variability is a significant step forward. It can standardize terms and create a more uniform dataset, which is essential for large-scale data analysis and machine learning applications.

6.5 Potential for Automation in Data Cleaning

The model's capabilities suggest potential for automating the cleaning and preprocessing of medical record data. Traditionally, this process is labor-intensive

and prone to human error. With an automated system, the data can be cleaned more consistently, reducing the risk of errors and saving valuable time and resources.

6.6 Applications in Clinical Decision Support Systems

The model can be integrated into clinical decision support systems (CDSS) to assist healthcare providers in making informed decisions. By accurately identifying medical terms and negative words, the model can help in flagging critical information, such as contraindications or potential drug interactions, thus improving patient safety.

6.7 Enhancing Research and Drug Discovery

In research settings, particularly in drug discovery, the model can aid in the analysis of clinical trial data and literature. It can help in identifying patterns, adverse effects, and patient outcomes more accurately, accelerating the drug development process and potentially leading to more effective treatments.

6.8 Challenges and Limitations

Despite the potential benefits, there are challenges and limitations to consider. The model's performance is dependent on the quality of the training data. If the training data is biased or incomplete, the model may not perform as expected. Additionally, the model must be regularly updated to adapt to new medical knowledge and terminology.

6.9 Ethical Considerations

The use of such a model also raises ethical considerations, particularly regarding data privacy and security. Ensuring that patient data is handled in

compliance with regulations like HIPAA and GDPR is crucial. The model must be designed with robust security measures to protect sensitive health information.

6.10 Future Developments

Looking ahead, the model can be further developed to include more advanced natural language processing capabilities, such as context-aware entity recognition and sentiment analysis. This would allow for even more nuanced understanding of medical records, capturing not only what is said but also the implications and emotions behind the text.

In conclusion, the architecture model that combines Scispacy and fuzzywuzzy offers a promising approach to improving the recognition accuracy of medical terminology in unstructured data. Its application has the potential to significantly enhance the quality and efficiency of medical record data analysis, with far-reaching implications for clinical practice, research, and patient care. As the field of artificial intelligence continues to advance, such models will play an increasingly vital role in healthcare, driving innovation and improving outcomes. However, it is essential to address the challenges and ethical considerations to ensure that these technologies are used responsibly and effectively.

7 CONCLUSION

In conclusion, the experiment successfully demonstrates the efficacy of an architecture model that integrates Scispacy and fuzzywuzzy for extracting and labeling medical terms and negative words from medical record data. This innovative approach offers a robust solution for the accurate and efficient analysis of medical records, which is a critical component in the realm of healthcare informatics.

The development of this model signifies a significant advancement in the field, providing a new tool and method for medical record data analysis. By leveraging the strengths of Scispacy's Named Entity Recognition capabilities and the flexibility of fuzzywuzzy's string matching, the model is able to accurately identify and label medical terms, even in the presence of variations and inconsistencies in data recording. This capability is particularly valuable given the complex and nuanced nature of medical language, which often includes jargon, abbreviations, and evolving terminology.

The implications of this model extend beyond mere data extraction. It has the potential to significantly improve the quality and efficiency of medical services. For instance, accurate extraction of medical terms can lead to more precise patient diagnosis, better treatment planning, and enhanced patient care. It can also facilitate faster and more reliable clinical decision-making by providing healthcare professionals with relevant, structured information from unstructured medical records.

Moreover, this model can contribute to the advancement of medical research by enabling the analysis of large volumes of medical records to identify trends, correlations, and patterns that might otherwise be obscured in raw data. This can lead to new insights into disease mechanisms, treatment efficacy, and patient outcomes, ultimately accelerating the discovery of new medical knowledge.

The model also holds promise for improving public health initiatives by aiding in the surveillance of disease outbreaks and the tracking of health trends within populations. This can inform public health policies and interventions, leading to more effective management of public health crises.

However, the successful implementation of this model also requires consideration of several challenges. These include ensuring data privacy and security, addressing potential biases in the model, and maintaining the model's accuracy as medical knowledge and terminology evolve over time. It is also crucial to ensure that the model's outputs are interpretable and understandable to healthcare professionals, to facilitate trust and adoption in clinical settings.

Looking ahead, the model can be further refined and expanded to include additional functionalities, such as real-time analysis of medical records during clinical encounters or integration with other types of medical data for a more comprehensive patient profile. Continuous updates and improvements will be necessary to adapt to the ever-changing landscape of healthcare and to incorporate new medical discoveries and terminologies.

In summary, the experiment showcases an architecture model that stands at the forefront of medical data analysis, offering a powerful tool for improving the quality and efficiency of medical services. As this technology continues to evolve, it has the potential to revolutionize healthcare by enhancing clinical decision-making, driving medical research, and ultimately, improving patient outcomes. It is a testament to the power of combining cutting-edge natural language processing techniques with practical clinical applications, and it represents a significant step towards the realization of smarter, data-driven healthcare systems.

Wang Chunmei

REFERENCES

1. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University. (2023). Online Mendelian Inheritance in Man, OMIM®. Baltimore (MD). Retrieved [Date of Access], from https://omim.org/
2. Köhler S, Vasilevsky NA, Engelstad M, Foster E, McMurry JA, Ayme S, Baynam G, et al. (2023). The Human Phenotype Ontology in 2023. Nucleic Acids Research, 51(D1), D1203-D1213. doi:10.1093/nar/gkac1005. Retrieved [Date of Access], from https://hpo.jax.org/app/
3. Qiu, D., et al. (2020). Deep learning fusion model for Alzheimer's disease diagnosis using multimodality imaging and clinical data. Journal of Alzheimer's Disease, 75(3), 711-722.
4. Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine, 25(1), 37-43.
5. Shen, D., et al. (2017). Deep learning in medical image analysis. Annual Review of Biomedical Engineering, 19, 221-248.
6. Shen, W., et al. (2015). Multi-scale convolutional neural networks for lung nodule classification. IEEE Transactions on Biomedical Engineering, 62(9), 2063-2071.
7. Yosinski, J., et al. (2014). How transferable are features in deep neural networks? In Advances in Neural Information Processing Systems (pp. 3320-3328).
8. Topol, E. J. (2019). Digital medicine: A new era for healthcare. Foreign Affairs, 98(5), 83-93.
9. Wang, X., et al. (2021). A review of medical data visualization: Opportunities and challenges. Journal of Biomedical Informatics, 117, 103736.
10. Irvin, J. et al. (2019). CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. [Article]

11. Rajpurkar, P. et al. (2017). CheXNet: A Deep-Learning-Enabled Radiologist for Chest X-Rays. [Article]
12. Cheng, J. Z. et al. (2016). Deep Learning for Clinical Time Series Data: A Case Study on Sepsis Prediction. [Article]
13. Topol, E. J. (2019). High-Performing Medicine: The Next Healthcare Disruption? [Book]
14. Yosinski, J. et al. (2014). How Transferable Are Features in Deep Neural Networks? [Article]
15. Guo, Y., Liu, Y., Oerlemans, A., et al. (2016). Deep learning for visual understanding: A review. Neurocomputing, 187, 27-48. doi: 10.1016/j.neucom.2015.09.116.
16. Hayman, S. (1999). The McCulloch-Pitts model. In International Joint Conference on Neural Networks (pp. 4438-4439). Washington, USA: IEEE.
17. LeCun, Y., Bottou, L., Bengio, Y., et al. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278–2324. doi: 10.1109/5.726791.
18. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In International Conference on Neural Information Processing Systems (pp. 1097-1105). Lake Tahoe, USA: Curran Associates.
19. Shen, W., Zhou, M., Yang, F., et al. (2015). Multi-scale convolutional neural networks for lung nodule classification. Inf Process Med Imaging, 24, 588–599. doi: 10.1007/978-3-319-19992-4_46.