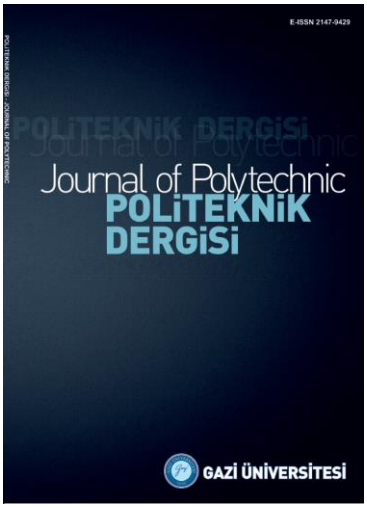

POLİTEKNİK DERGİSİ

JOURNAL of POLYTECHNIC

ISSN: 1302-0900 (PRINT), ISSN: 2147-9429 (ONLINE)
URL: http://dergipark.org.tr/politeknik

Semantic and structural analysis of MIMIC-CXR radiography reports with NLP methods

MIMIC-CXR radyoloji raporlarının DDİ yöntemleriyle anlamsal ve yapısal analizi

Yazar(lar) (Author(s)): *Ege Erberk USLU[1], Emine SEZER[2], Zekeriya Anıl GÜVEN[3]*

ORCID[1]: 0000-0001-9119-8574

ORCID[2]: 0000-0003-4776-6436

ORCID[3]: 0000-0002-7025-2815

To cite to this article: Uslu E. E, Sezer E. and Güven Z. A., "Semantic and structural analysis of MIMIC-CXR radiography reports with NLP methods", ***Journal of Polytechnic***, 27(5): 1955-1969, (2024).

Bu makaleye şu şekilde atıfta bulunabilirsiniz: Uslu E. E, Sezer E. and Güven Z. A., "Semantic and structural analysis of MIMIC-CXR radiography reports with NLP methods", ***Journal of Polytechnic***, 27(5): 1955-1969, (2024).

Erişim linki (To link to this article): http://dergipark.org.tr/politeknik/archive

DOI: 10.2339/politeknik.1395811

Semantic and Structural Analysis of MIMIC-CXR Radiography Reports with NLP Methods

Highlights

- *This article presents the first textual analysis of the MIMIC-CXR dataset.*
- *Conventional and AI-driven methods were implemented to analyze the dataset.*
- *The study conducted an analysis of the dataset utilizing n-gram and NER techniques.*
- *Analysis of NER shows the importance of lemmatization and POS tagging.*
- *TTR and Entropy values show lexical variety in MIMIC-CXR results in related fields.*

Graphical Abstract

Syntactic and semantic analysis of the MIMIC-CXR dataset, a collection of chest radiographs and accompanying radiology reports, is conducted to prepare the data for training a large language model.

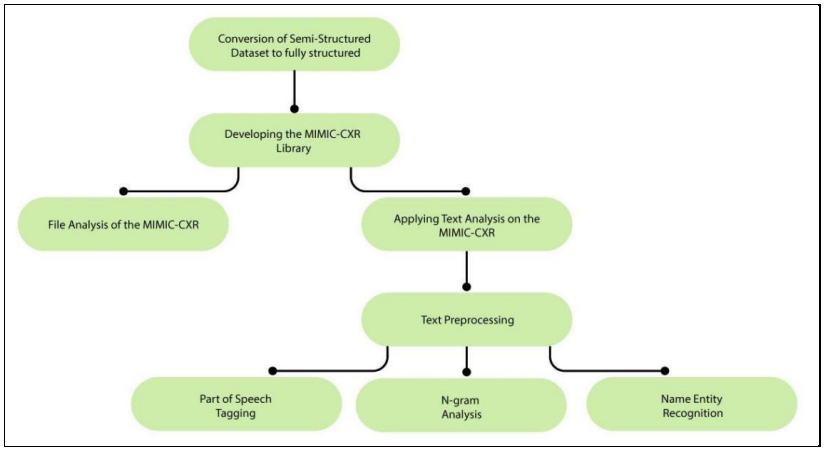

Figure. Procedure stages of the methodology

Aim

The study aims to provide valuable insights for NLP researchers by conducting a thorough semantic and structural analysis of the data prior to language model design.

Design & Methodology

The textual features of the MIMIC-CXR dataset are evaluated through the utilization of "Part of Speech Tagging", "n-Gram", and "NER" methodologies after applying various text preprocessing techniques, including sentence and word tokenization, punctuation and stop word removal, number deletion, lowercasing, and lemmatization.

Originality

The study is unique in that it provides valuable information to NLP researchers by performing a comprehensive semantic and structural analysis of the data before language model design through large semi-structured data sets. Another uniqueness is that it includes a detailed analysis on the MIMIC-CXR dataset.

Findings

The comparative effectiveness of a generic NER methodology vis-à-vis the n-gram technique in extracting word frequencies is relatively lower.

Conclusion

NER and N-gram analysis play crucial roles for semantic and structural analysis of semi-structured big datasets in the development of domain-specific large language models.

Declaration of Ethical Standards

The author(s) of this article declare that the materials and methods used in this study do not require ethical committee permission and/or legal-special permission.

Semantic and Structural Analysis of MIMIC-CXR radiography reports with NLP Methods

Araştırma Makalesi / Research Article

Ege Erberk USLU[1], Emine SEZER[1*], Zekeriya Anıl GÜVEN[2]

[1]Faculty of Engineering, Department of Computer Engineering, Ege University, Izmir, Turkiye
[2]Faculty of Engineering and Architecture, Department of Computer Engineering, Bakırçay University, Izmir, Turkiye

(Geliş/Received : 26.11.2023 ; Kabul/Accepted : 25.12.2023; Erken Görünüm/Early View : 02.02.2024)

ABSTRACT

Artificial intelligence that aims to imitate human decision-making processes, using human knowledge as a foundation, is a critical research area with various practical applications in different disciplines. In the health domain, machine learning and image processing techniques are increasingly being used to assist in diagnosing diseases. Many healthcare reports, such as epicrisis summaries prepared by clinical experts, contain crucial and valuable information. In addition to information extraction from healthcare reports, applications such as automatic healthcare report generation are among the natural language processing research areas based on this knowledge and experience. The primary goals are to reduce the workload of clinical experts, minimize the likelihood of errors, and save time to speed up the diagnosis process. The MIMIC-CXR dataset is a huge dataset consisting of chest radiographs and reports prepared by radiology experts related to these images. Before developing a natural language processing-based model, preprocessing steps were applied to the dataset, and the results of syntactic and semantic analyses performed on unstructured report datasets are presented. The results show that most examined words and phrases exhibit minimal semantic inference disparities. The generic named entity recognition method demonstrates comparatively lower effectiveness than the n-gram technique in extracting word frequencies. However, named entity recognition facilitated the identification of medical entities within the dataset. This study is expected to provide insights for developing language models, particularly for developing a natural language processing model on the MIMIC-CXR dataset.

Keywords: Natural language processing, MIMIC-CXR, chest radiology report, structural analysis, semantic analysis.

MIMIC-CXR Radyoloji Raporlarının DDİ Yöntemleriyle Anlamsal ve Yapısal Analizi

ÖZ

Yapay zeka, insan karar verme süreçlerini taklit etmeyi ve insan bilgisini temel almayı amaçlayan, farklı disiplinlerde çeşitli pratik uygulama alanına sahip kritik bir araştırma alanıdır. Sağlık alanında, makine öğrenimi ve görüntü işleme teknikleri hastalıkların teşhisine yardımcı olmak için giderek daha fazla kullanılmaktadır. Klinik uzmanlar tarafından hazırlanan epikriz özetleri gibi birçok sağlık raporu kritik ve değerli bilgiler içermektedir. Sağlık raporlarından bilgi çıkarmaya ek olarak, otomatik sağlık raporu oluşturma gibi uygulamalar, bilgi ve deneyime dayalı doğal dil işleme araştırma alanlarından biridir. Bu tür uygulamaların temel hedefleri, klinik uzmanların iş yükünü azaltmak, hata olasılığını en aza indirmek ve tanı sürecini hızlandırmak için zamandan tasarruf etmektir. MIMIC-CXR veri seti, radyoloji uzmanları tarafından çekilen göğüs röntgenleri ve bu görüntülerle ilgili raporlardan oluşan bir büyük veri setidir. Doğal dil işleme tabanlı bir model geliştirilmeden önce veri setine ön işleme adımları uygulanmış ve yapılandırılmamış rapor veri setleri üzerinde gerçekleştirilen sözdizimsel ve anlamsal analizlerin sonuçları sunulmuştur. Sonuçlar, incelenen kelimelerin ve ifadelerin çoğunun en az düzeyde anlamsal çıkarım eşitsizliği sergilediğini göstermektedir. Genel adlandırılmış varlık tanıma yöntemi, n-gram tekniğine göre kelime sıklıklarını çıkarmada nispeten daha düşük etkinlik göstermektedir. Ancak adlandırılmış varlık tanıma, veri kümesi içindeki tıbbi varlıkların tanımlanmasını kolaylaştırmaktadır. Bu çalışmanın, dil modellerinin geliştirilmesi, özellikle MIMIC-CXR veri seti üzerinde bir doğal dil işleme modelinin geliştirilmesi için araştırmacılara ışık tutması beklenmektedir.

Anahtar Kelimeler: Doğal dil işleme, MIMIC-CXR, göğüs radyoloji raporu, yapısal analiz, anlamsal analiz.

1. INTRODUCTION

The exponential growth of data in the digital age has led to the increasing importance of efficient and effective Natural Language Processing (NLP) approaches for information systems in text summarization, sentiment analysis, information extraction, named entity identification, association extraction, social media monitoring, text mining, language translation programs, and question-answering systems [1, 2]. NLP is a computational technique that applies different levels of linguistic analysis to transform natural language into a valuable representation for further processing [3]. In computer science and artificial intelligence, NLP is considered as a complex subject for understanding human natural language relies on words and the relationships between those words to create correct meaning. NLP approaches involve processing texts or documents by saving storage space, minimizing directory

*Corresponding Author
e-mail : emine.sezer@ege.edu.tr

size, and understanding the information provided to meet the user's needs. In addition, it increases the effectiveness of good documentation and information retrieval processes [4].

In addition to the structured data typically found in electronic health records (EHRs), these records also contain rich text information, including reports written by experts in their native language. This makes EHRS a paramount source of information for the health domain, and facilitating access to this information would likely accelerate the diagnosis and treatment process, reduce labor and time costs, and improve the standard of health services [5].

Correct diagnosis is essential for effective treatment planning, and immediate diagnosis is critical in some diseases to prevent fatal or permanent disabilities. Radiology is the cornerstone of contemporary health care, providing detailed clinical information for disease detection, staging, and treatment planning, while also playing an indispensable role in monitoring and predicting outcomes. Radiology reports are unstructured free text, and therefore require effective and automated information extraction solutions to transform them into computer-manageable presentations for large-scale analysis [6]. With the help of NLP tools, the transformation of unstructured text into encoded data becomes feasible, facilitating the automated identification and extraction of information from radiology reports. This process proves valuable in various clinical applications, including diagnostic surveillance, cohort creation, quality assessment, and computer vision labeling [7,8].

Clinicians require access to extensive patient data and medical literature to provide high-quality, quantitative health care. However, the vast majority of this data is stored in an unstructured format [9]. This unstructured data, which contains essential information, is interpreted by clinicians and used in the diagnosis and treatment process. However, the conclusions drawn from this data can vary from clinician to clinician, which can impact the diagnosis and treatment process. Variability in clinician interpretation of unstructured data can lead to inaccurate diagnoses and treatments, decreased efficiency, and increased costs. NLP aids clinicians in diagnosis and treatment by improving the accuracy and efficiency of these processes.

Despite its benefits, NLP remains an underutilized technique for extracting data from large volumes of radiology reports in both research and clinical practice settings due to the high cost of development and the difficulty of generalizing models. Dictionary-based [10] and rule-based [11] analysis account for most best-performing NLP techniques. Although this approach may be successful for a specific application, it necessitates a significant amount of manual effort to tailor the appropriate methodologies to a particular study's case and dataset.

In recent years, deep learning has begun to provide solutions that give researchers tools to create automatic classification models for medical images that are widely adaptable and do not require human input, or manual contribution [12]. In spite of this, for clinical reports, the ambiguity of expression in free text, lexical fluctuations, non-grammatical steps, and frequent abbreviations make it difficult to apply deep learning algorithms to extract information from text [13].

Chest radiology is the most widely used medical imaging technique globally for assessing the thoracic cavity. Chest radiographs are used in medical research to diagnose acute and chronic cardiopulmonary diseases, confirm the accurate placement of devices such as pacemakers, central lines, chest tubes, and gastric tubes, and as well as the identification of acute and chronic cardiopulmonary conditions [14]. The MIMIC Chest X-ray (MIMIC-CXR) dataset is a large publicly available database of labeled chest radiographs. It contains patient data consisting of chest radiology images and reports, meeting the requirements of the National Health Insurance and Personal Data Protection Authority [10, 11].

This article presents a comprehensive statistical and semantic analysis of the MIMIC-CXR dataset, uncovering its distinctive characteristics through an examination of the technical information, patient numbers, chest radiology examinations, and text analysis. The study provides valuable insights for NLP researchers by conducting a thorough semantic and structural analysis of the data prior to language model design. The article is structured as follows: the second section presents a detailed explanation of the methodology employed in this study, followed by the third section, which describes the experiments conducted on the MIMIC-CXR dataset, along with the results of these experiments and their implications for future research and studies in the final section.

2. MATERIAL AND METHOD

Firstly, the semistructured MIMIC-CXR dataset was converted to a fully-structured format in accordance with the guidelines provided on the MIMIC-CXR website [11]. Subsequently, a Pandas library-based tool was developed to facilitate various functions such as reading, modifying, and managing the complete structured data set. The created library was used for file and text analysis. During the text analysis process, the preprocessing steps described in the later sections of the article were applied. The N-gram method was employed for data analysis, specifically comparing the uni(1)-gram, bi(2)-gram, and tri(3)-gram approaches. The named entity recognition method utilized a biomedical-ner-all model [15] based on DistilBERT. All preprocessing steps were used in this method as well. A comparative investigation was conducted to examine the utilization and non-utilization of root analysis (lemmatization) and word type tagging techniques as preprocessing steps in the implementation of named entity recognition. The "FINDINGS" and

"IMPRESSION" sections of free text reports were used in the experiments. The procedure stages of this study are shown in Figure 1.

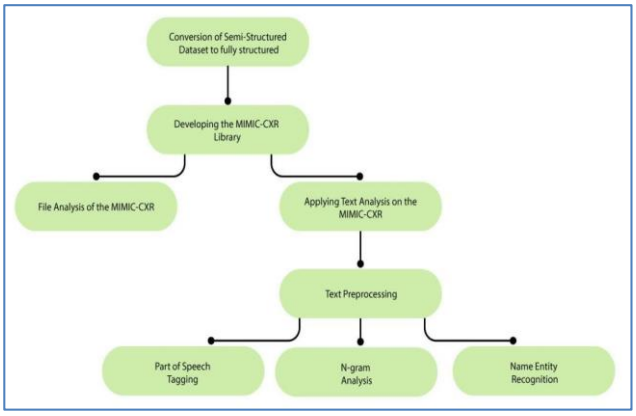

Figure 1. Procedure stages of the methodology

2.1. MIMIC-CXR Dataset

The publicly available MIMIC-CXR dataset, which has volume, variety, and value features determined as the priority characteristics of big data, has been referenced from nearly 300 articles since its publication in 2019. To demonstrate the conducted studies, an investigation, focused on diagnosing pneumothorax accurately using image and NLP methods revealed that NLP techniques improve diagnostic accuracy [16]. Another study used deep learning techniques to develop a hierarchical approach for labeling chest X-ray abnormalities [17].

The novel framework developed to automatically generate impressions and summarize key information in radiology reports significantly reduced the workload of radiologists [18].

The MIMIC-CXR dataset was constructed using chest X-ray images from patients in the emergency department of Beth Israel Deaconess Medical Center between 2011 and 2016. This dataset consists of 227,835 image studies, including 377,110 radiology images of 65,379 patients, with semi-structured free-text radiology reports which comply with health insurance and data protection regulations. Access to the data requires user registration, authentication, and acceptance of a data usage policy. A sample from the dataset are given in Figure 2 [10].

Within the semi-structured reports in the MIMIC-CXR dataset, there are fields named with eleven labels as “IMPRESSION”, “FINDINGS”, “LAST_PARAGRAPH”, “COMPARISON”, “INDICATION”, “EXAMINATION”, “TECHNIQUE”, “HISTORY”, “NOTIFICATION”, “RECOMMENDATIONS”, and “WET READ”.

The MIMIC-CXR dataset organizes files by classifying them based on the first two digits of patient numbers, starting from p10 and ending with p19. As a result, the patients' reports are distributed among ten distinct folders. In order to protect patient anonymity, the patients are assigned numbers in accordance with the MIMIC-CXR mockup. These numbers are used in the lower layer of the folder hierarchy. Within each patient folder, one can find the radiologists' reports pertaining to the corresponding patient's radiological images. To provide a clear understanding of the organization, Figure 3 illustrates the file and folder hierarchy.

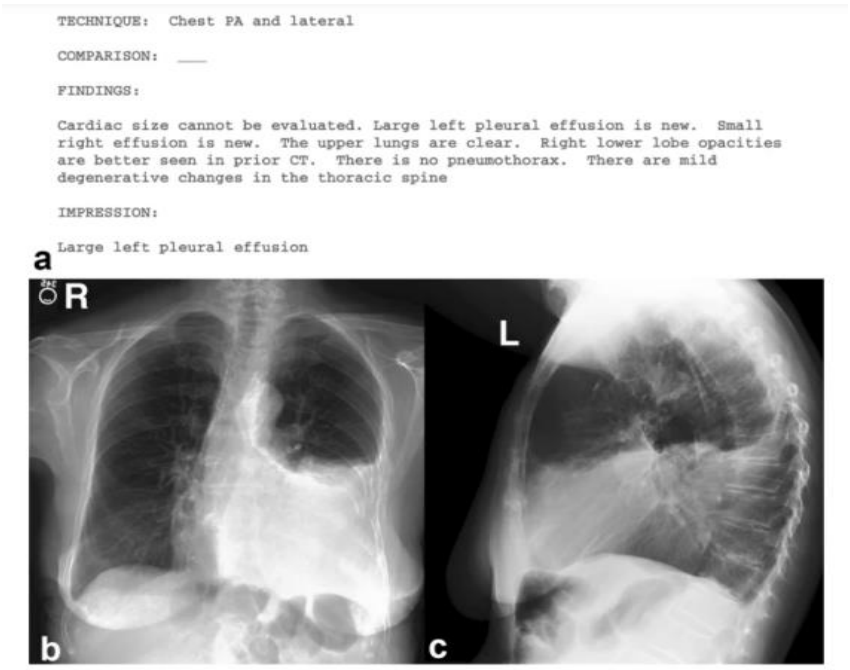

Figure 2. A sample study from MIMIC-CXR. (a), the radiology report provides the interpretation of the image. Personal patient information has been removed and replaced with three underscores (_ _ _). Two chest X-rays are shown for this study: (b) front view (left image) and (c) side view (right image)

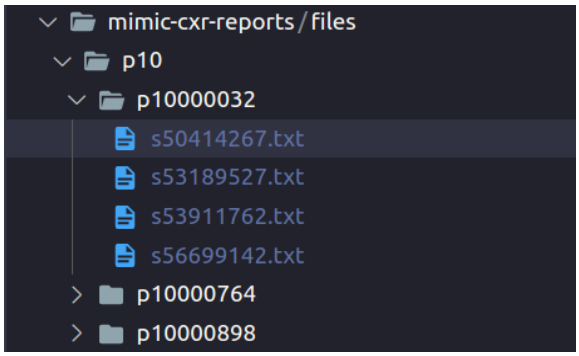

Figure 3. Hierarchy of files of the MIMIC-CXR Dataset

The MIMIC-CXR dataset was converted from a semi-structured to a fully structured format to improve the efficiency of data processing. The dataset was reviewed to identify limitations, such as the presence of typos and inconsistent spelling. These limitations were addressed by fixing typos, providing index values for inseparable partitions, and creating a list of identical names and correct spellings. A library was created using the Pandas Library to facilitate data processing. This library allows for reading and storing data in CSV format, and it integrates methods for further data analysis explained in following sections.

2.2. File Analysis of the MIMIC-CXR

The data stored with the help of created library includes eleven fields' information. Statistics on the presence of information in these fields have been revealed. The data is columnar, with eleven field information, patient number, and report number. All entries of the eleven fields were checked for emptiness. Blank entries in the "FINDINGS" sections were reviewed for emptiness. These blank entries were not included in the study but were included in statistical information. The frequency-finding technique revealed the total number of patients, the number of reports per patient, and the number of patients who came once or more than once. Further elaboration on the file analysis component of the MIMIC-CXR dataset can be found in experiment in Section 3.1.

2.3. Text Analysis on the MIMIC-CXR

The text preprocessing steps in this study include sentence and word tokenization, punctuation and stop word removal, number deletion, lowercasing, and lemmatization [19] as illustrated in Figure 4. Sentence tokenization divides text into sentences, while word tokenization separates it into words. Punctuation marks are removed for clarity. Stop words, which do not change meaning, are eliminated. Numbers are deleted. Characters are converted to lowercase for consistency. Lemmatization identifies the root form of words. This process is like finding a transformation for normalizing a word [20]. Table 1 shows an example of the preprocessing steps.

The MIMIC-CXR dataset's textual characteristics were analyzed through the utilization of "Part of Speech Tagging," "n-Gram," and "NER" methodologies.

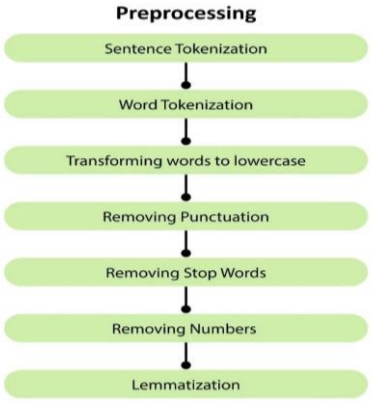

Figure 4. Preprocessing phases

Part-of-speech (POS) tagging assigns labels to words in a sentence, indicating their grammatical function, such as noun, verb, adjective, or conjunction. This tagging procedure assigns labels to the different parts of speech based on the surrounding context [17]. This undertaking can prove to be difficult on account of polysemy, wherein a word can possess multiple meanings. For example, the term "hot" can pertain to both the temperature of a meal and its spicy flavor. Nevertheless, ambiguity often dissipates when a word is used in conjunction with other words [17]. For instance, the phrase "hot delivery" unambiguously indicates that "hot" pertains to temperature. This study made use of the NLTK library for the word type labeling and utilized nouns derived from the lemmatization procedure. Adjectives, conjunctions, and other components were excluded as they were considered less relevant to the characteristics of the dataset.

N-gram-based approaches are widely used in contemporary NLP applications. It is a phrase consisting of n consecutive elements in specific textual data. N-grams are combinations of ‘n’ adjacent elements obtained from textual data. The number ‘n’ denotes the total number of words in an n-gram. These consecutive elements can be words, characters, POS labels, or any other item [18]. In this study, the CountVectorizer function from the scikit-learn library was used to extract n-grams. However, a crucial consideration arises when calculating n-grams. Due to the removal of sentence-ending punctuation marks in the preprocessing stage, the sentences may appear incomplete. This can lead to the generation of extraneous n-grams. For example, in the sentence "John has not a symptom called edema. In that case, he is healthy," the bigram "edema case" is not present. But after removing punctuation, the sentence becomes "John has not a symptom called edemaIn that case, he is healthy," which is seen as a single sentence by the CountVectorizer library, creating the false bigram "edema case." To avoid this error, individual sentences were separated and stored in arrays before processing. This eliminated the potential for the error.

Table 1. Textual Preprocessing Phases with example

Text	***Preprocessing Algorithm***	***After Preprocessing***
John does not have a symptom called edema in his lungs. This leads that he is in good health.	*Sentence Tokenization*	*'John does not have a symptom called edema in his lungs.', 'This leads that he is in good health.'*
'John does not have a symptom called edema in his lungs.', 'This leads that he is in good health.'	*Word Tokenization*	*'John', 'does', 'not', 'have', 'a', 'symptom', 'called', 'edema', 'in', 'his', 'lungs', '.', 'This', 'leads', 'that', 'he', 'is', 'in', 'good', 'health', '.'*
'John', 'does', 'not', 'have', 'a', 'symptom', 'called', 'edema', 'in', 'his', 'lungs', '.', 'This', 'leads', 'that', 'he', 'is', 'in', 'good', 'health', '.'	*Transforming words to lowercase*	*'john', 'does', 'not', 'have', 'a', 'symptom', 'called', 'edema', 'in', 'his', 'lungs', '.' ,'this', 'leads', 'that', 'he', 'is', 'in', 'good', 'health', '.'*
'john', 'does', 'not', 'have', 'a', 'symptom', 'called', 'edema', 'in', 'his', 'lungs', '.' ,'this', 'leads', 'that', 'he', 'is', 'in', 'good', 'health', '.'	*Removing Punctuation*	*'john', 'does', 'not', 'have', 'a', 'symptom', 'called', 'edema', 'in', 'his', 'lungs', 'this', 'leads', 'that', 'he', 'is', 'in', 'good', 'health'*
'john', 'does', 'not', 'have', 'a', 'symptom', 'called', 'edema', 'in', 'his', 'lungs', 'this', 'leads', 'that', 'he', 'is', 'in', 'good', 'health'	*Removing Stopwords*	*'john', 'symptom', 'called', 'edema', 'lungs', 'leads', 'good', 'health'*
'john', 'symptom', 'called', 'edema', 'lungs', 'leads', 'good', 'health'	*Removing Numbers*	*'john', 'symptom', 'called', 'edema', 'lungs', 'leads', 'good', 'health'*
'john', 'symptom', 'called', 'edema', 'lungs', 'leads', 'good', 'health'	*Lemmatization*	*'john', 'symptom', 'called', 'edema', 'lung', 'lead', 'good', 'health'*

Named entity recognition (NER) is an extracting information process that aims to identify and classify specific types of information units referred to as Named Entities (NEs) [21]. The process consists of two main steps: entity identification and entity classification. In the first step, entity identification, words or phrases in texts are recognized, while in the second step, entity classification, the recognized entities are assigned to specific categories [22]. NER tasks often rely on deep learning models, such as the BERT language model. BERT is a transformer-based language model that has bidirectional language representation, which makes it well-suited for analyzing text [23].

In this study, a DistilBERT-based model was used to identify names of biomedical entities for named entity detection. The model was trained on the Maccrobat 2018 dataset, which is based on the English language, and it is capable of distinguishing 107 medical entities from texts [15]. Unlike the previous stages where individual sentences were separated and analyzed independently, sentence separation was not performed in this stage to maintain consistency within the model. Pre-processing steps, including word segmentation, punctuation removal, stop word removal, and number removal, were carried out as in the previous stages. The efficacy of root analysis and part-of-speech tagging operations would be comparatively evaluated. A sample report with Biomedical-NER is showed in Figure 5.

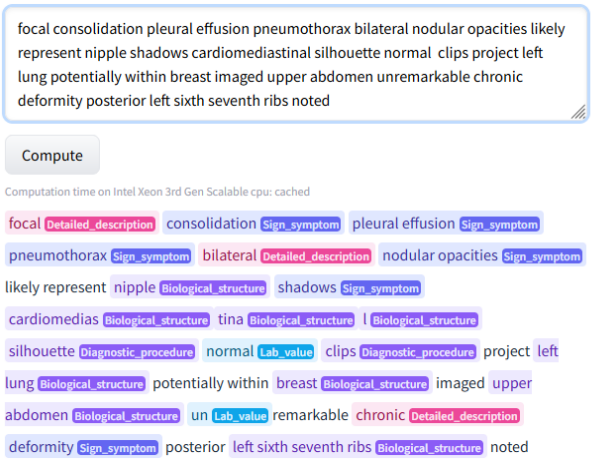

Figure 5. Application of Biomedical-NER to a sample report

3. THE RESEARCH FINDINGS AND DISCUSSION

This section presents the experiments performed on the MIMIC-CXR dataset and their results. Experiments include statistical, structural, and semantic analyses to prepare free text datasets such as the MIMIC-CXR dataset for NLP studies.

3.1. File Analysis of the MIMIC-CXR

Among the total of 227,835 files in MIMIC-CXR dataset, 227,781 files include the factors that would be considered in the disease analysis. However, 54 files lack information on the disease due to the inability to extract any relevant data from the file structure, as explained in Section 2.1.

The chest radiography reports encompass a total of non-mandatory 11 fields which provide valuable explanations from radiologists regarding the radiography image of the patient. Table 2 presents the quantity of fields found in the entirety of the files, along with their corresponding ratios. Figure 6 visually illustrates the distribution ratios of the files based on the fields that are present in the radiology reports of the patients. Furthermore, the term "Valid" denotes files that contain pertinent fields, whereas "Invalid" signifies files without such fields. According to the findings depicted in Figure 6, the fields denoted as "IMPRESSION", "FINDINGS", "COMPARISON", "INDICATION", "EXAMINATION", "TECHNIQUE", and "HISTORY" hold significant importance. However, the remaining categories, namely "LAST PARAGRAPH", "NOTIFICATION", "RECOMMENDATION", and "WET READ" do not possess sufficient quantity to efficiently be included in the studies.

Another aspect that can be derived from the dataset files is the count of radiological image reports pertaining to patients. This provides insight into the frequency of chest radiology for each individual. The total number of patient records amounts to 65374. Approximately 49.98% of these patients possess merely a solitary report, while the remaining 50.02% have multiple reports. Table 3 showcases the initial five patients who possess the highest number of radiology reports, along with the count of reports attributed to each patient, and the ratio of these reports in relation to the overall number of reports.

The information within radiological imaging reports can convey a meaning that extends beyond mere numerical values. However, when evaluated on a patient basis, it becomes evident that the significance of this information becomes apparent. For instance, within a given dataset, it becomes imperative for a particular patient to have access to the information regarding "Pleural Effusion" in their respective reports. Yet, if this information is absent in the reports of other patients, it becomes inconsequential within the dataset as a whole. It was observed that almost half of the patients came to the clinic once, while the other half came more than twice, as shown in Figure 7. However, patients with more than one visit displayed an accumulation range between 1 and 5. Although this may initially appear to complicate patient-based analysis, the number of patients falling within other ranges is likely sufficient. Thus, the MIMIC-CXR dataset has facilitated the execution of person-based data analysis in research studies.

Table 2. File distribution

Attribute	*IMPRESSION Field Count*	*FINDINGS Field Count*	*LAST_PARAGRAPH Field Count*	*COMPARISON Field Count*	*INDICATION Field Count*	*EXAMINATION Field Count*	*TECHNIQUE Field Count*	*HISTORY Field Count*	*NOTIFICATION Field Count*	*RECOMMENDATION Field Count*	*WET_READ Field Count*
Invalid	*3822*	*72065*	*217267*	*63991*	*61965*	*128177*	*146427*	*170779*	*222066*	*225989*	*227704*
Valid	*189561*	*155716*	*10514*	*16379*	*165816*	*99604*	*81354*	*57002*	*5715*	*1792*	*77*
Invalid Ratio (%)	*16.78*	*31.64*	*95.38*	*28.09*	*27.20*	*56.27*	*64.28*	*74.98*	*97.49*	*99.21*	*99.97*
Valid Ratio (%)	*83.22*	*68.36*	*4.62*	*71.91*	*72.80*	*43.73*	*35.72*	*25.02*	*2.51*	*0.79*	*0.03*

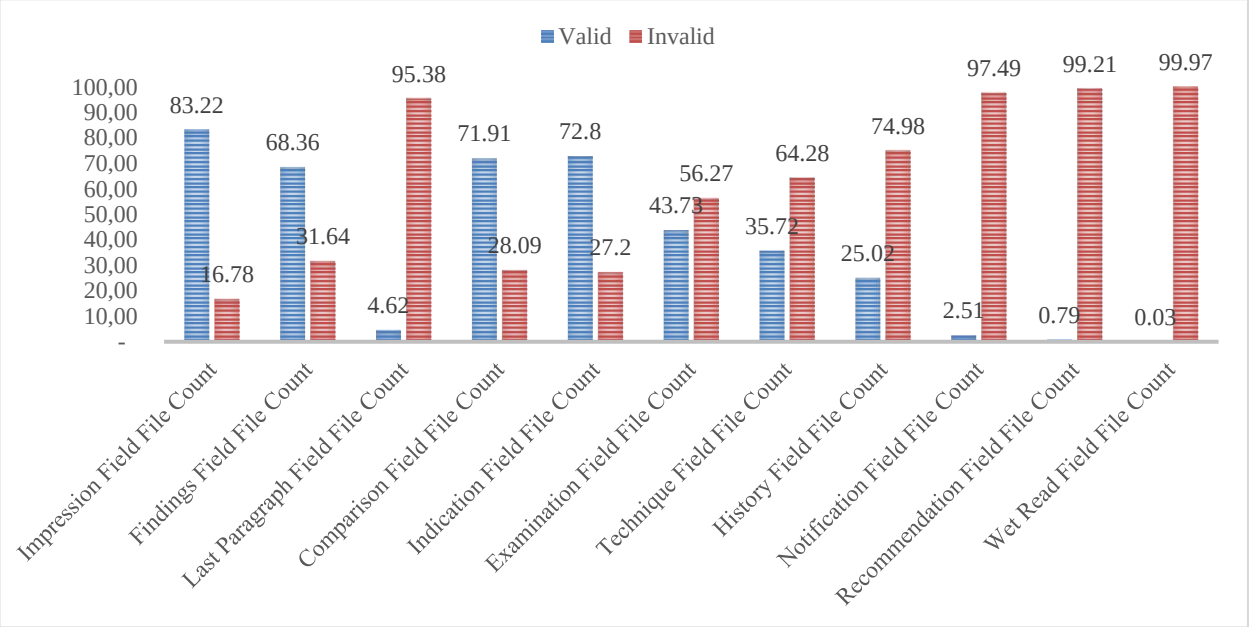

Figure 6. File distribution ratios according to the sections (%)

Table 3. Patient distribution of MIMIC-CXR Dataset

Patient ID	*Number of Patient Reports*	*Ratio of Number of Patient Reports to Total Number of Reports*
16454913	*158*	*0.07%*
15936063	*131*	*0.06%*
18295542	*127*	*0.06%*
12043836	*124*	*0.05%*
19674244	*108*	*0.05%*
Total top 5	*648*	*0.28%*
Total	*227,781*	

The information within radiological imaging reports can convey a meaning that extends beyond mere numerical values. However, when evaluated on a patient basis, it becomes evident that the significance of this information becomes apparent. For instance, within a given dataset, it becomes imperative for a particular patient to have access to the information regarding "Pleural Effusion" in their respective reports. Yet, if this information is absent in the reports of other patients, it becomes inconsequential within the dataset as a whole. It was observed that almost half of the patients came to the clinic once, while the other half came more than twice, as shown in Figure 7. However, patients with more than one visit displayed an accumulation range between 1 and 5. Although this may initially appear to complicate patient-based analysis, the number of patients falling within other ranges is likely sufficient. Thus, the MIMIC-CXR dataset has facilitated the execution of person-based data analysis in research studies.

The MIMIC-CXR dataset consists of 14 distinct findings related to thoracic ailments. The Chexpert system [24] assigns suitable labels to each patient report based on the corresponding findings. These data were classified into three separate labels of -1, 0, and 1 by Chexpert. The binary digit '1' signifies the presence of a specific observation, whereas the digit '0' indicates its absence. Conversely, a numerical value of '-1' indicates a lack of elucidation for the corresponding observation. Results that remained unresolved and were outside the scope of this study were considered inconclusive. The assignment of the value '0' was made to variables that hold the value of '-1'. Reports containing multiple findings were excluded from the dataset as per the parameters of this study. Figure 8 showcases data pertaining to reports that contain individual results.

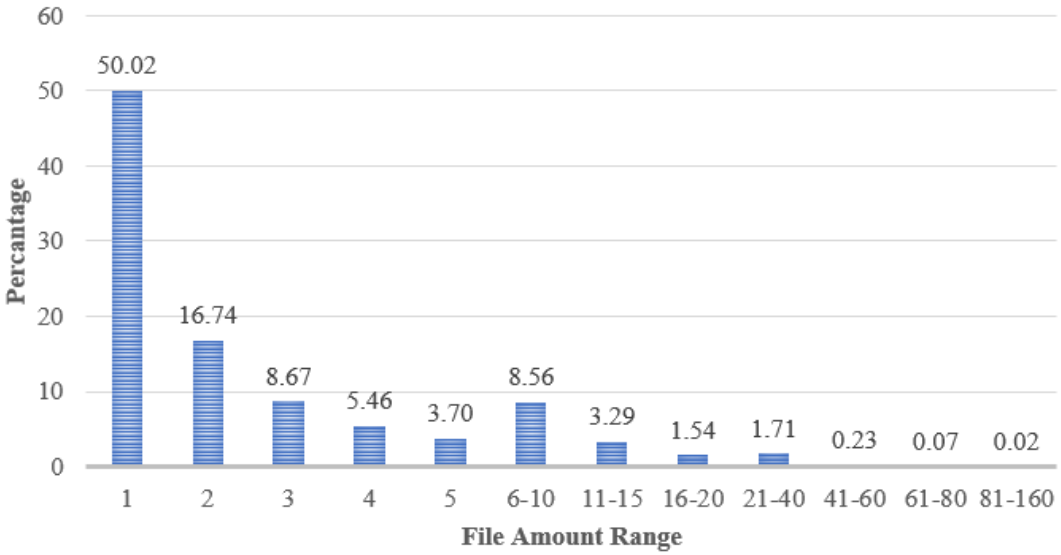

Figure 7. Patient's report number range distribution

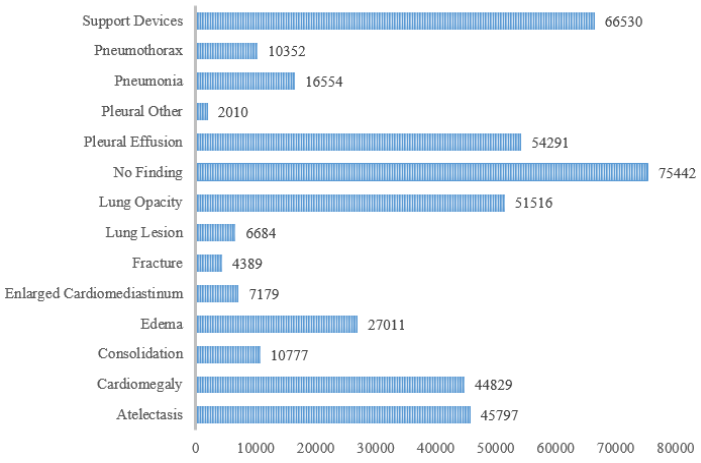

Figure 8. File distribution of impressions in reports

3.2. Text Analysis

To reveal the textual properties of the dataset, n-gram and named entity recognition methods were used in this study. Thanks to these methods, the information about what the data wants to tell and what it contains has been reached.

N-gram analysis was performed as unigram, bi-gram, and trigram. These analyses were performed on the dataset's findings, impressions, and data containing both domains. An example of unigram analysis about MIMIC-CXR is given in Figure 9.

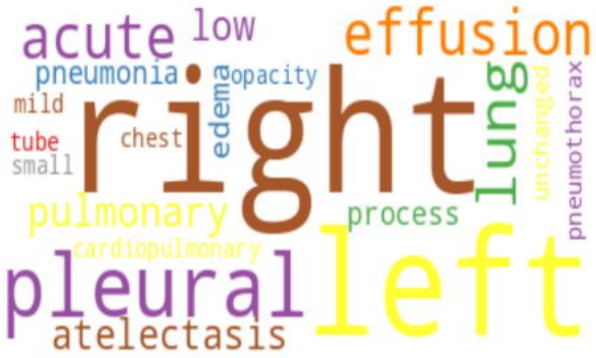

Figure 9. Word cloud of unigram "IMPRESSION"

Before analyzing the word groups, it is crucial to grasp the independent meanings of individual words. For this purpose, a single-word analysis of radiology image reports was conducted. The most frequently occurring words extracted from the "FINDINGS" fields are presented in Figure 10. The most prevalent terms include ‘effusion’, ‘lung’, ‘pleural’, ‘right’, and ‘pneumothorax’. These frequently repeated words encompass medical findings, anatomical locations, organs, and descriptors of the findings' states.

As a result of the unigram analysis on the "IMPRESSION" fields, the most frequently repeated words are ‘right’, ‘left’, ‘pleural’, ‘effusion’, and ‘lung’. In this section, the words expressing the anatomical location and localization are more common. However, it is observed that the exact words are frequently repeated in the general findings and impression parts. As a result of the unigram analysis of the common "FINDINGS" fields, the most repeated words are shown in Table 4. It is seen that the most frequently repeated words are ‘effusion’, ‘lung’, and ‘pleural’. Words with diagnostic values appear in the common findings section. Words expressing anatomical location and localization are seen in lower ranks. It is seen that similar features are seen when compared with the reports in which only the findings are obtained.

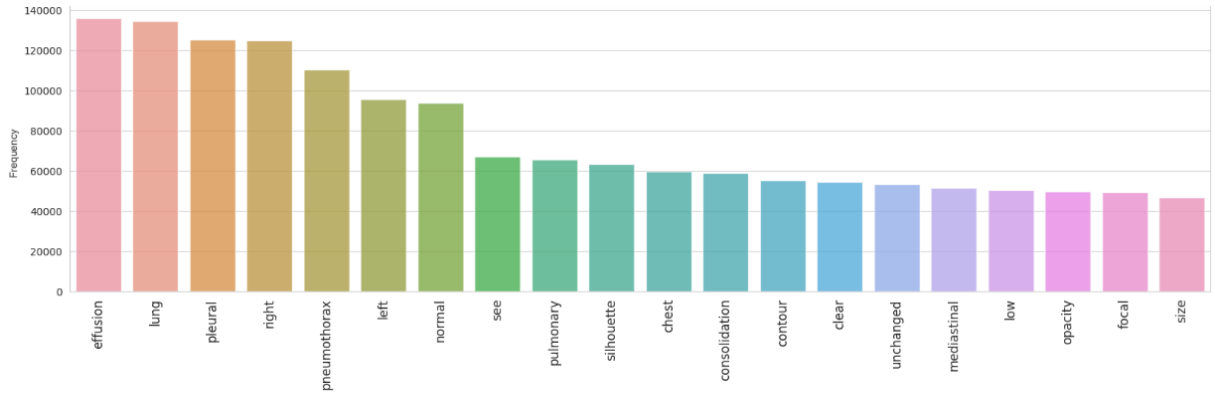

Figure 10. Word distribution of unigram "FINDINGS"

Table 4. Unigram joint "FINDINGS"

Index	*Words*	*Frequencies*	*Percentage*
1	*effusion*	*116700*	*3.059*
2	*lung*	*114360*	*2.997*
3	*pleural*	*106118*	*2.781*
4	*right*	*100668*	*2.638*
5	*pneumothorax*	*100288*	*2.629*
6	*normal*	*86229*	*2.260*
7	*left*	*74471*	*1.952*
8	*see*	*59087*	*1.549*
9	*consolidation*	*55857*	*1.464*
10	*silhouette*	*53884*	*1.412*

Due to unigram analysis of common "IMPRESSION" sections, the most repeated words are shown in Figure 11. These include ‘acute’, ‘process’, ‘cardiopulmonary’, ‘right’, and ‘left’. The "IMPRESSION" section is characterized by a higher prevalence of terms denoting the situation, anatomical location, and localization. Similar features were observed when compared to reports with only impressions. Unigram analysis of words belonging to the findings and impression fields in the dataset revealed a relationship between these two fields. However, to reveal this relationship more clearly, an examination of reports containing both fields in the dataset showed minimal differences in the overall findings section vocabulary. Conversely, semantic differences in the impression section were remarkable. While terms expressing anatomical localization were previously more prevalent in this area, this analysis revealed a shift towards medical findings.

Unigram analysis provides insights into the semantic relationships within the dataset. However, the resulting data does not fully capture these relationships nor the precise intended meaning of the fields. Therefore, bigram analysis emerges as a potential approach for achieving more meaningful results. These analyses revealed the emergence of words that contribute to medical diagnoses. The most frequently occurring bigrams of the "FINDINGS" fields are presented in Table 5. These include ‘pleural effusion’, ‘effusion pneumothorax’, ‘focal consolidation’, and ‘lung clearance’. These recurring bigrams correspond to findings that could potentially lead to more accurate diagnoses. The frequently repeated bigrams in the "IMPRESSION" field include ‘pleural effusion’, ‘acute cardiopulmonary’, ‘pulmonary edema’, and ‘cardiopulmonary process’. These findings in the "IMPRESSION" sections can aid in making more accurate diagnoses.

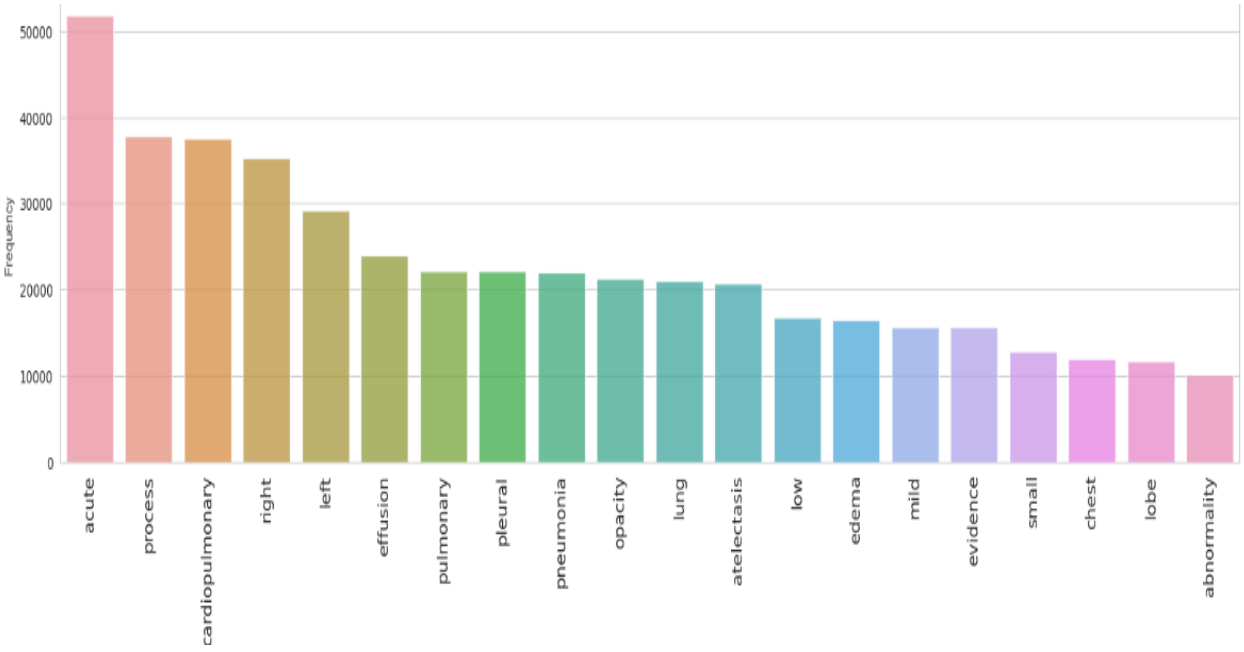

Figure 11. Word distribution of unigram joint "IMPRESSION"

Table 5. Bigram Joint "FINDINGS"

Index	*Words*	*Frequencies*	*Percentage*
1	*pleural effusion*	*111778*	*2.894*
2	*effusion pneumothorax*	*61013*	*1.580*
3	*focal consolidation*	*40361*	*1.045*
4	*lung clear*	*34990*	*0.906*
5	*hilar contour*	*31436*	*0.814*
6	*cardiomediastinal silhouette*	*31030*	*0.803*
7	*pulmonary edema*	*29410*	*0.761*
8	*heart size*	*26330*	*0.682*
9	*view chest*	*26251*	*0.679*
10	*lung volume*	*25111*	*0.650*

As a result of the bigram analysis on the common "FINDINGS" fields, the most repeated words are shown in Figure 12. The most frequently repeated words are ‘pleural effusion’, ‘effusion pneumothorax’, ‘focal consolidation’, and ‘lung clear’. Findings and impressions share similar characteristics in this regard. In general, findings that will guide the diagnosis are also encountered in this area. The results of analyzing common (joint) impression areas using bigrams are analyzed. This analysis's most frequently repeated words are acute cardiopulmonary, cardiopulmonary process, pleural effusion, pulmonary edema, and low lobe. These findings and impressions are similar to what is typically encountered in this field. Generally, the findings that will aid in diagnosis can also be found in this area.

Given that bigram analysis revealed expressions with deeper meanings, trigram analysis was utilized to further refine the data. Trigram analysis demonstrated that medical diagnoses identified through bigram were accompanied by both condition and localization. This observation suggests that the application of trigram analysis in the healthcare domain holds promise for enhanced effectiveness. The most frequently encountered trigrams in the "FINDINGS" sections include ‘pleural effusion pneumothorax’, ‘lateral chest view’, and ‘mediastinal hilar contour.’ These trigrams allow for the extraction of detailed diagnostic and localization information.

Upon analyzing the "IMPRESSION" fields using a trigram approach, Figure 13 shows the most repeated words. The most frequently repeated words are ‘acute cardiopulmonary process’, ‘right pleural effusion’, ‘left pleural effusion’, and ‘bilateral pleural effusion’. The words with the localization of the diagnosis are remarkable. This increases the specificity of findings and diagnoses.

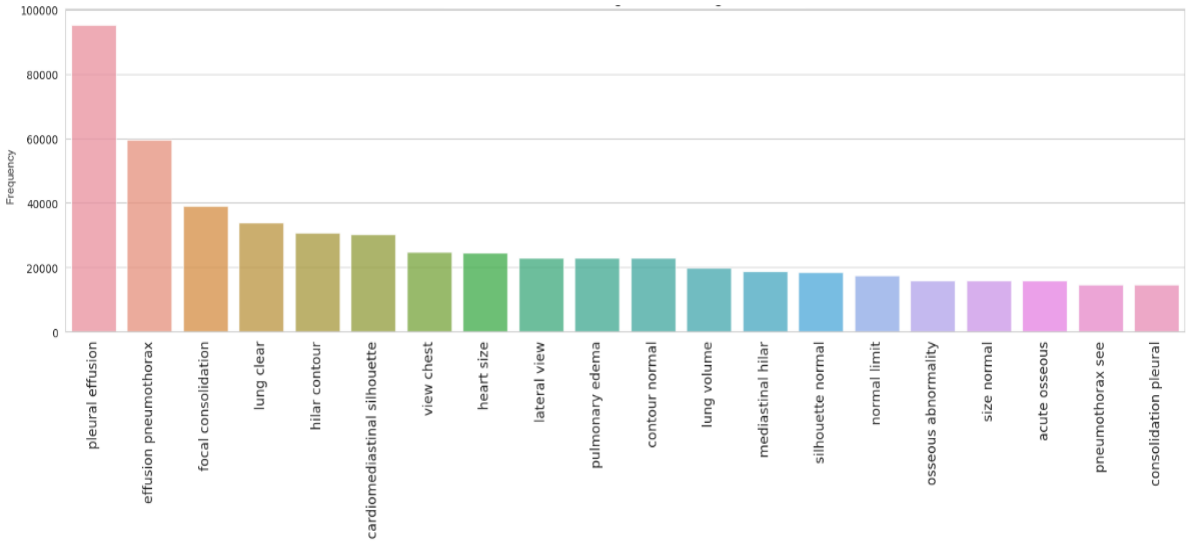

Figure 12. Word Distribution of Bigram Joint "FINDINGS"

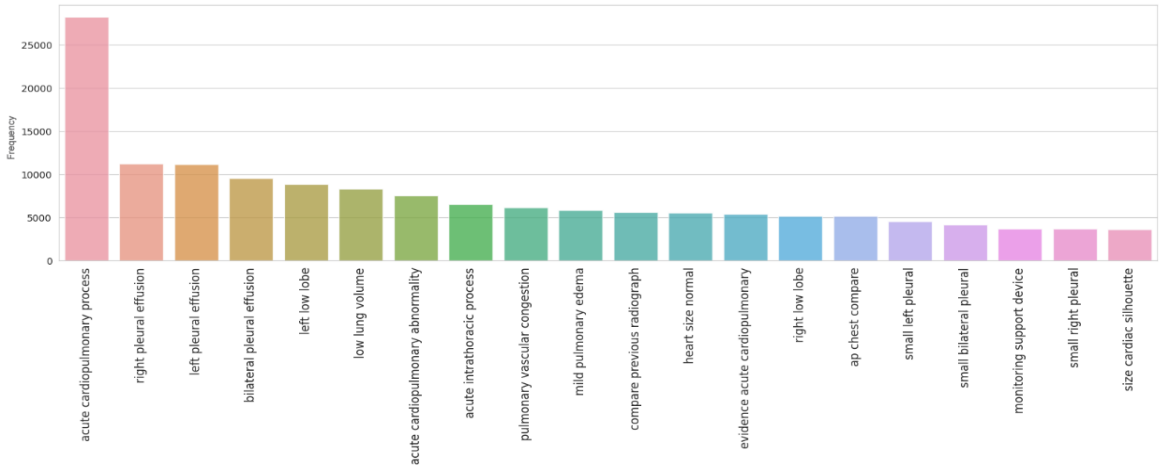

Figure 13. Word distribution of trigram "IMPRESSION"

Upon analyzing "FINDINGS" section utilizing trigrams, it has been determined that the words with the highest frequency are presented in Table 6. Linguistic expressions that provide further elaboration on the concept of localization are encountered. The characteristics exhibited are analogous to those of discoveries. A result has been obtained by analyzing frequently used "IMPRESSION" fields using trigram methodology. Figure 14 displays the words that have been repeated the most.

Table 6. Trigram Joint "FINDINGS"

Index	*Words*	*Frequencies*	*Percentage*
1	*pleural effusion pneumothorax*	*48429*	*1.983*
2	*lateral view chest*	*19244*	*0.788*
3	*mediastinal hilar contour*	*18411*	*0.754*
4	*acute osseous abnormality*	*15751*	*0.645*
5	*cardiomediastinal silhouette normal*	*15410*	*0.631*
6	*consolidation pleural effusion*	*14605*	*0.598*
7	*heart size normal*	*13740*	*0.563*
8	*focal consolidation pleural*	*13350*	*0.547*
9	*hilar contour normal*	*13335*	*0.546*
10	*pa lateral view*	*10594*	*0.434*

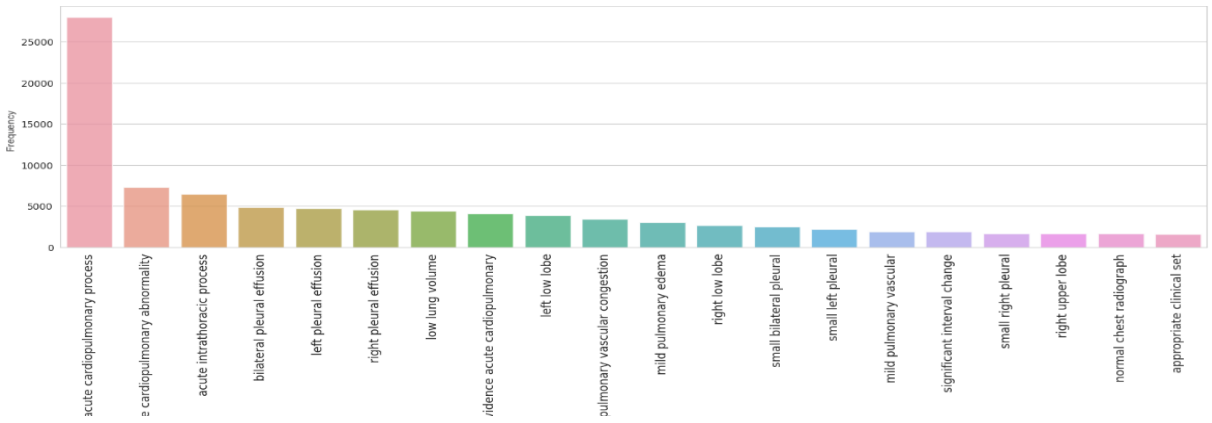

Figure 14. Word Distribution of Trigram Joint "IMPRESSION"

The specified n-grams are assessed in terms of time, entropy, frequency number, unique value, and lexical diversity. Entropy indicates the number of bits that can be conveyed in words pertaining to each domain. Lexical variation pertains to the range of a textual dataset. The increase in the number of bits signifies the diversity within the data. The greater the number of words in a text, the more bits will be used to express the data. The linear variation of the lexical diversity with the number of bits is observed. Although the quantity of textual data pertaining to the impression sections in the dataset is lower compared to that of the findings sections, it is evident that the lexical diversity is higher as a consequence of all n-gram analyses. It has been deduced that it would be more fitting to use in studies encompassing the overall sense impressions as they generally exhibit greater textual diversity. Statistical evidence of these explanations is demonstrated in Table 7. Before conducting NER analysis, it is imperative to comprehend the semantics and contents of the data. To accomplish this, n-gram analysis was used in this study. Upon evaluating the acquired data, the named entity labeling model for this data was selected [20]. The model entails relatively comprehensive terminology within the medical domain. The rationale behind opting for this model is the desire to obtain a holistic perspective of the data rather than a specific one. Consequently, the general characteristics of the data are ascertained. It becomes possible to extract information about the medical category to which the words belong. For instance, as illustrated in Table 8, medical structure and sign symptom groups predominantly fall under the "FINDINGS" sections. It was observed that for this study, which encompasses textual reports of chest radiological imaging, a narrower range of terminology.

An experiment was conducted in the context of named entity tagging and recognition methods to investigate the utilization of root analysis and word type tagging processes. As seen from Table 9, there was no significant difference in semantic inference from the data in most words and phrases. The efficiency of extracting word frequencies through a general NER approach is comparatively lower than that of the n-gram technique. The application of NER in this study reveals the presence of medical groups within the dataset. Consequently, word groups such as ‘organs’, ‘findings’, ‘diagnostic procedures’, ‘disease’, and ‘severity (condition)’ appear to hold considerable value in this dataset.

Table 7. Overall N-Gram Analysis Results

	N-GRAM											
	"FINDINGS"			*"IMPRESSION"*			*Common*					
							1		*2*		*3*	
	1	*2*	*3*	*1*	*2*	*3*	*"FINDINGS"*	*"IMPRESSION"*	*"FINDINGS"*	*"IMPRESSION"*	*"FINDINGS"*	*"IMPRESSION"*
Time (Minutes/ Seconds)	*0.765*	*0.761*	*0.727*	*0.561*	*0.541*	*0.536*	*0.616*	*0.261*	*0.601*	*0.258*	*0.595*	*0.252*
Entropy bits/Word	*8.19*	*12.13*	*14.94*	*8.41*	*12.68*	*15.54*	*8.05*	*8.11*	*11.79*	*12.06*	*14.51*	*14.77*
Lexical Diversity	*0.002*	*0.047*	*0.184*	*0.0025*	*0.069*	*0.242*	*0.002*	*0.004*	*0.048*	*0.092*	*0.183*	*0.290*
Type-token-ratio (TTR)												

Table 8. NER Analysis without using Lemmatization and Pos Tagging

"FINDINGS"				*"IMPRESSION"*				*Common*			
Words	*Word Frequency*	*Word Group*	*Word Group Frequency*	*Words*	*Word Frequency*	*Word Group*	*Word Group Frequency*	*Words*	*Word Frequency*	*Word Group*	*Word Group Frequency*
normal	*71089*	*Biological_structure*	*726338*	*acute*	*47398*	*Detailed_description*	*401091*	*normal*	*66626*	*Biological_structure*	*623407*
lungs	*55561*	*Sign_symptom*	*593982*	*pulmonary*	*42759*	*Biological_structure*	*361961*	*lungs*	*53069*	*Sign_symptom*	*497811*
chest	*50630*	*Diagnostic_procedure*	*538585*	*pneumonia*	*39375*	*Sign_symptom*	*321687*	*consolidation*	*46094*	*Diagnostic_procedure*	*440559*
consolidation	*48839*	*Detailed_description*	*498068*	*atelectasis*	*32661*	*Disease_disorder*	*311443*	*clear*	*45874*	*Lab_value*	*414091*
clear	*47655*	*Lab_value*	*490827*	*mild*	*30577*	*Diagnostic_procedure*	*266230*	*chest*	*45691*	*Disease_disorder*	*402395*
unchanged	*46817*	*Disease_disorder*	*489813*	*chest*	*29814*	*Lab_value*	*236411*	*focal*	*42672*	*Detailed_description*	*395488*
pulmonary	*46500*	*Severity*	*159790*	*edema*	*25605*	*Severity*	*139422*	*pulmonary*	*37494*	*Severity*	*129345*
focal	*45907*	*Therapeutic_procedure*	*80103*	*small*	*25507*	*Therapeutic_procedure*	*93342*	*unchanged*	*36775*	*Therapeutic_procedure*	*56124*
mild	*34776*	*Qualitative_concept*	*8505*	*right*	*24345*	*Coreference*	*5175*	*mild*	*29532*	*Qualitative_concept*	*7508*
atelectasis	*32586*	*Coreference*	*5667*	*cardiopulmonary*	*24334*	*Clinical event*	*4628*	*silhouette*	*28100*	*Coreference*	*4493*

Table 9. NER Analysis with using Lemmatization and Pos Tagging

"FINDINGS"				"IMPRESSION"				Common			
Words	Word Frequency	Word Group	Word Group Frequency	Words	Word Frequency	Word Group	Word Group Frequency	Words	Word Frequency	Word Group	Word Group Frequency
lung	96231	Biological_structure	735106	acute	47710	Detailed_description	436968	lung	83885	Biological_structure	631149
normal	70442	Sign_symptom	558176	pulmonary	43543	Biological_structure	366193	normal	65897	Sign_symptom	467927
chest	51148	Detailed_description	548487	pneumonia	39304	Disease_disorder	322312	consolidation	47036	Diagnostic_procedure	446717
consolidation	49665	Diagnostic_procedure	543409	lung	38976	Sign_symptom	308642	chest	46092	Detailed_description	436309
unchanged	47863	Disease_disorder	513882	atelectasis	32475	Diagnostic_procedure	267568	clear	45377	Disease_disorder	420934
pulmonary	47461	Lab_value	465421	mild	30572	Lab_value	227888	focal	42603	Lab_value	392822
clear	47261	Severity	157977	chest	29755	Severity	139625	pulmonary	38137	Severity	127608
focal	45853	Therapeutic_procedure	86829	right	28291	Therapeutic_procedure	96614	unchanged	37570	Therapeutic_procedure	61348
pleural effusion	40263	Qualitative_concept	10310	small	26198	Coreference	5841	pleural effusion	33266	Qualitative_concept	9143
mild	34692	Coreference	6287	left	25337	Qualitative_concept	4984	mild	29466	Coreference	5054

4. RESULTS

In this article, MIMIC-CXR dataset is analyzed using MIMIC-CXR file and text analysis, which are explicated comprehensively. The study examines the quantitative and qualitative data within the dataset through statistical analysis and textual interpretation. The present exposition pertains to data consisting of eleven distinct sections within radiology reports, encompassing information about these respective fields. The dataset being analyzed consists of eleven discrete data fields, including variables such as patient and report identification numbers. The segment discussing the distribution of files suggests that the data exhibits an imbalanced distribution. The total count of medical records pertaining to patients is 65,374. According to the data provided, it is highly improbable that around half of the patients (49.98%) have a single record in the clinic, while the other half (50.02%) have multiple records. This allows for the analysis of individual documents. For example, a patient must review the information on "Pleural Effusion" contained within the medical records of a specific individual within a dataset. However, the significance of the dataset would be compromised in the absence of such data in the remaining instances.

Despite the aforementioned challenge, an adequate quantity of cases across various ranges will facilitate analysis based on patients. The implementation of the MIMIC-CXR dataset has enabled the execution of person-centered data analysis in academic research. The current research utilized n-gram and named entity recognition methodologies to detect the textual characteristics of the dataset. The utilization of these methodologies has facilitated the acquisition of insight regarding the significance and substance of the data. The analysis of unigrams revealed a limited range of linguistic diversity in the "FINDINGS" section.

On the other hand, the modifications in semantics within the perceptual domain are significant from an academic standpoint. The current analysis highlights the importance of medical breakthroughs in anatomical localization terminology, which had previously enjoyed greater prevalence. Examining the individual words occurring in the fields of "FINDINGS" and "IMPRESSION" within the dataset revealed a significant association between these two fields. To clarify this correlation, an analysis was conducted on the dataset's records encompassing both variables. The analysis of individual words, also known as unigrams, significantly comprehends the semantic associations present in the provided dataset.

5. CONCLUSION AND FUTURE WORK

This article assessed a methodology for annotating and identifying proper nouns, focusing on root analysis and word classification tagging techniques. The research results show that the majority of the words and phrases analyzed do not exhibit significant disparities in semantic inference. The comparative effectiveness of a generic NER methodology vis-à-vis the n-gram technique in extracting word frequencies is relatively lower. The application of NER has facilitated the identification of medical entities in the given dataset. Implementing NER has yielded significant advantages in identifying word phrases about organs, diagnoses, diagnostic procedures, medical conditions, and the corresponding severity levels of such conditions in the provided dataset. Diverse outcomes can be derived from text analysis by using distinct language models. Various observations can be presented through the utilization of diverse NLP techniques for text analysis.

A comprehensive review of the peer-reviewed literature revealed no published work on statistical and semantic analysis of the MIMIC-CXR data set. Therefore, it was not possible to compare the results obtained in this study with those of previous research.

As future work, a language model for chest radiology reports can be designed considering the semantic and statistical results presented in this article. Another research topic could be determining and analyzing how many different clusters can be identified in the MIMIC-

CXR dataset. Additionally, it may be possible to evaluate and analyze the effectiveness of clustering on the MIMIC-CXR dataset using various combinations of dimensionality reduction, topic modeling, and LM. Another issue may be evaluating the effectiveness of multiple classification algorithms on the MIMIC-CXR dataset. Textual data can be used to develop an AI model to predict 14 findings in chest radiology reports in the MIMIC-CXR dataset. It may be possible to develop an AI model that aims to predict patient reports based on their medical findings and assist healthcare professionals in clinical decision-making.

DECLARATION OF ETHICAL STANDARDS

The author(s) of this article declare that the materials and methods used in this study do not require ethical committee permission and/or legal-special permission.

AUTHORS' CONTRIBUTIONS

Ege Erberk USLU: Performed the experiments and analyse the results.

Emine SEZER: Analyse the results and wrote the manuscript.

Zekeriya Anıl GÜVEN: Analyse the results and wrote the manuscript.

CONFLICT OF INTEREST

There is no conflict of interest in this study.

REFERENCES

[1] Bilen, B., and Horasan, F., "LSTM Network based Sentiment Analysis for Customer Reviews", ***Journal of Polytechnic***, 25(3):959-66, (2022).

[2] Alnawas, A., and Arıcı, N., "The Corpus Based Approach to Sentiment Analysis in Modern Standard Arabic and Arabic Dialects: A Literature Review". ***Journal of Polytechnic***, 21(2):461-70, (2018). doi:10.2339/politeknik.403975.

[3] Khurana, D., Koli, A., Khatter, K. et al., "Natural language processing: state of the art, current trends and challenges", ***Multimed Tools Appl***, 82: 3713–3744, (2023). https://doi.org/10.1007/s11042-022-13428-4

[4] Hallinan, J. T. P. D., Feng, M., Ng, D., Sia, S. Y., Tiong, V. T. Y., Jagmohan, P., Makmur, A., Thian, Y. L., "Detection of Pneumothorax with Deep Learning Models: Learning From Radiologist Labels vs Natural Language Processing Model Generated Labels", ***Academic Radiology,*** 29(9): 1350–1358, (2022). https://doi.org/10.1016/j.acra.2021.09.013

[5] Névéol, A., Deserno, T. M., Darmoni, S. J., Güld, M. O., and Aronson, A. R., "Natural language processing versus content-based image analysis for medical document retrieval", ***Journal of the American Society for Information Science and Technology***, 60(1):123-134, (2009).

[6] Banerjee, I., Chen, M. C., Lungren, M. P., Rubin, D.L., "Radiology report annotation using intelligent word embeddings: Applied to multi-institutional chest CT cohort", ***J Biomed Inform***., 2018 Jan;77:11-20, (2018). doi: 10.1016/j.jbi.2017.11.012.

[7] Kalra, A., Chakraborty, A., Fine, B., and Reicher, J., "Machine learning for automation of radiology protocols for quality and efficiency improvement", ***Journal of the American College of Radiology***, 17(9): 1149-115, (2020).

[8] Abro, A. A. , Talpur, M. S. H. & Jumani, A. K., "Natural Language Processing Challenges and Issues: A Literature Review", ***Gazi University Journal of Science***, 36(4):1522-1536, (2023). doi: 10.35378/gujs.1032517.

[9] López-Úbeda, P., Martín-Noguerol, T., Juluru, K., and Luna, A., "Natural Language Processing in Radiology: Update on Clinical Applications", ***Journal of the American College of Radiology***, 19(11): 1271-1285 (2022).

[10] Johnson, A.E.W., Pollard, T.J., Berkowitz, S.J., Greenbaum, Nathenial R., Lungren, Matthew P., Deng, Chih-ying, Mark, Roger G., Horng, Steven., "MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports" ***Sci Data,*** 6: 317 (2019). https://doi.org/10.1038/s41597-019-0322-0.

[11] MIMIC-CXR Database, Retrieved January 3, 2023, from https://physionet.org/content/mimic-cxr/2.0.0/

[12] Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P. C., Mark, R., Mietus, J. E., Moody, G. B., Peng, C.K, and Stanley, H. E., "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals" :***irculation [Online]***, 101 (23):e215–e220, (2000).

[13] Kundeti, S. R., Vijayananda, J., Mujjiga, S., and Kalyan, M., "Clinical named entity recognition: Challenges and opportunities", ***IEEE International Conference on Big Data (Big Data),*** 1937-1945, (2016).

[14] Rajpurkar P, Irvin J, Ball RL, Zhu K, Yang B, Mehta H, et al., "Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists", ***PLoS Med,*** 15(11): e1002686. (2018). https://doi.org/10.1371/journal.pmed.1002686

[15] d4data/biomedical-ner-all Hugging Face. (n.d.). Retrieved February 5, 2023, from https://huggingface.co/d4data/biomedical-ner-all.

[16] Liu, H., Christiansen, T., Baumgartner, W.A., Verspoor, Karin., "BioLemmatizer: a lemmatization tool for morphological processing of biomedical text", ***J Biomed Semant,*** 3, 3 (2012). https://doi.org/10.1186/2041-1480-3-3

[17] Cutting, D., Kupiec, J., Pedersen, J., and Sibun, P., "A practical part-of-speech tagger", ***In Third conference on applied natural language processing***, 133-140, (1992, March).

[18] Sidorov, G., Velasquez, F., Stamatatos, E., Gelbukh, A., and Chanona-Hernández, L., "Syntactic N-grams as machine learning features for natural language processing", ***Expert Syst. Appl.***, 41: 853-860, (2014).

[19] Donnelly, L. F., Grzeszczuk, R., and Guimaraes, C. V., "Use of natural language processing (NLP) in evaluation of radiology reports: an update on applications and technology advances", ***Seminars in Ultrasound, CT and MRI,*** 43(2): 176-181, WB Saunders, (2022).

[20] Plisson, J., Lavrac, N., and Mladenic, D., A Rule based Approach to Word Lemmatization, (2004).

[21] Sharnagat, R., "Named entity recognition: A literature survey", ***Center For Indian Language Technology***, 1-27, (2014).

[22] Marrero, M., Urbano, J., Sánchez-Cuadrado, S., Morato, J., and Gómez-Berbís, J. M., "Named entity recognition: fallacies, challenges and opportunities", ***Computer Standards & Interfaces***, 35(5): 482-489, (2013).

[23] Devlin, J., Chang, M. W., Lee, K., and Toutanova, K., "Bert: Pre-training of deep bidirectional transformers for language understanding", ***arXiv preprint***, (2018). arXiv:1810.04805.

[24] Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund H., Haghgoo, B., Ball, R., Shpanskya, K., Seekings, J., Mong, D. A., Halabi, S. S., Sandberg, J. K., Jones, R., Larson, D. B., Langlotz, C. P., Patel, B. N., Lungren, M. P., and Ng, A. Y., "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison", ***In Proceedings of the AAAI conference on artificial intelligence,*** 33(1): 590-597, (2019, July).