JSS Mahavidyapeetha
JSS SCIENCE AND TECHNOLOGY UNIVERSITY
JSS Technical Institutions' Campus, Mysuru – 570006

AUTOMATED IDENTIFICATION OF MEDICAL PRESCRIPTION

Thesis submitted in partial fulfillment of curriculum prescribed for the award of the degree of

BACHELOR OF ENGINEERING
IN
COMPUTER SCIENCE AND ENGINEERING

by

Aditya Raj
01JST20CS180

Harshit Sharma
01JST20CS065

Pavan Prakash
(01JST20CS109)

Shania Vijay
01JST20CS147

Under the Guidance of

Mr. VIJAY M B
Assistant Professor,
Department of Computer Science and Engineering,
JSS STU, Mysuru.

……………………………………………

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
May 2024

JSS MAHAVIDYAPEETHA
JSS SCIENCE AND TECHNOLOGY UNIVERSITY
SRI JAYACHAMARAJENDRA COLLEGE OF ENGINEERING

- Constituent College of JSS Science and Technology University
- Approved by A.I.C.T.E
- Governed by the Grant-in-Aid Rules of Government of Karnataka
- Identified as lead institution for World Bank Assistance under TEQIP Scheme

CERTIFICATE

This is to certify that the work entitled "AUTOMATED IDENTIFICATION OF MEDICAL PRESCRIPTION" is a bonafide work carried out by Aditya Raj(01JST20CS180), Harshit Sharma(01JST20CS065), Pavan Prakash(01JST20CS109), and Shania Vijay(01JST20CS147) in Partial fulfillment for the award of the Degree of BACHELOR OF ENGINEERING IN COMPUTER SCIENCE AND ENGINEERING of Sri Jayachamarajendra College of Engineering, JSS Science and Technology University, Mysuru, during the year 2024. It is certified that all corrections / suggestions indicated during CIE have been incorporated in the report. The project report has been approved as it satisfies the academic requirements in respect of project work prescribed for the Bachelor of Engineering degree.

Guide
Mr. Vijay M B
Assistant Professor,
Dept. of Computer Science
and Engineering,
SJCE, JSS STU, Mysuru.

Head of the Department
Dept. of Computer
Science and Engineering
SJCE, JSS STU, Mysuru.

Examiners: 1. ………………………

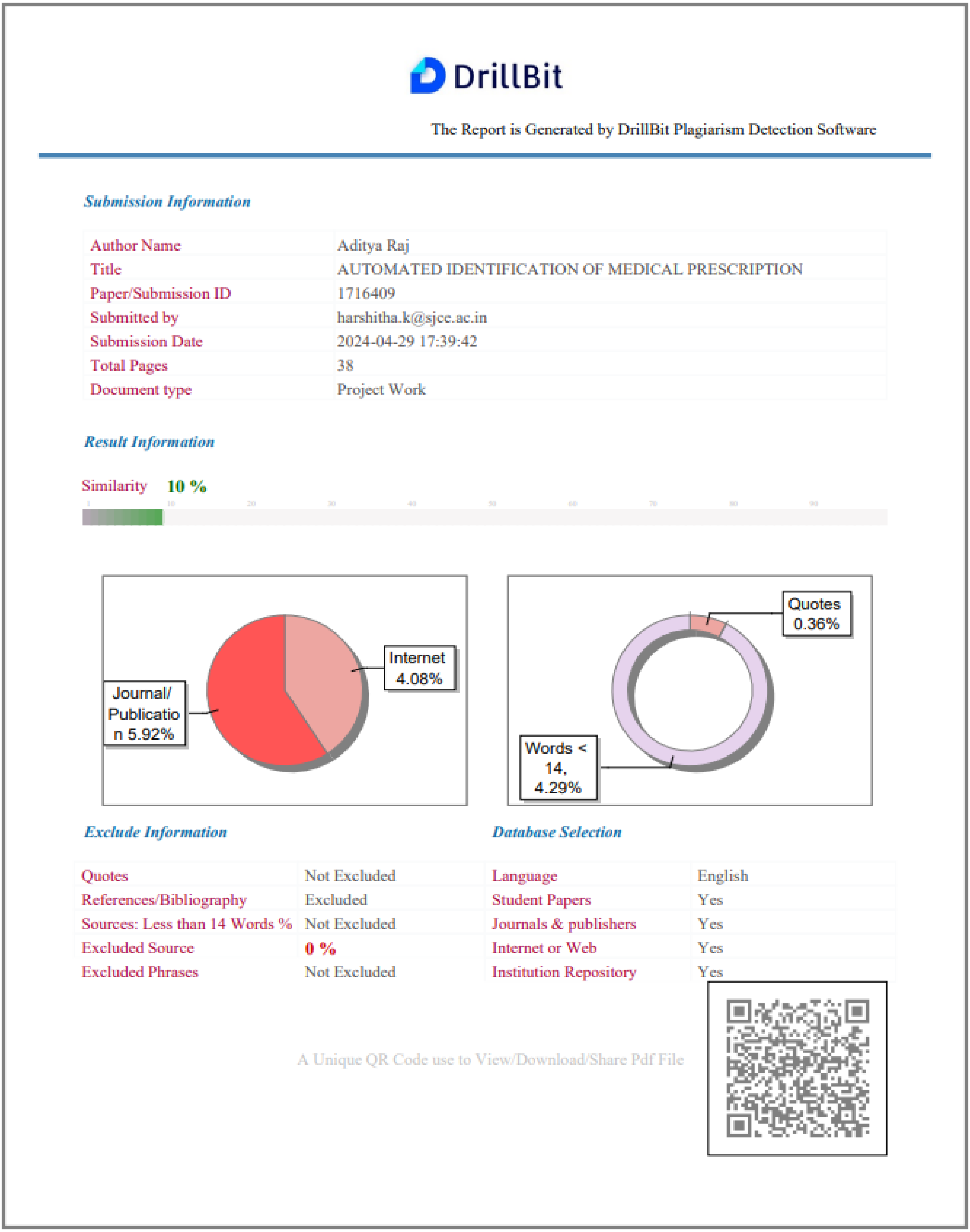
Place: Mysuru

2. ……………………...

Date:

3. ……………………...

PLAGIARISM CHECK CERTIFICATE



DrillBit

The Report is Generated by DrillBit Plagiarism Detection Software

Submission Information

Author Name	Aditya Raj
Title	AUTOMATED IDENTIFICATION OF MEDICAL PRESCRIPTION
Paper/Submission ID	1716409
Submitted by	harshitha.k@sjce.ac.in
Submission Date	2024-04-29 17:39:42
Total Pages	38
Document type	Project Work

Result Information

Similarity **10 %**

1
10
20
30
40
50
60
70
80
90
Journal/ Publication 5.92%
Internet 4.08%
Quotes 0.36%
Words < 14, 4.29%

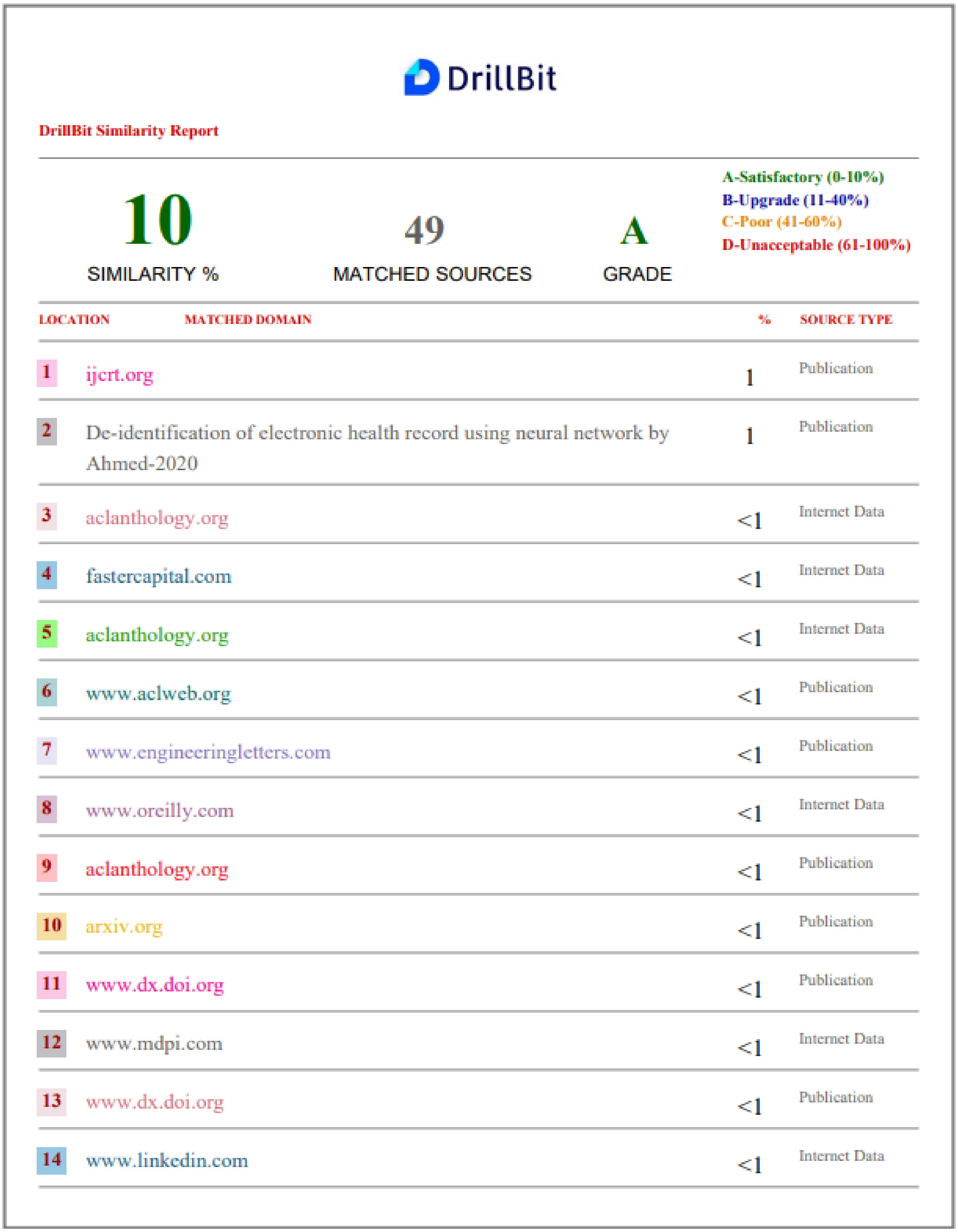
Exclude Information

Quotes	Not Excluded
References/Bibliography	Excluded
Sources: Less than 14 Words %	Not Excluded
Excluded Source	**0 %**
Excluded Phrases	Not Excluded

Database Selection

Language	English
Student Papers	Yes
Journals & publishers	Yes
Internet or Web	Yes
Institution Repository	Yes

A Unique QR Code use to View/Download/Share Pdf File



DrillBit

DrillBit Similarity Report

10	49	A	A-Satisfactory (0-10%) B-Upgrade (11-40%) C-Poor (41-60%) D-Unacceptable (61-100%)
SIMILARITY %	MATCHED SOURCES	GRADE	

LOCATION	MATCHED DOMAIN	%	SOURCE TYPE
1	ijcrt.org	1	Publication
2	De-identification of electronic health record using neural network by Ahmed-2020	1	Publication
3	aclanthology.org	<1	Internet Data
4	fastercapital.com	<1	Internet Data
5	aclanthology.org	<1	Internet Data
6	www.aclweb.org	<1	Publication
7	www.engineeringletters.com	<1	Publication
8	www.oreilly.com	<1	Internet Data
9	aclanthology.org	<1	Publication
10	arxiv.org	<1	Publication
11	www.dx.doi.org	<1	Publication
12	www.mdpi.com	<1	Internet Data
13	www.dx.doi.org	<1	Publication
14	www.linkedin.com	<1	Internet Data

DECLARATION

I. We certify that the work contained in this report has been done by us under the guidance of our supervisor Prof. Vijay M B, Assistant Professor, Department of Computer Science and Engineering, JSS Science and Technology University, Sri Jayachamarajendra College of Engineering.

II. The work has not been submitted to any other Institute for any degree or diploma.

III. We have conformed to the norms and guidelines given in the Ethical Code of Conduct of the Institute.

IV. Whenever we have used materials (data, theoretical analysis, figures, and text) from other sources, we have given due credit to them by citing them in the text of the thesis and giving their details in the references. Further, we have taken permission from the copyright owners of the sources, whenever necessary.

V. The dissertation work is submitted in partial fulfillment of the requirements for the award of degree of Bachelor of Engineering in Computer Science and Engineering for the academic year 2020 — 2024.

Name:	USN:	Signature:
Name:	USN:	Signature:
Name:	USN:	Signature:
Name:	USN:	Signature:

Date:

Place:

ABSTRACT

This report describes the finds and results of a project aimed at creating an automated system for identifying and deciphering prescriptions. The project addresses the growing demand for precise and efficient prescription filling in healthcare facilities. The automated identification system seeks to improve patient safety by streamlining prescription management, reducing human mistakes, and utilizing advances in machine learning and natural language processing techniques to efficiently identify the drugs in medical prescriptions.

The report begins with an introduction on current difficulties with manually processing prescriptions, including the possibility of mistakes, delays, and misinterpretation. It underlines the need for automating this procedure to boost effectiveness and keep the patient informed about the drugs being consumed and by also suggesting alternatives.

Next, the report describes the methodology adopted to develop the automated identification system. The process involved gathering data, preprocessing it, extracting features, and training the model. The system was trained on three different datasets.

Further, the findings and the performance of all the models on various datasets are discussed. We have also curated a custom dataset using 10B tagging which is further explained in the report.

Furthermore, the report describes the potential applications and benefits of implementing the automated identification system in real-world healthcare settings. It highlights the system's potential to improve prescription accuracy, reduce human errors, enhance patient safety, and expedite prescription processing.

In conclusion, this report underscores the importance of automated identification systems in medical prescription processing. The project's findings demonstrate the feasibility and effectiveness of employing advanced machine learning techniques. This project holds immense promise in revolutionizing and benefiting healthcare professionals and patients alike.

ACKNOWLEDGEMENT

First, we are indebted to our college JSS Science and Technology University, Mysuru for providing us all the facilities needed for the successful completion of our project work.

We would like to express our sincere gratitude to all those who have contributed to the successful completion of this final year project.

We are deeply thankful to our project guide (supervisor), Prof. Vijay M B, for his invaluable guidance, expertise, and continuous support throughout the entire duration of this project. His instructive feedback, constructive criticism, and encouragement have been instrumental in shaping the direction and quality of this work.

We would also like to extend our appreciation to the faculty members of JSS Science and Technology University, for their valuable inputs, suggestions, and encouragement during the course of this project. Their expertise and knowledge have significantly enriched my skill development in the field.

To all those mentioned above and to anyone else who has directly or indirectly contributed to this project, we extend our heartfelt appreciation. Your support and assistance have been invaluable, and this project would not have been possible without you.

Thank you.

CONTENTS

LIST OF FIGURES

Page No.

LIST OF TABLES

Page No.

Chapter 1

INTRODUCTION

1.1 INTRODUCTION

Automating healthcare processes has become increasingly crucial in recent years due to technological advancements and the imperative for enhanced efficiency and patient safety. One critical aspect ripe for automation is the identification of medical prescriptions. Accurate identification of prescriptions is paramount for ensuring patient safety, minimizing errors, and optimizing healthcare workflows. However, manual identification is time-consuming, error-prone, and susceptible to variations in handwriting styles and prescription formats. Thus, developing an automated system for prescription identification holds significant promise for advancing healthcare practices.

The primary objective of this project is to design and develop an automated identification system for medical prescriptions. The system aims to precisely extract pertinent information from prescriptions, including medication names, dosages, instructions, and patient details. By automating this process, the project endeavors to enhance efficiency, reduce errors, and bolster patient safety in healthcare settings. Furthermore, it aims to standardize prescription identification practices and facilitate seamless incorporation with pre-existing electronic health record (EHR) systems, thereby enabling comprehensive patient treatment and data management.

The potential benefits of automating prescription identification are vast. By diminishing the time and effort required for manual identification, healthcare providers can devote more attention to patient care and critical tasks, consequently improving workflow efficiency and productivity. Additionally, automation can mitigate medication errors, ensuring that patients receive the correct medications and dosages, thereby augmenting patient safety and adherence.

This project endeavors to contribute to the broader field of healthcare automation by tackling the specific challenges and specifications related to automated prescription identification. Through the creation of a reliable and precise system, the project aims to enhance healthcare practices, standardize identification processes, and facilitate smooth incorporation with established healthcare systems.

In the subsequent sections of this report, we will delve into the techniques, methodologies, and considerations involved in developing the automated prescription identification system. We will explore the data collection and annotation process, the training and fine-tuning of models, evaluation metrics, and the application of the system in real-world healthcare scenarios. Additionally, we will discuss the challenges faced during the project, ethical considerations, and potential future directions for automated prescription identification.

By undertaking this project, we aim to contribute to the advancement of healthcare technology, improve patient care and safety, and pave the way for more efficient and accurate prescription identification processes in healthcare settings.

1.2 ABOUT THE PROJECT

1.2.1 Motivation

With the rapid advancement of technology, there is a growing imperative for automation across various industries, including healthcare. Automating the identification of medical prescriptions holds the potential to streamline processes, enhance accuracy, and improve patient safety. Manual identification of medical prescriptions is inherently prone to errors, but by automating this process, the likelihood of errors can be significantly diminished, ultimately leading to enhanced patient care and outcomes. Moreover, automating prescription identification can contribute to better management of medication records and patient histories. By accurately capturing and organizing prescription data, healthcare providers can easily retrieve and analyze information, leading to more informed decision-making and improved patient care.

Automated prescription identification systems can serve as valuable tools to support healthcare professionals by providing accurate and reliable information about prescriptions, including drug names, dosages, and administration instructions. These systems can assist healthcare providers in making informed decisions and reducing the likelihood of errors. Furthermore, automation in prescription identification offers the opportunity for scalability and standardization across healthcare facilities. By implementing a consistent and reliable system, the challenges associated with deciphering various handwriting styles or prescription formats can be overcome, ensuring accuracy and efficiency. In addition, the report explains the possible uses and advantages of incorporating automated identification systems in real-life healthcare scenarios.

It emphasizes the potential of the system to enhance precision in prescribing medications, minimize human errors, improve patient safety, and speed up prescription processing.

To conclude, this report highlights the significance of automated identification systems in medical prescription processing. The research findings demonstrate that the utilization of advanced machine learning techniques is both feasible and effective.

This project has the potential to revolutionize and benefit healthcare professionals and patients alike.

Considering these motivations, we have developed a compelling rationale for undertaking a project on the automated identification of medical prescriptions.

1.2.2 Problem Definition

The project aims to address the inefficiencies, errors, and potential hazards linked with manual prescription identification procedures in healthcare settings. It seeks to develop a reliable and accurate automated system capable of effectively identifying and extracting relevant information from medical prescriptions and patient details, including medication names, dosages, and instructions.

Key aspects of the problem definition include:

a. Inefficiency and Time Constraints: Manual identification of medical prescriptions is time-consuming, requiring healthcare professionals to review and decipher handwritten or complex prescriptions. This could result in delays in patient treatment and impede the productivity of healthcare professionals.

b. Error-Prone Identification: Manual prescription identification is prone to errors, such as misinterpretation of handwriting or unclear instructions. These errors can result in medication mistakes, potentially compromising patient safety and leading to adverse drug reactions.

c. Lack of Standardization: Medical prescriptions can vary in format, language, and writing style, making the identification process challenging and prone to inconsistencies. This lack of standardization further exacerbates the risk of errors and hinders efficient prescription management.

d. Patient Safety and Adherence: Errors in prescription identification can have significant ramifications for patient safety., including incorrect medication administration or dosage. Automating the identification process can aid in mitigating these risks, ensuring patients receive the correct medications and dosages.

e. Complex Prescription Structures: Medical prescriptions can contain complex structures, including abbreviations, dosage instructions, and medication frequency. Manual identification of these structures can be challenging, requiring domain expertise and increasing the likelihood of errors.

f. Scalability and Consistency: In larger healthcare settings or across multiple healthcare facilities, maintaining consistent and accurate prescription identification becomes increasingly difficult. Automating the identification process can provide scalability and standardization, ensuring consistent and reliable identification regardless of the volume or variety of prescriptions.

By defining these problem areas, the project aims to create a solution that overcomes the shortcomings of manual prescription identification, enhances efficiency, reduces errors, improves patient safety, and integrates seamlessly with existing healthcare systems.

1.2.3 Challenges

a. Complex Prescription Formats: Medical prescriptions can have complex structures and formats, including abbreviations, symbols, dosage instructions, and frequency details. Deciphering and extracting the relevant information accurately from these complex structures pose a significant challenge for automated systems.

b. Handling Prescription Abbreviations: Medical prescriptions often utilize various abbreviations specific to the healthcare domain. These abbreviations can be ambiguous and context-dependent, requiring sophisticated algorithms to accurately interpret their intended meaning and context.

c. Limited Training Data and Domain-Specific Challenges: Obtaining a sufficient amount of accurately annotated training data for prescription identification can be challenging, especially for specialized domains or rare medical conditions. Building robust and accurate models might necessitate substantial data gathering efforts and expertise in domain-specific challenges.

d. Ongoing Model Updates and Maintenance: Medical knowledge, drug information, and prescription practices continuously evolve. Therefore, maintaining and updating the automated identification systems to keep up with new medications, dosage guidelines, and emerging practices is essential. Regular model updates and system maintenance should be planned to ensure the system remains accurate and up to date.

e. Indian Dataset for Medicine: Lack of an annotated dataset for Indian medicine names. The absence of localized data can raise concerns about the automated identification of local medicine prescriptions.

Addressing these challenges requires a combination of advanced machine learning and natural language processing techniques, domain expertise, and extensive data collection. Overcoming these obstacles will lead to the development of robust and reliable automated identification systems for medical prescriptions, ultimately improving patient safety and healthcare efficiency.

1.2.4 Objectives

- Developing an Accurate and Reliable Prescription Identification System: The principal objective of this project is designing and creating an automated mechanism capable of accurately identifying and extracting pertinent information from medical prescriptions. The system should demonstrate exceptional accuracy in deciphering prescription details, encompassing medication names, dosages, instructions, and patient information.
- Enhancing Patient Safety and Reducing Medication Errors: Another critical aim of the project is to bolster patient safety by minimizing drug mistakes linked to manual prescription identification. By precisely identifying prescriptions and ensuring the accuracy of medications and dosages, the system can mitigate medication mistakes, adverse drug reactions, and potential harm to patients.
- Exploring Potential for Scalability and Generalizability: This project seeks to explore the scalability and generalizability of the automated prescription identification system. It aims to evaluate the system's performance across diverse healthcare settings, encompassing hospitals, clinics, and pharmacies, with varying patient demographics, prescription patterns, and languages.
- Continuously Improving and Updating the System: An essential objective is to establish a framework for continuous improvement and system updates. The

project endeavors to monitor system performance, gather feedback from users, and integrate new data and knowledge to enhance the accuracy, reliability, and adaptability of the prescription identification system over time.

- Ensuring Compliance with Privacy and Ethical Guidelines: This project prioritizes patient privacy, data security, and compliance with ethical guidelines throughout the development and implementation of the automated prescription identification system. It aims to institute robust data governance practices, adhere to pertinent regulations, and ensure the ethical utilization of patient information.

By addressing these objectives, the project aims to advance healthcare technology, enhance patient care and safety, optimize workflow efficiency, and establish standardized practices in the identification of medical prescriptions.

1.3 EXISTING SOLUTION

In current medical prescription identification systems, notable areas require enhancement. There's a need for improved capabilities in handling diverse handwriting styles for accurate Optical Character Recognition (OCR). Deeper semantic understanding is necessary for interpreting contextual nuances within prescriptions, improving accuracy in medication details and patient information extraction. Universality in adapting to various prescription layouts and structures is crucial. Real-time processing capabilities need improvement for immediate decision-making in healthcare settings. Smooth incorporation with EHR systems must be ensured for efficient data exchange. Robust measures for patient privacy and data security are pivotal. User-friendly interfaces should be enhanced for better user experience. Systems should adeptly handle prescriptions in multiple languages. Improving machine learning model generalization across healthcare contexts is essential. Continuous learning and adaptation capabilities are crucial for staying current with evolving prescription patterns and healthcare practices. Addressing these aspects will develop more robust, adaptable, and user-friendly medical prescription identification systems.

1.4 PROPOSED SOLUTION

The proposed automated prescription identification system aims to surpass current limitations by integrating advanced technologies to achieve more accurate and efficient medical prescription processing. By leveraging Optical Character Recognition (OCR) with enhanced capabilities, the system ensures precise text extraction from diverse handwriting styles. Additionally, a focus on deep semantic understanding through Natural Language Processing (NLP) enhances the accuracy of medication details and patient information extraction.

The system boasts a flexible architecture designed to universally adapt to various prescription layouts and structures, facilitating real-time processing for immediate identification and decision-making in healthcare settings. Moreover, smooth incorporation with EHR systems ensures efficient data exchange, while robust privacy measures uphold data security standards.

With a user-friendly interface, the system enhances usability for healthcare professionals and accommodates prescriptions in multiple languages. Machine learning models are fine-tuned to improve generalization, ensuring adaptability to diverse healthcare contexts. Crucially, the system is designed for continuous learning and adaptation to evolving prescription patterns and changes in healthcare practices, ensuring sustained relevance and effectiveness.

Overall, the proposed system aims to establish a new standard by addressing existing limitations and harnessing innovative technologies to enhance precision and efficiency in medical prescription processing.

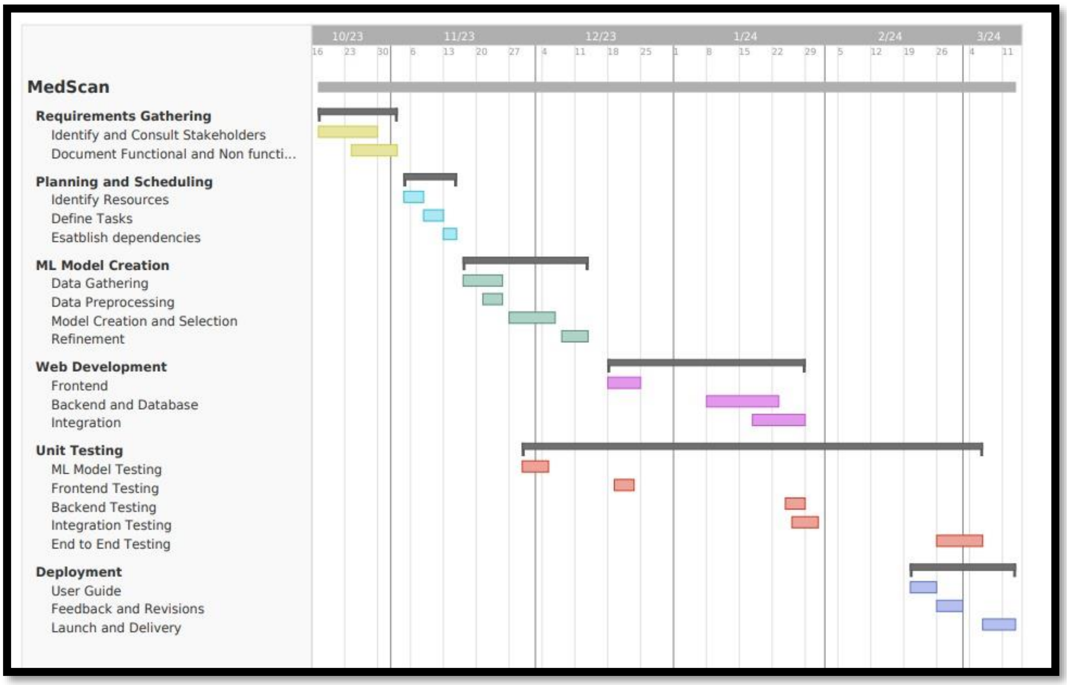
1.5 GANTT CHART

Phase 1: Requirements Gathering and Planning (Sep 2023-Nov 2023): In this initial phase, the project team meticulously gathered and analyzed requirements from key stakeholders, outlining the intricacies of the automated prescription identification system. An elaborate project blueprint was formulated, encompassing timelines, milestones, and resource allocation. The team identified essential technologies and resources vital for the project's success, setting a robust foundation for subsequent phases.

Phase 2: Design and Development (Nov 2023-Feb 2024): Building on the gathered requirements, the project entered the designing and development stage. The focus was on crafting an intuitive and efficient user interface, laying the groundwork for the system's core functionalities. Development activities concentrated on implementing prescription extraction mechanisms, enhancing categorization features, and establishing a secure backend for efficient data management. The system's architecture and database structures were carefully designed to ensure scalability and security.

Phase 3: Testing and Refinement (Feb 2024): Rigorous testing and quality assurance processes took center stage in this phase. The application underwent comprehensive assessments to validate functionality, usability, and security. User feedback played a pivotal role in refining features and optimizing the overall user experience. This iterative process ensured the creation of a polished and seamless automated prescription identification system.

Phase 4: Deployment and Launch (Feb 2024 - x 2024): The automated prescription identification system reached its culmination with deployment to a production environment, making it accessible to end-users. Concurrently, the project team crafted comprehensive documentation and user guides to facilitate user adoption. Training materials were prepared to empower users with the knowledge needed to effectively utilize the system, ensuring a successful launch and seamless integration into healthcare workflows.



Chapter 2

LITERATURE REVIEW

2.1 Medical Prescription Classification: An NLP Based Approach

Authors: Viincenza Carchiolo; Alessandro Longheu; Giuseppa Reitano; Luca Zagarella
Year of Publication: 2019

The digitization of healthcare data has become essential in recent years to manage the vast amount of information generated by healthcare organizations. In addition to enabling a number of related applications, including clinical text mining, predictive modeling, survival analysis, patient similarity evaluation, and genetic data analysis, this approach is an essential resource for enhancing the delivery of healthcare services. In order to make the process of obtaining healthcare permission and medical expense reimbursement easier, this project focuses on the digitization of medical prescriptions. Using Natural Language Processing (NLP) and machine learning techniques, the proposed system first extracts text from scanned medical prescriptions and then uses that text to classify the prescriptions based on embedded terms and categories related to patient/doctor personal data, symptoms, pathology, diagnosis, and recommended treatments. After identifying the kind and format of the medical prescription by analyzing the input image, pertinent strings are isolated using text extraction and rectification. After then, these strings are categorized according to previously gathered data, which establishes whether the prescription is eligible for additional approval. The suggested system's first module focuses on differentiating between various kinds of prescription drugs. Text extraction is done after the image is pre-processed, and the result is a string dictionary with each item representing a field of the medical prescription. Subsequently, spelling correction is applied to address residual errors commonly found in OCR software, particularly in scenarios with low-quality scanned images or reduced font sizes.

The goal of information classification is to determine the grantability of a given prescription. This process utilizes syntactic rules and rule-based tagging NLP techniques. Syntactic rules model valid grammar sequences, while rule-based approaches use contextual information to assign tags to unknown or ambiguous words. These rule-based triggers may require supervised training, although different methods are available.

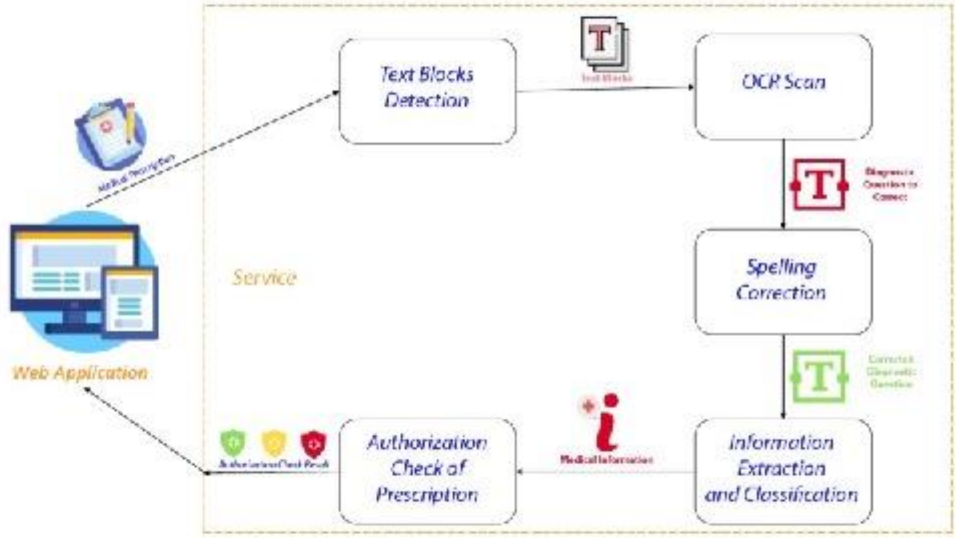
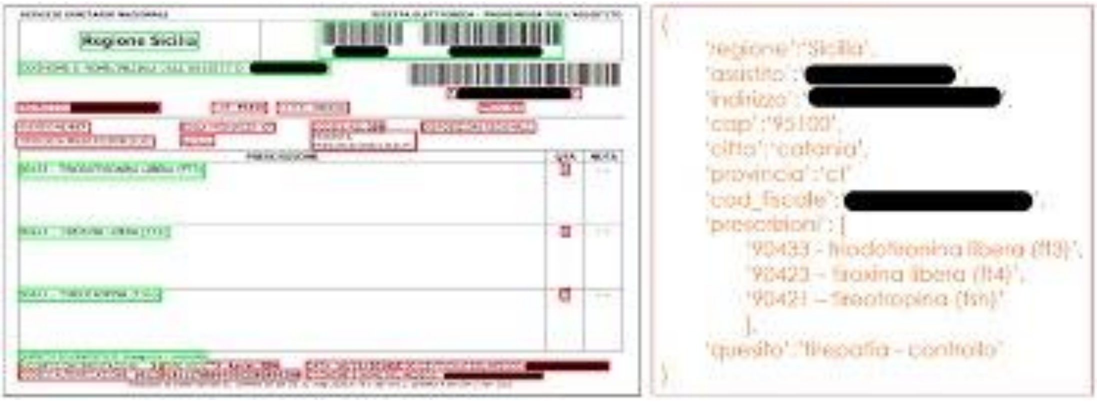

Text Blocks Detection
Text Blocks
OCR Scan
Diagnostic Question to Correct
Spelling Correction
Corrected Diagnostic Question
Information Extraction and Classification
Medical Information
Authorization Check of Prescription
Authorization Check Result
Service
Medical Prescription
Web Application

Figure 2.1.1



```
{
    'regione':'Sicilia',
    'assistito': '[redacted]',
    'indirizzo': '[redacted]',
    'cap':'95100',
    'citta':'catania',
    'provincia':'ct'
    'cod_fiscale': '[redacted]',
    'prescrizioni' : [
        '90433 - triiodotironina libera (ft3)',
        '90423 – tiroxina libera (ft4)',
        '90421 – tireotropina (tsh)'
        ],
    'quesito':'tireopatia - controllo'
}
```

Figure 2.1.2

Finally, during the testing phase, the performance of spelling correction and information classification algorithms is evaluated to establish whether an input medical prescription can be classified as grantable. Results demonstrate that in most cases, the system allows automatic classification of text and prescribed medicine, with errors occurring at a rate of 5%.

The principal goal of the proposed system is to create a service for the analysis and authorization of medical prescriptions. Results suggest that the system can automatically classify text and prescribed medicine in most cases, with a low error rate of 5%.

Existing Knowledge: The 2019 paper focuses on classifying medical prescriptions using NLP. It involves error correction and classification based on syntactic rules.

Gap: While the paper addresses prescription classification, the project needs to explore the specific challenges related to medical prescription extraction and identification, including the intricacies of medication details and patient information.

2.2 Efficient Estimation of Word Representations in Vector Space

Authors: Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean

Year of Publication: 2013

The technique for learning word embeddings—word representations in a high-dimensional vector space—is presented in this study. For learning these representations, Continuous Bag-of-Words (CBOW) and Skip-Gram, two neural network-based models, are suggested. Whereas the Skip-Gram model predicts the context given a target word, the CBOW model predicts the target word given its context. Using negative sampling, which is a computationally efficient method in contrast to others like hierarchical SoftMax, both models are trained on a sizable corpus of text. The authors show that their trained word representations outperform earlier state-of-the-art models by evaluating their quality on a range of natural language processing (NLP) tasks, including as language modeling, part-of-speech tagging, and named entity recognition.

The study also highlights the benefits of word embeddings, including their versatility in a variety of NLP applications and their capacity to capture syntactic and semantic links between words. In the field of NLP, this technique has gained widespread acceptance and is now considered standard practice.

In a continuous vector space, word embeddings encode words as high-dimensional vectors with the goal of encapsulating their meanings for use as input in machine learning models, especially neural networks. Words are mapped into dense vectors, with different words far away and semantically equivalent words close together. For instance, because of their different semantic connotations, "cat" and "feline" would be close, but "cat" and "car" would be far apart. In a similar vein, terms like "lion" and "king" would be far apart, but "queen," "prince," and "princess" would be close.

Applications for word embeddings can be found in machine translation, sentiment analysis, named entity recognition, part-of-speech tagging, language modeling, and other NLP tasks. They are used as input for tasks like text classification, question answering, and text synthesis in neural networks and other machine learning models.

Furthermore, word embeddings are useful for pre-training models in transfer learning for natural language processing tasks; popular methods like word2vec, GloVe, and BERT are easily accessible for use in the field.

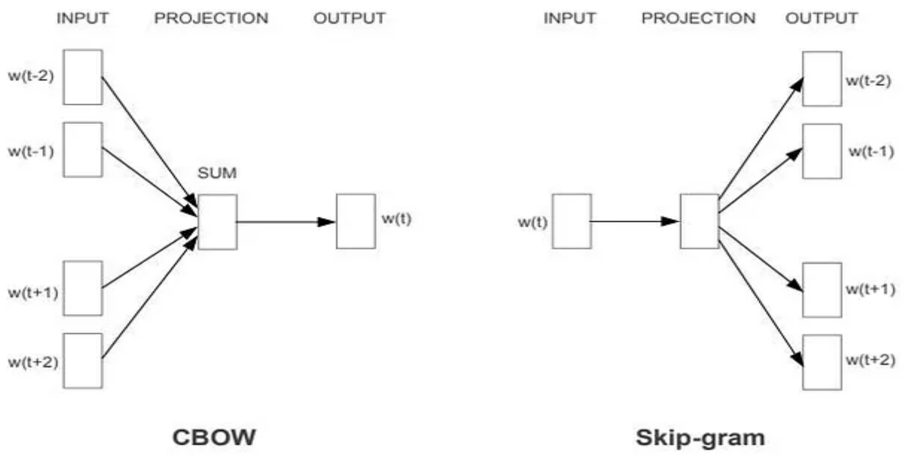


Figure 1: New model architectures. The CBOW architecture predicts the current word based on the context, and the Skip-gram predicts surrounding words given the current word.

Figure 2.2.1

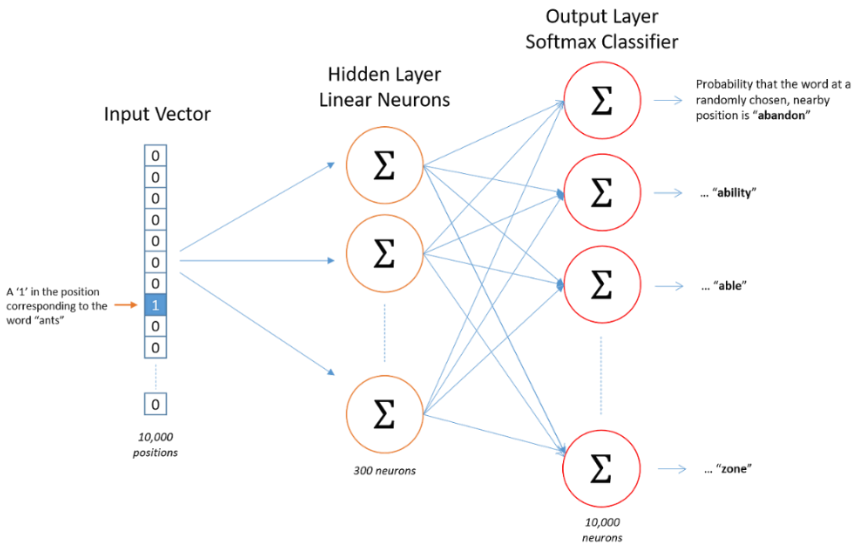
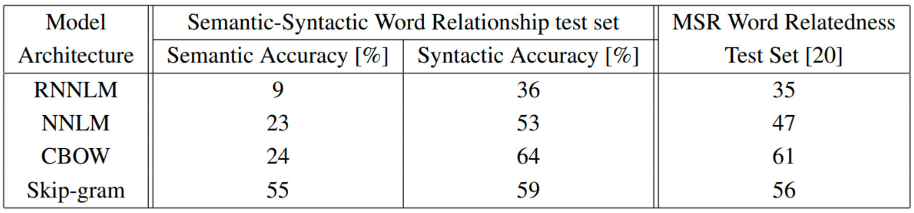


Figure 2.2.2

Table 3: *Comparison of architectures using models trained on the same data, with 640-dimensional word vectors. The accuracies are reported on our Semantic-Syntactic Word Relationship test set, and on the syntactic relationship test set of [20]*



Model	Semantic-Syntactic Word Relationship test set		MSR Word Relatedness
Architecture	Semantic Accuracy [%]	Syntactic Accuracy [%]	Test Set [20]
RNNLM	9	36	35
NNLM	23	53	47
CBOW	24	64	61
Skip-gram	55	59	56

Table 2.2.1

Existing Knowledge: The 2013 paper introduces efficient techniques for learning word embeddings, essential for Natural Language Processing tasks such as language modeling and part-of-speech tagging.

Gap: The project could benefit from incorporating and adapting word embedding techniques to capture the semantics of medical terms and improve the understanding of prescription content.

2.3 Recent Trends in Named Entity Recognition (NER)

Author: Arya Roy

Year of Publication:2021

A critical component of natural language processing (NLP) is named entity recognition (NER), which finds and classifies named entities in text, such as individuals, groups, places, and dates. Several essential steps are usually involved in the NER process:
1. Tokenization: Text is divided into discrete words, or tokens, which are the fundamental building blocks that NER systems use to function.
2. Part-of-speech (POS) tagging: This helps distinguish words with different meanings by identifying each token's grammatical function inside a phrase.
3. Chunking: To make it easier to identify named entities made up of several words, words are grouped into meaningful chunks or noun phrases.
4. Named Entity Recognition: Using a variety of techniques, such as rule-based systems, statistical models, and machine learning-based methods, the system recognizes and classifies named entities.
5. Evaluation: The system's performance is assessed using labeled test data, typically measured by metrics like the F1 score, which balances precision and recall.

A number of approaches can be used to put NER into practice:
1. Rule-based systems: To recognize named things, these systems use established rules based on text patterns such as affixes and capitalization. Even though they are accurate, their applicability to new languages or domains is difficult and their rule coverage is limited.
2. Statistical models: To categorize named things, models such as Conditional Random Fields (CRFs) and Hidden Markov Models (HMMs) use the statistical characteristics of text and context, which can effectively grasp word context after being trained on

labelleddata.

3. Systems based on machine learning: These systems use methods like neural networks, decision trees, and support vector machines (SVMs) to identify patterns in text that correspond to identified things. They train on large amounts of labeled data, which leads to their outstanding performance.

4. Hybrid systems: These systems use a combination of machine learning and rule-based techniques to categorize named things using machine learning algorithms after preprocessing text with rules. By utilizing the advantages of both approaches, they provide accuracy and robustness.

5. Deep Learning-based systems: These systems learn rich representations of words and context to precisely identify and categorize named entities. They do this by utilizing neural networks and its derivatives, such as LSTM, GRUs, and transformer-based models, like BERT. They are excellent at encapsulating intricate relationships in words. Each methodology has its strengths and weaknesses, making the choice dependent on factors like task requirements, available resources, and desired performance levels.

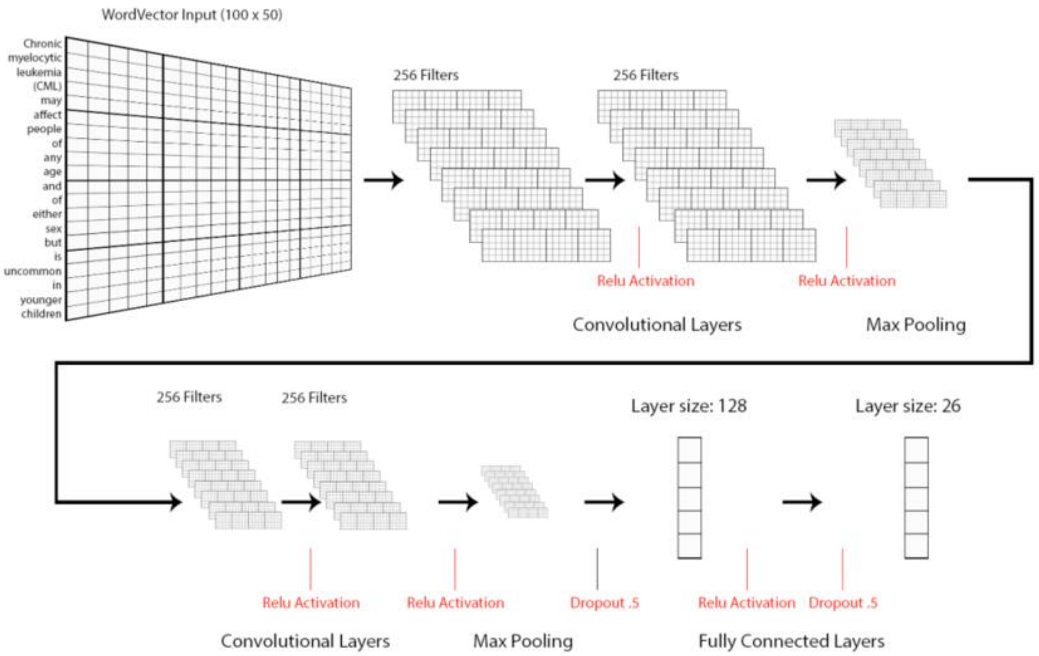


Figure 2.3.1

Existing Knowledge: The 2021 paper discusses various NER methods, including rule-based, statistical, machine learning-based, hybrid, and deep learning-based systems, with a focus on entities in text.

Gap: While the paper provides insights into NER methods, the project needs to specifically explore NER challenges and techniques relevant to medical prescriptions, considering the unique entities and structures within healthcare documents.

2.4 A Neural Attention Model for Sentence Summarization

Authors: Alexander M. Rush, Sumit Chopra, Jason Weston

Year of Publication: 2015

This article introduces a novel neural attention-based model designed specifically for summarizing sentences. The authors propose a unique combination of neural networks and attention mechanisms to condense text while preserving its core concepts.

Summarization poses a significant challenge in natural language understanding, aiming to distill the essential meaning of a text into a shorter form. While many existing summarization systems rely on extractive methods, which involve selecting and combining segments of the text, abstractive summarization seeks to generate summaries that may contain new information not explicitly present in the original text.

The model presented in this paper follows an encoder-decoder architecture. The encoder transforms the input sentence into a continuous representation, while the decoder generates the summary by attending to the encoded input. The attention mechanism allows the model to focus on specific parts of the input sentence, enhancing the informativeness of the generated summary.

Key components of the model include:

Encoder: This employs a bidirectional recurrent neural network (RNN) to encode the input text into a fixed-length vector representation, considering context in both forward and backward directions.

Attention Mechanism: Using a "global attention" mechanism, this assigns weights to each word in the input text, guiding the decoder in selecting words for summary generation.

Decoder: This generates the summary based on the encoded input text and attention weights, determining which words to include and their order.

The authors evaluate their model on various datasets, including Gigaword and DUC, demonstrating its superiority over state-of-the-art models. They emphasize the model's ability to handle both long and short sentences effectively.

The attention mechanism emerges as a crucial aspect of the model, enabling it to focus on informative parts of the input sentence and produce coherent and accurate summaries. This approach outperforms traditional methods that rely on heuristics or rule-based approaches.

In conclusion, the paper highlights the model's achievements in sentence summarization and underscores the significance of the attention mechanism in improving summary quality through focused information extraction.

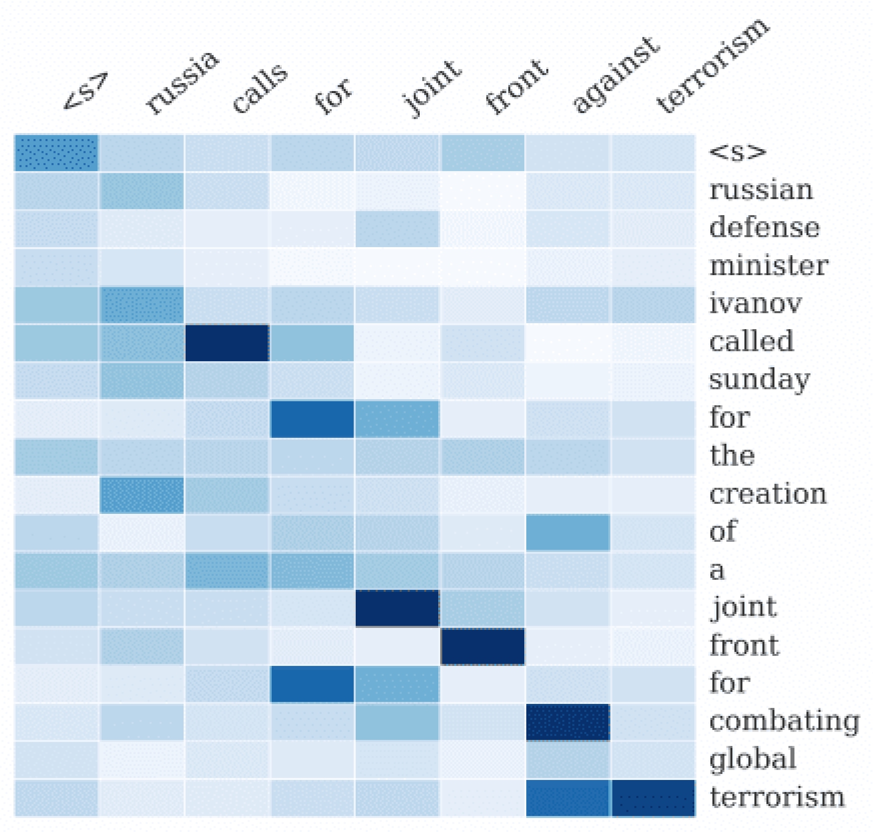


Figure 2.4.1

Existing Knowledge: The 2015 paper introduces a neural attention model for sentence summarization, demonstrating its effectiveness on various datasets.

Gap: The project could explore the application of attention mechanisms to extract critical information from prescriptions, enabling the system to focus on key details and enhance the accuracy of information extraction.

2.5 Attention is all you need

Authors: Ashish Vaswani

Year of Publication: 2017

Recurrent neural networks (RNNs), long short-term memory networks (LSTMs), and gated RNNs are commonly employed in Sequence Modelling tasks such as machine translation and language modeling. However, these RNN-based models process sequences sequentially, which limits parallelization and creates challenges with capturing long-range dependencies.

Recent advancements have made progress in improving computational efficiency and model performance, but they have not fully resolved the issue of sequential computation bottleneck. Attention mechanisms offer a solution by enabling dependency modeling without being constrained by sequence distance, making them crucial for sequence modeling and transduction tasks. However, attention mechanisms are typically integrated with recurrent networks.

This paper introduces the Transformer, a model architecture entirely based on attention mechanisms for capturing global dependencies between input and output. The Transformer enables extensive parallelization and significantly enhances translation quality, achieving noteworthy results with just twelve hours of training on eight GPUs. In the Transformer, both the encoder and decoder employ stacked self-attention and pointwise, fully connected layers. Self-attention is chosen due to its lower computational complexity per layer, increased parallelizability, and shorter path lengths between long-range dependencies compared to recurrent layers.

The Transformer utilizes two types of attention functions: Scaled Dot-Product Attention, which computes attention on multiple queries simultaneously, and multi-head attention, allowing the model to attend to different representation subspaces jointly.

A key advantage of the Transformer is its ability to connect all positions in constant time, unlike the linear complexity of recurrent layers. This makes self-attention layers faster for shorter sequences, which is common in machine translation tasks.

Although the Transformer has mainly been applied to transduction models, the authors intend to extend its application to tasks involving non-text modalities such as images, audio, and video in the future. Since its introduction, the Transformer has attracted significant attention due to its remarkable enhancements in translation quality and its versatility in various NLP tasks.

Existing Knowledge: The 2017 paper introduces the Transformer architecture, which relies entirely on attention mechanisms for capturing global dependencies between input and output, significantly improving translation quality, and enabling extensive parallelization.

Gap: The project could investigate the adaptation of the Transformer architecture to the task of medical prescription classification, leveraging attention mechanisms to capture intricate dependencies within prescriptions and improve classification accuracy.

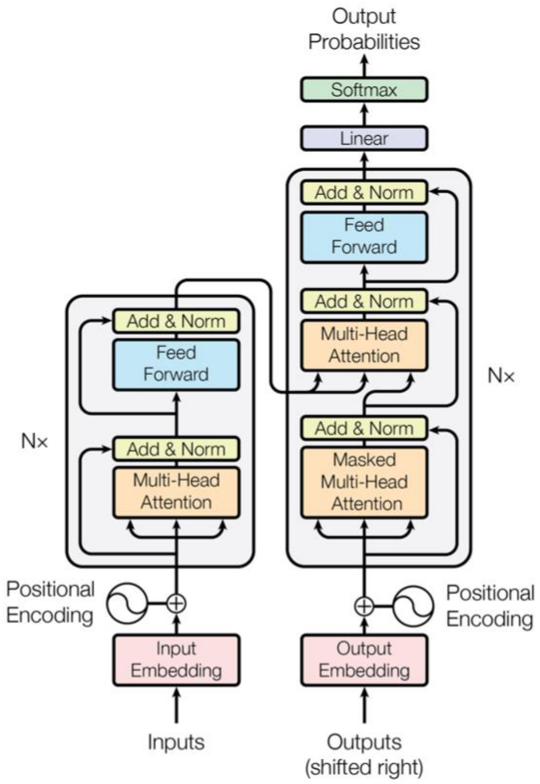


Figure 2.5.1

2.6 BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Authors: Jacob Devlin, Ming-Wei Cheng, Kenton Lee, Kristina Toutanova

Year of Publication: 2018

The BERT (Bidirectional Encoder Representations from Transformers) paper, introduced by Google AI researchers, represents a significant advancement in natural language processing (NLP). BERT utilizes a bidirectional Transformer model to generate word representations that capture extensive contextual information. Unlike previous approaches that relied on unidirectional models, BERT considers both preceding and succeeding words, enhancing its understanding of sentence meanings and connections. The model undergoes pre-training on large text datasets, employing masked language modeling and next sentence prediction tasks.

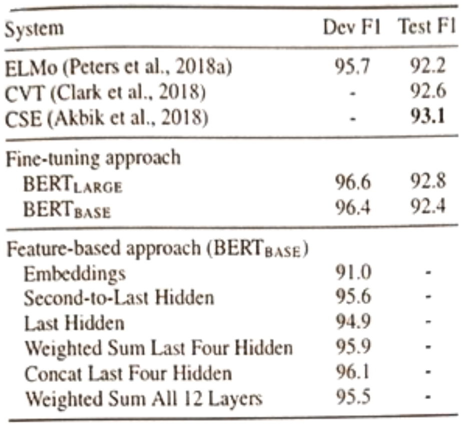
BERT's key innovation lies in its pre-training and fine-tuning phases. During pre-training, the model learns to predict masked words within sentences and whether two sentences are consecutive in a document, fostering a deep understanding of syntax, semantics, and discourse structure. Following pre-training, BERT is fine-tuned on specific tasks using labeled data, making it adaptable and powerful across various NLP tasks.

BERT demonstrates remarkable performance across a wide range of NLP tasks, surpassing previous state-of-the-art models. Its success stems from its ability to capture both local and global contextual cues, enabling precise predictions and comprehension of complex sentence structures. The model's bidirectional nature and attention mechanism facilitate effective handling of long-range dependencies, enhancing word representations in context.

Beyond task-specific performance, BERT has spurred advancements in NLP research, particularly in transfer learning and contextual word representations. Its pre-training and fine-tuning methodology have inspired the development of numerous models that achieve state-of-the-art results across diverse applications. Additionally, the release of the BERT paper has spurred the creation of large-scale pre-trained language models, driving significant progress in NLP research and applications.

BERT's influence extends across academia and industry, stimulating innovation and unlocking new possibilities for natural language understanding and generation.

Gap: Despite BERT's success in various NLP tasks, there is a need to explore its application in specific healthcare-related tasks, such as medical text classification or entity recognition, to evaluate its effectiveness in domain-specific contexts and potentially enhance healthcare-related NLP applications.



System	Dev F1	Test F1
ELMo (Peters et al., 2018a)	95.7	92.2
CVT (Clark et al., 2018)	-	92.6
CSE (Akbik et al., 2018)	-	**93.1**
Fine-tuning approach		
BERT$_{\text{LARGE}}$	96.6	92.8
BERT$_{\text{BASE}}$	96.4	92.4
Feature-based approach (BERT$_{\text{BASE}}$)		
Embeddings	91.0	-
Second-to-Last Hidden	95.6	-
Last Hidden	94.9	-
Weighted Sum Last Four Hidden	95.9	-
Concat Last Four Hidden	96.1	-
Weighted Sum All 12 Layers	95.5	-

Table 2.6.1

Chapter 3

REQUIREMENT SPECIFICATION

3.1 FUNCTIONAL REQUIREMENTS

- Data Input and Preprocessing
    - Accept input as a medical prescription image.
    - Preprocess input data to remove stop words, common words and whitespaces.
- Custom Dataset Generation
    - Creation of a small custom annotated dataset with Indian drug names.
- User Interface
    - User friendly interface where the user can select the trained model and upload the image of the prescription.
    - After the image is processed and the drug names are identified, it should give an output with a hyperlink to more knowledge on the drug.
- Implement a custom Named Entity Recognition (NER) model to identity medical terms and medicines from the input data.

3.2 NON-FUNCTIONAL REQUIREMENTS

- System should be highly accurate in identifying medical terms and chemicals with minimal false positives/negatives.
- System's response time should be minimum for processing input and generating output images.
- System should be reliable and available for use 24/7 with minimal downtime Ensure data privacy and security by handling medical prescription data in compliance with relevant regulations (e.g., HIPAA).
- System should be scalable to handle increasing volumes of medical prescriptions as the application grows.

Chapter 4

TOOLS AND TECHNOLOGIES USED

4.1 HARDWARE REQUIREMENTS

- Minimum CPU: Intel Core i5 or equivalent
- Minimum RAM: 8 GB
- Minimum storage: 256 GB SSD
- Recommended GPU: NVIDIA GeForce GTX 1060 or equivalent for
- accelerated NLP processing
- Internet connectivity with adequate bandwidth for retrieving and processing data

4.2 SOFTWARE REQUIREMENTS

- Operating System: Windows 10 or Linux Ubuntu 18.04
- Python 3.7 or higher for software development and deployment
- Libraries and APIs: NLTK, SpaCy, TensorFlow, Keras for NLP processing,
- and image generation
- Integrated Development Environment (IDE): PyChann or equivalent
- Version Control System: Git for code versioning and collaboration

Chapter 5

SYSTEM DESIGN

5.1 WORKFLOW DIAGRAM

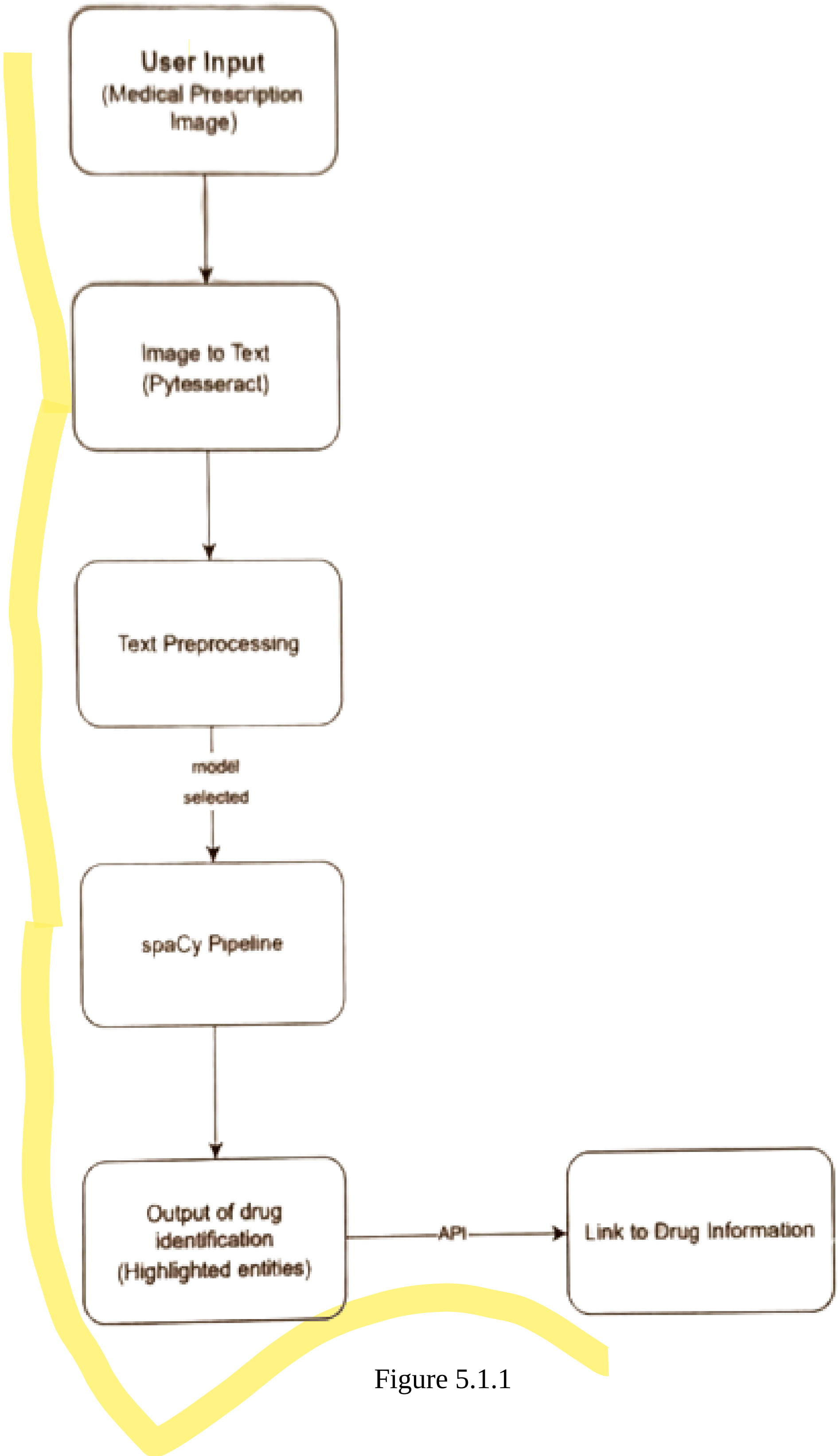


Figure 5.1.1

1. User gives the input - a medical prescription image.

2. We pass the image through Pytesseract to perform OCR and convert the image to text.

3. Text preprocessing is performed where all the stop words, whitespaces and unwanted text is removed.

4. Pre-processed text is taken as input to the selected model and passed through the spacy pipeline.

5. Output is generated where the drug names are the highlighted entities.

6. Identified drug entities are taken as input to an API to fetch drug related information from drugs.com.

5.2 SEQUENCE DIAGRAM

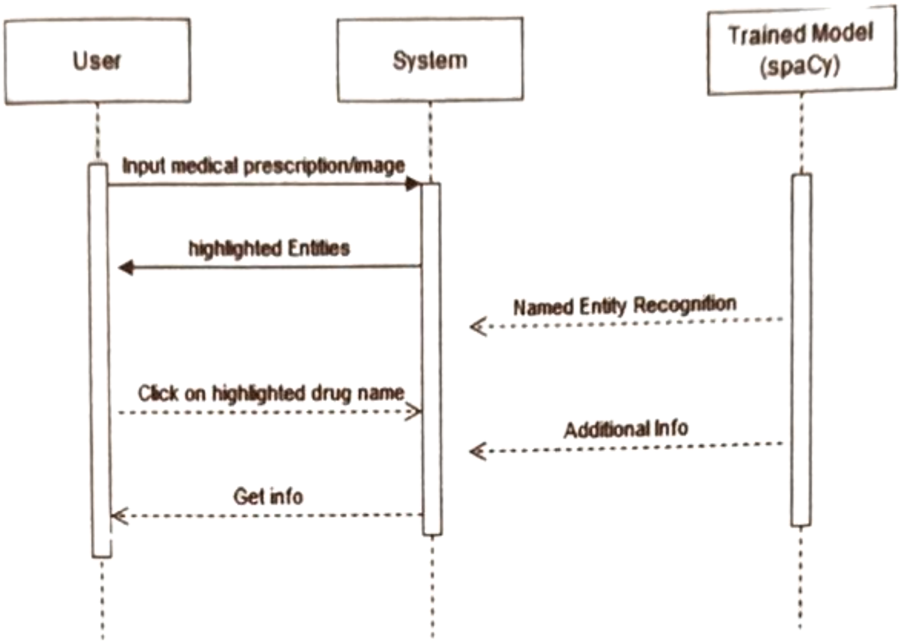


Figure 5.2.1

5.3 SPACY ARCHITECTURE

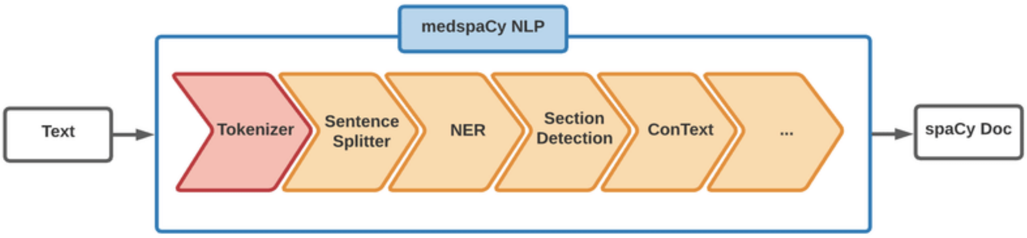


Figure 5.3.1

Chapter 6

SYSTEM IMPLEMENTATION

6.1 METHODOLOGY

6.1.1 Data Collection

The initial step in our project is to gather data from various sources. The data should be diverse and representative of the domain we are targeting.

1. **Data Sources:** Identify likely sources from which prescription data can be collected. We have identified three suitable datasets for this project.

BC5CDR Dataset:

- The BC5CDR dataset is a manually annotated corpus designed for tasks involving biomedical named entity recognition (NER) and relation extraction. It comprises PubMed abstracts and full-text articles that have undergone annotation to identify mentions of chemicals and diseases, along with their relationships.
- This dataset is divided into a training set, consisting of 1,500 PubMed abstracts, and a test set, comprising 500 PubMed abstracts. Each entry in the corpus is annotated by at least two annotators to ensure accuracy and reliability.
- Annotations in the BC5CDR corpus are structured in the standoff format, where each entity and relation is assigned a unique identifier and stored separately from the original text. This format facilitates comparison between annotations and enables evaluation of machine learning models on the dataset.
- The BC5CDR corpus serves as a benchmark for evaluating various machine learning models in the biomedical NER and relation extraction domains. It is widely utilized by researchers in the biomedical NLP community, contributing to advancements in this field.
- The en_ner_bc5cdr_md model is a pre-trained spacy model specifically designed for biomedical NER tasks using the BC5CDR dataset. Trained to recognize entities pertaining to chemicals and diseases in biomedical text, this model aids in automating entity recognition processes.

Here are some key details about the BC5CDR corpus:

1. The corpus comprises 2,000 PubMed abstracts, divided into a training set of 1,500 abstracts and a test set of 500 abstracts.

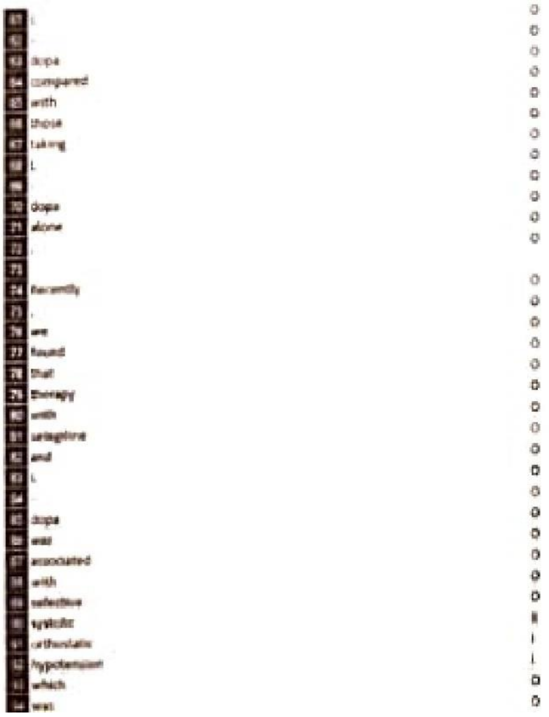
2. Each abstract has undergone annotation by at least two annotators, with annotations cross-verified for accuracy.

3. Annotations cover mentions of chemical and disease entities, as well as their relations.

4. Annotations are stored in the standoff format, assigning unique identifiers to entities and relations, separate from the original text.

5. The corpus serves as a benchmark for evaluating machine learning models in biomedical named entity recognition and relation extraction tasks.

It's essential to acknowledge that machine learning model performance on the BC5CDR corpus can vary based on factors like task specificity, data quantity and quality, and evaluation criteria.



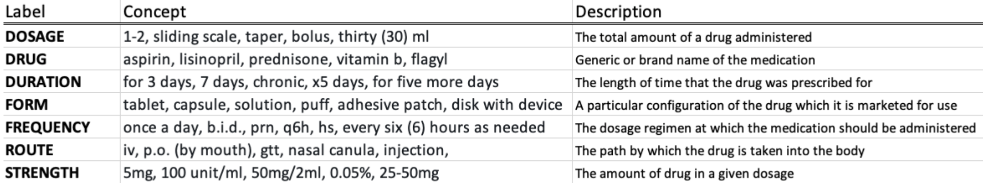
```
61 L            O
62 -            O
63 dopa         O
64 compared     O
65 with         O
66 those        O
67 taking       O
68 L            O
69 -            O
70 dopa         O
71 alone        O
72 .            O
73
74 Recently     O
75 ,            O
76 we           O
77 found        O
78 that         O
79 therapy      O
80 with         O
81 selegiline   O
82 and          O
83 L            O
84 -            O
85 dopa         O
86 was          O
87 associated   O
88 with         O
89 selective    O
90 systolic     B
91 orthostatic  I
92 hypotension  I
93 which        O
94 was          O
```

Figure 6.1.1

Med7 Dataset:

The Med7 dataset is a corpus of annotated clinical text with named entities related to medication use. Below are some key features of the Med7 dataset:

- The Med7 dataset contains a total of 5,000 clinical notes, split into a training set (4,000 notes) and a test set (1,000 notes).
- The clinical notes are derived from the MIMIC-III database, which is a large, publicly available database of de-identified electronic health records.
- The annotations in the Med7 dataset include seven types of medication-related named entities: medication name, potency, and form of dosage, route of administration, frequency, duration, and reason for use.
- The Med7 dataset has seen extensive utilization for developing and evaluating machine learning models for medication-related named entity recognition(NER) in clinical text.
- ML model performance on the Med7 dataset varies depending on the specific model architecture, training data, and evaluation metrics used.
- In a recent study, a transformer-based model achieved cutting edge performance on the Med7 dataset, achieving an F1 score of 0.903 on the test set.
- The Med7 dataset was introduced in a 2018 paper by Dernoncourt et al., titled "Training and Evaluating a Named Entity Recognizer for Electronic Health Records".
- The dataset is available in two formats: a raw text format, which includes the original clinical notes and corresponding annotations, and a tokenized format, which includes the text and annotations in a standardized format appropriate for utilization with machine learning algorithms.
- The Med7 dataset has been utilized in several exploratory studies on different aspects of medication-related named entity recognition, including using different types of neural network architectures, the effect of pretraining on large corpora, and the benefit of active learning to improve model performance.



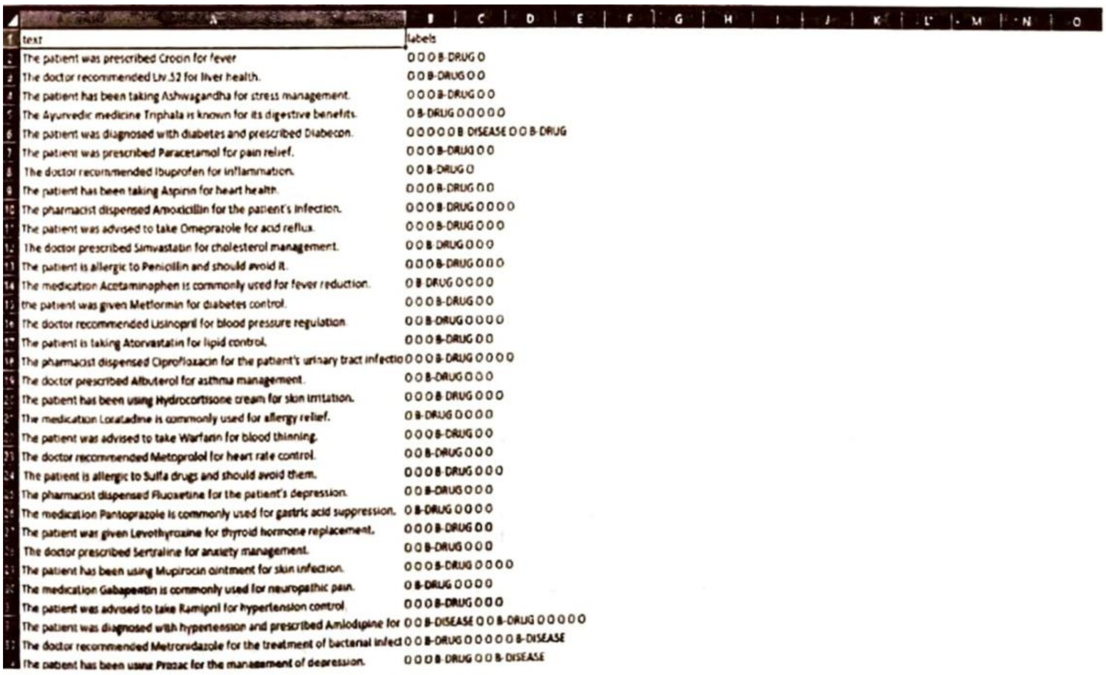
Label	Concept	Description
DOSAGE	1-2, sliding scale, taper, bolus, thirty (30) ml	The total amount of a drug administered
DRUG	aspirin, lisinopril, prednisone, vitamin b, flagyl	Generic or brand name of the medication
DURATION	for 3 days, 7 days, chronic, x5 days, for five more days	The length of time that the drug was prescribed for
FORM	tablet, capsule, solution, puff, adhesive patch, disk with device	A particular configuration of the drug which it is marketed for use
FREQUENCY	once a day, b.i.d., prn, q6h, hs, every six (6) hours as needed	The dosage regimen at which the medication should be administered
ROUTE	iv, p.o. (by mouth), gtt, nasal canula, injection,	The path by which the drug is taken into the body
STRENGTH	5mg, 100 unit/ml, 50mg/2ml, 0.05%, 25-50mg	The amount of drug in a given dosage

Figure 6.1.2

Custom Dataset:

The Custom Indian dataset is a collection of data obtained by scraping websites specific to the Indian context. Here are some key features of the dataset:

- The Custom Indian Web Scraping dataset focuses on information gathered from diverse online sources in India. It aims to provide valuable information and insights related to a specific domain or research area relevant to the Indian context.
- The dataset is collected by utilizing web scraping techniques to extract relevant data from websites in India. The data collection process follows ethical considerations, respects website terms of service, and adheres to data protection regulations.
- The dataset captures information related to healthcare providers in India, including doctors, clinics, hospitals, and pharmacies. This aspect helps in analyzing prescription patterns, healthcare access, and provider-specific insights within the Indian healthcare landscape.



text	labels
The patient was prescribed Crocin for fever	O O O B-DRUG O
The doctor recommended Liv.52 for liver health.	O O B-DRUG O O
The patient has been taking Ashwagandha for stress management.	O O O B-DRUG O O
The Ayurvedic medicine Triphala is known for its digestive benefits.	O B-DRUG O O O O O
The patient was diagnosed with diabetes and prescribed Diabecon.	O O O O O B DISEASE O O B-DRUG
The patient was prescribed Paracetamol for pain relief.	O O O B-DRUG O O
The doctor recommended Ibuprofen for inflammation.	O O B-DRUG O
The patient has been taking Aspirin for heart health.	O O O B-DRUG O O
The pharmacist dispensed Amoxicillin for the patient's infection.	O O O B-DRUG O O O O
The patient was advised to take Omeprazole for acid reflux.	O O O B-DRUG O O O
The doctor prescribed Simvastatin for cholesterol management.	O O B-DRUG O O O
The patient is allergic to Penicillin and should avoid it.	O O O B-DRUG O O O
The medication Acetaminophen is commonly used for fever reduction.	O B-DRUG O O O O
the patient was given Metformin for diabetes control.	O O O B-DRUG O O
The doctor recommended Lisinopril for blood pressure regulation	O O B-DRUG O O O O
The patient is taking Atorvastatin for lipid control.	O O O B-DRUG O O
The pharmacist dispensed Ciprofloxacin for the patient's urinary tract infectio	O O O B-DRUG O O O O
The doctor prescribed Albuterol for asthma management.	O O B-DRUG O O O
The patient has been using Hydrocortisone cream for skin irritation.	O O O B-DRUG O O O
The medication Loratadine is commonly used for allergy relief.	O B-DRUG O O O O
The patient was advised to take Warfarin for blood thinning.	O O O B-DRUG O O
The doctor recommended Metoprolol for heart rate control.	O O B-DRUG O O O
The patient is allergic to Sulfa drugs and should avoid them.	O O O B-DRUG O O O
The pharmacist dispensed Fluoxetine for the patient's depression.	O O B-DRUG O O O
The medication Pantoprazole is commonly used for gastric acid suppression.	O B-DRUG O O O O
The patient was given Levothyroxine for thyroid hormone replacement.	O O O B-DRUG O O
The doctor prescribed Sertraline for anxiety management.	O O B-DRUG O O O
The patient has been using Mupirocin ointment for skin infection.	O O O B-DRUG O O O O
The medication Gabapentin is commonly used for neuropathic pain.	O B-DRUG O O O O
The patient was advised to take Ramipril for hypertension control.	O O O B-DRUG O O O
The patient was diagnosed with hypertension and prescribed Amlodipine for	O O B-DISEASE O O B-DRUG O O O O O
The doctor recommended Metronidazole for the treatment of bacterial infect	O O B-DRUG O O O O O O B-DISEASE
The patient has been using Prozac for the management of depression.	O O O B-DRUG O O B-DISEASE

Figure 6.1.3

2. **Data Diversity:** Ensure that the collected data represent a diverse range of prescription types, including different medical conditions, medication classes, and patient demographics. This diversity helps train the automated system to handle various prescription formats, abbreviations, and language variations.

3. **Annotated Data:** Annotate the collected prescription data with the relevant entities to create a labeled dataset. The annotations may include medication names,

dosages, instructions, patient details, and other pertinent information. Annotation can be performed by domain experts or trained annotators who are familiar with prescription structures and terminology.

When we talk about annotation, especially in natural language processing (NLP) tasks like named entity recognition (NER), the IOB (Inside-Outside-Beginning) tagging scheme is commonly used. IOB tagging allows annotators to label entities in text, indicating the position of each token within an entity. Following are some variants of annotation commonly used with respect to IOB tagging.

Named Entity Recognition (NER) Annotation:

NER annotation involves identifying and labeling specific named entities in text, such as persons, organizations, locations, dates, or any other predefined entity types. Using IOB tagging, annotators assign labels to each token, indicating whether it is inside an entity (I), at the beginning of an entity (B), or outside an entity (O).

IOB Tagging:

- IOB tagging is a labeling scheme used to annotate named entities, chunks, relations, events, or semantic roles in natural language text.
- The acronym "IOB" stands for Inside, Outside, and Beginning, representing the three possible tags assigned to each token in the text.
- Each token is labeled with an IOB tag indicating its position relative to an entity or chunk. The three possible tags are "B" (beginning), "I" (inside), and "O" (outside).
- The "B" tag is assigned to the first token of an entity or chunk, indicating the beginning of a new entity or chunk.
- The "I" tag is assigned to tokens that are inside an entity or chunk, occurring after the beginning token.
- The "O" tag is assigned to tokens that are outside any entity or chunk, indicating that they do not belong to any labeled entity or chunk.
- IOB tagging allows for the representation of multiple entities or chunks within a single sentence by transitioning between "B" and "I" tags.
- The IOB tagging scheme enables the precise identification and differentiation of entities or chunks, facilitating subsequent processing or analysis tasks.

- IOB tags can be extended with additional labels or attributes to represent more complex structures or relations, depending on the specific task requirements.
- IOB tagging is widely used in various NLP tasks, and it serves as a standard for representing entity boundaries, event mentions, semantic roles, and other relevant linguistic structures within text data.

6.1.2 Data Preprocessing

Post data collection, we need to preprocess it to remove irrelevant information and clean it for further analysis. This may include eliminating stop words, common words, white spaces, and any other unnecessary characters of symbols. This step makes sure that our data is clean and ready for NLP processing.

6.1.3 NLP Pipeline

Next, we build a blank NLP pipeline that will consist of various processing steps such as tokenization, part-of-speech tagging, and dependency parsing. These steps will be handled using libraries such as NLTK, Spacy, or other relevant NLP libraries.

1. **Text Preprocessing:** The first task is to preprocess the prescription text to remove noise, normalize the text, and handle any special characters or formatting issues. This may include steps such as lowercasing the text, removing punctuation, and handling whitespace.
2. **Tokenization:** The prescription text is divided into individual tokens, such as words or sub words, using tokenization techniques. Tokenization allows the system to process the text on a granular level, facilitating subsequent analysis and extraction.
3. **Part-of-Speech (POS) Tagging:** POS tagging assigns grammatical labels (such as noun, verb, adjective) to each token in the prescription text. POS tags provide knowledge regarding the role and syntactic category of words, aiding in subsequent analysis and extraction steps.
4. **Named Entity Recognition (NER):** NER is a critical step in the pipeline. It identifies and extracts specific entities from the prescription text, such as medication names, dosages, instructions, and patient details. NER models are trained to recognize predefined entity types and label them accordingly.
5. **Entity Normalization:** After identifying entities, normalization techniques can be applied to standardize the extracted information. For example, medication names

may be mapped to a standard drug database, dosages may be converted to a consistent format, and patient details may be validated against known naming conventions.

6. **Post-processing and Validation:** After the extraction of entities, post-processing steps are performed to validate and refine the extracted information. This may involve consistency checks, rule-based validation, or additional external validation sources to ensure accuracy and integrity of the extracted prescription details.

6.1.4 User Input Preprocessing

When a user provides input to our system, we will process it by tokenizing the input text, tagging named entities, and applying other NLP processing steps from our pipeline. This will allow us to extract relevant information from the user input.

6.1.5 Passing Entities to API

Once the highlighted entities are received, they are passed to the API to drugs.com which then searches its database to find the drug in question. If it exists, it returns the user a clickable link that the user can visit to view more information.

6.1.6 Option to use custom dataset

Users have an option to choose between three datasets we have developed. Individuals have the opportunity to test our custom dataset for Indian Medications to compare and review results.

6.2 APPLICATION DESIGN

- **User Interface:** Design an intuitive and user-friendly interface that allows users to input medical prescriptions. The interface should be visually appealing and user-friendly, with explicit guidance for users on how to input the prescription.
- **Prescription Input Processing:** Develop a module that can take in the medical prescription as input and preprocess it to eliminate any useless information, such as patient names or addresses. This module should also handle different formats of prescriptions, such as scanned images or digital text files, and convert them into a standardized format for further processing.

- **NLP Pipeline:** Build a comprehensive NLP pipeline that includes steps such as tokenization, stop word removal, and entity recognition. This pipeline should be optimized for medical text and tailored to identify medical terms and chemicals involved in the prescription. This can be achieved by training a custom NER model on preprocessed data specific to the medical domain.
- **Output Presentation:** Design a module that presents the output image along with the identified medical terms and chemicals in a meaningful way to the users. This can involve displaying the output image with annotated entities, providing a summary of the recognized entities, and offering options for further actions, such as saving the image or exporting the results.
- **Error Handling:** Implement error handling mechanisms to handle any exceptions or bugs that may arise during the processing of prescriptions. This can involve providing informative error messages to users, logging errors for troubleshooting, and implementing fallback options in case of failures.
- **Deployment:** Develop a plan for deploying the application in a production environment, considering factors such as hosting options, server configuration, and system integration. Ensure that the deployment plan includes thorough testing and validation to ensure the system's stability and reliability in a production setting.

Chapter 7

RESULTS AND DISCUSSION

7.1 RESULTS

We will now examine the performance of the discussed models. When evaluating an NER model, three parameters are considered:

Precision: This metric evaluates the accuracy of the NER system's predictions. Precision measures the ratio of true positive predictions (correctly identified named entities) to the total number of predicted named entities. In essence, precision assesses the proportion of predicted named entities that are genuinely correct.

Precision scores help assess the system's capacity to accurately identify and classify named entities without generating numerous false positive errors. A higher precision score indicates a lower rate of false positives, indicating greater precision in predictions. Precision can be calculated using the following formula:

Precision = (True Positives) / (True Positives + False Positives)

Where:

True Positives(TP) are the count of correctly predicted named entities.

False Positives(FP) are the count of incorrectly predicted named entities.

Recall: This metric evaluates the comprehensiveness of the NER system's predictions. It measures the ratio of true positive predictions to the total number of actual named entities present in the text. In essence, recall assesses the proportion of actual named entities correctly identified by the system.

The recall score aids in assessing the system's effectiveness in capturing named entities without overlooking many of them. A higher recall score indicates a lower rate of false negatives, indicating greater comprehensiveness in identifying named entities.

Recall can be calculated using the following formula:

Recall = (True Positives) / (True Positives + False Negatives)

Where:

True Positives(TP) are the count of correctly predicted named entities.

False Negatives(FN) are the count of actual named entities that were not predicted by the system.

NER F-Score: Also referred to as the F1 score, it merges precision and recall offering a balanced assessment of the NER system's performance. The F1 score represents the harmonic mean of precision and recall, computed using the formula:

F1 Score = 2 * (Precision * Recall) / (Precision + Recall)

Considering all three metrics is crucial for gaining a comprehensive grasp of the NER system's performance, as each metric contributes valuable insights into various aspects of the system's functionality.

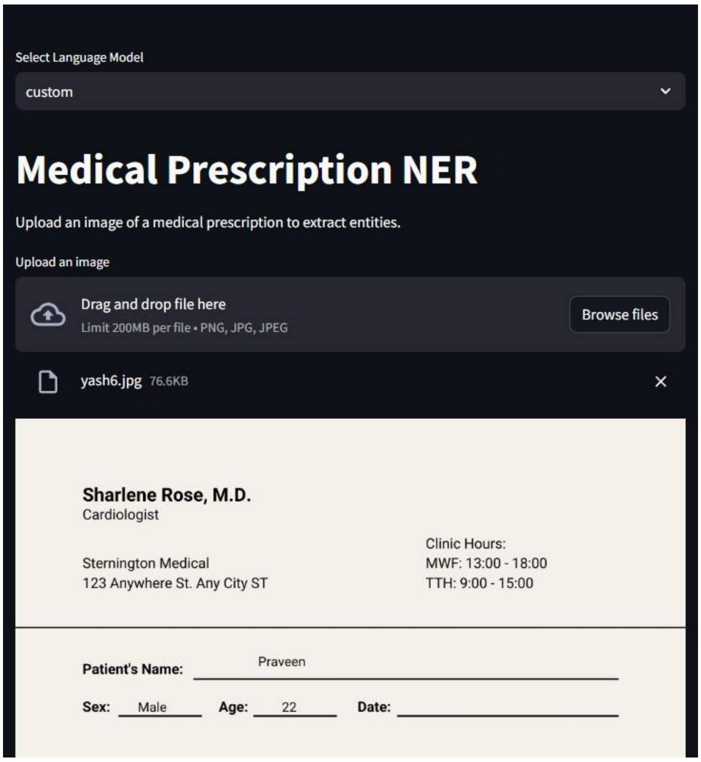
	Precision	Recall	F1
Med7	0.865	0.889	0.887
BC5CDR	0.884	0.873	0.884
CUSTOM	0.655	0.697	0.675

Table 7.1

7.2 PROJECT EXECUTION

7.2.1 Input and User Interface

A simple web-based interface where the user can choose the desired dataset(language model) and can upload an image of the medical prescription from the local system.



Select Language Model: custom

Medical Prescription NER

Upload an image of a medical prescription to extract entities.

Upload an image

Drag and drop file here

Limit 200MB per file • PNG, JPG, JPEG

yash6.jpg 76.6KB

Sharlene Rose, M.D.
Cardiologist

Sternington Medical
123 Anywhere St. Any City ST

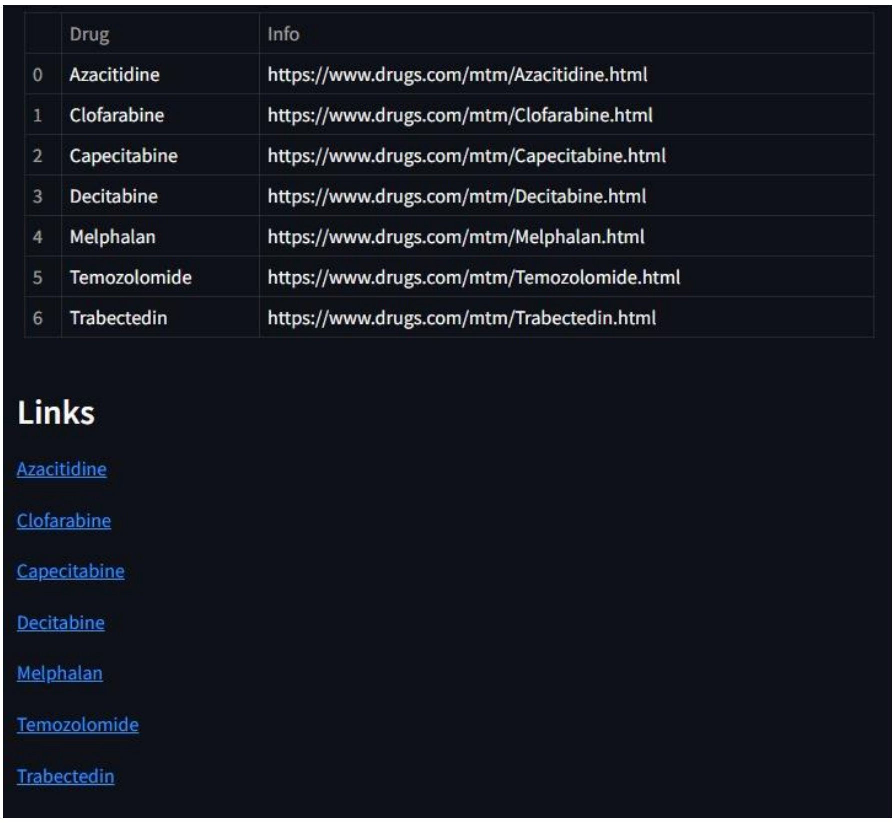
Clinic Hours:
MWF: 13:00 - 18:00
TTH: 9:00 - 15:00

Patient's Name: Praveen

Sex: Male Age: 22 Date:

Figure 7.2.1

7.2.2 Output



	Drug	Info
0	Azacitidine	https://www.drugs.com/mtm/Azacitidine.html
1	Clofarabine	https://www.drugs.com/mtm/Clofarabine.html
2	Capecitabine	https://www.drugs.com/mtm/Capecitabine.html
3	Decitabine	https://www.drugs.com/mtm/Decitabine.html
4	Melphalan	https://www.drugs.com/mtm/Melphalan.html
5	Temozolomide	https://www.drugs.com/mtm/Temozolomide.html
6	Trabectedin	https://www.drugs.com/mtm/Trabectedin.html

Links

Azacitidine

Clofarabine

Capecitabine

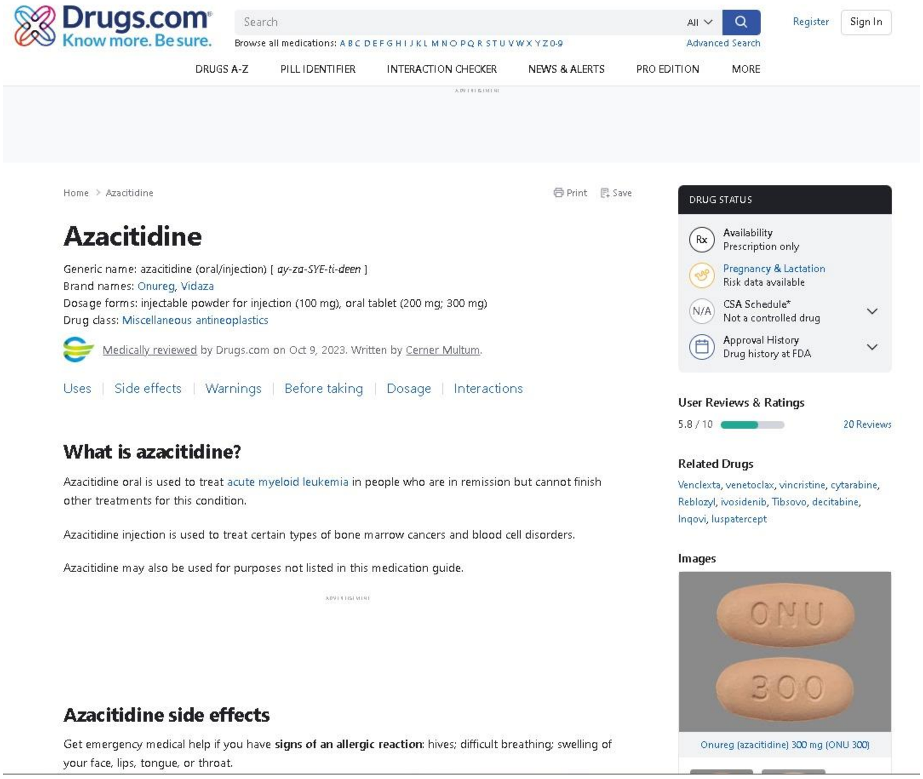
Decitabine

Melphalan

Temozolomide

Trabectedin

Figure 7.2.2



Azacitidine

Generic name: azacitidine (oral/injection) [*ay-za-SYE-ti-deen*]
Brand names: Onureg, Vidaza
Dosage forms: injectable powder for injection (100 mg), oral tablet (200 mg; 300 mg)
Drug class: Miscellaneous antineoplastics

Medically reviewed by Drugs.com on Oct 9, 2023. Written by Cerner Multum.

What is azacitidine?

Azacitidine oral is used to treat acute myeloid leukemia in people who are in remission but cannot finish other treatments for this condition.

Azacitidine injection is used to treat certain types of bone marrow cancers and blood cell disorders.

Azacitidine may also be used for purposes not listed in this medication guide.

Azacitidine side effects

Get emergency medical help if you have **signs of an allergic reaction**: hives; difficult breathing; swelling of your face, lips, tongue, or throat.

DRUG STATUS

Availability
Prescription only

Pregnancy & Lactation
Risk data available

CSA Schedule*
Not a controlled drug

Approval History
Drug history at FDA

User Reviews & Ratings

5.8 / 10 20 Reviews

Related Drugs

Venclexta, venetoclax, vincristine, cytarabine, Reblozyl, ivosidenib, Tibsovo, decitabine, Inqovi, luspatercept

Images

Onureg (azacitidine) 300 mg (ONU 300)

Figure 7.2.3

Chapter 8

CONCLUSION AND FUTURE WORK

8.1 CONCLUSION

The project focusing on the automation of medicine prescription identification has effectively addressed the necessity for a precise and efficient system to extract pertinent details from prescription documents. By utilizing natural language processing (NLP) methods and tailored named entity recognition (NER), the project has showcased the potential for streamlining and automating the prescription identification procedure.

Throughout the project duration, we have delved into various aspects of the NLP pipeline, encompassing text preprocessing, tokenization, part-of-speech tagging, named entity recognition, entity normalization, dependency parsing, contextual comprehension, and post-processing. Each of these stages has played a pivotal role in accurately pinpointing and extracting crucial entities like medication names, dosages, instructions, and patient particulars.

Through the assembly of a diverse and meticulously annotated dataset, we managed to train a bespoke NER model that exhibited promising performance in detecting prescription entities. The model underwent fine-tuning and evaluation processes to enhance its precision and adaptability. The continual refinement and incorporation of user input have further bolstered the system's efficacy in handling various prescription formats and complexities.

The project's advancement in the automated identification of medicine prescriptions signifies a significant stride in developing a resilient and efficient system for extracting prescription details. By harnessing the capabilities of NLP techniques, tailored NER models, and ongoing refinement, the project has set the stage for enhancing the efficiency, precision, and dependability of prescription processing in healthcare environments. This initiative paves the way for automating tasks, minimizing errors, and elevating patient care standards within the realm of medicine prescription management.

8.2 FUTURE SCOPE

The automated identification of medicine prescription project holds immense potential for future advancements and expansions. Here are some important aspects of future scope for continuous development and improvement:

1. **Custom Dataset for Indian Medicine:** Generating a custom dataset for Indian drugs holds very important potential for future R&D and applications. By creating a comprehensive and localized dataset, several benefits can be realized:
   a. Enhanced accuracy and relevance
   b. Improved drug identification
   c. Easy integration with Indian use cases
2. **Enhanced Entity Recognition:** Continuously refining and expanding entity recognition capabilities can lead to more accurate identification of medication names, dosages, instructions, and patient details. Incorporating advanced deep learning(DL) techniques, such as transformer-based models or domain-specific embeddings, can improve the ability of the system to handle complex prescription patterns and variations.
3. **Multilingual Support:** Expanding the system's capabilities to support multiple languages can enable its adoption in diverse healthcare settings worldwide. Developing language-specific models and incorporating translation techniques can facilitate automated identification of medicine prescriptions in different linguistic contexts.
4. **Handling Ambiguities and Contextual Understanding:** Enhancing the system's ability to handle ambiguities, context, and implicit information can improve accuracy and user experience. Implementing advanced contextual understanding techniques, such as coreference resolution and context-aware language models, can help disambiguate entities and generate deeper insights from prescription texts.
5. **Adapting to Evolving Prescription Practices:** The project can continuously adapt to changing prescription practices, new medications, and emerging trends. Regularly updating the system with new annotated data and monitoring real-world prescription patterns can ensure its relevance and effectiveness over time.
6. **Real-time Processing and Feedback:** Enabling real-time processing of prescription texts can streamline healthcare workflows and improve efficiency. Providing instant feedback to users, such as highlighting potential errors or missing

information in the prescription, can assist healthcare professionals in ensuring accuracy and completeness.

7. **Privacy and Security Enhancements:** Continuously enhancing privacy and security measures is crucial to protect sensitive patient information. Incorporating encryption techniques, access controls, and adherence to data privacy laws can ensure the confidentiality and maintaining the integrity of prescription data..

8. **Collaboration and Industry Adoption:** Collaborating with healthcare providers, institutions, and pharmaceutical companies can facilitate the adoption and integration of the automated identification system into existing healthcare systems. Partnering with industry stakeholders can lead to real-world implementation, user feedback, and further improvements based on practical use cases.

9. **Expanding Application Areas:** Exploring the application of automated identification of medicine prescription beyond healthcare settings, such as pharmaceutical research, clinical trials, and drug safety monitoring, can unlock new possibilities and avenues for the project.

10. **User-Friendly Interfaces:** Developing user-friendly interfaces, such as mobile applications or web-based platforms, can simplify the interaction between healthcare professionals and the automated identification system. Intuitive interfaces can enhance usability, accessibility, and adoption of the system in various healthcare settings.

In conclusion, the future scope of the automated identification of medicine prescription project lies in continuously improving accuracy, expanding language support, adapting to evolving practices, integrating with decision support systems, ensuring privacy and security, fostering collaboration, and exploring new application areas. By pursuing these avenues, the project can contribute to the advancement of medication management, patient safety, and healthcare efficiency.

APPENDIX A- Project Team Details

Project Title	Title of the project		
USN	**Team Members**	**Email ID**	**Mobile number**
01JST20CS180	Aditya Raj	adirajravi13@gmail.com	6205617602
01JST20CS065	Harshit Sharma	harshit2112003@gmail.com	7019126215
01JST20CS109	Pavan Prakash	ppavan250901@gmail.com	7795911522
01JST20CS147	Shania Vijay	shaniavijay14@gmail.com	8296811670

Aditya Raj

Harshit Sharma

Pavan Prakash

Shania Vijay

APPENDIX B- Cos, Pos and PSOs

Mapping for the project work(20CS83P)

Course Outcomes:

CO1: Formulate the problem definition, conduct literature review and apply requirements analysis.

CO2: Develop and implement algorithms for solving the problem formulated.

CO3: Comprehend, present and defend the results of exhaustive testing and explain the major findings.

Program Outcomes:

PO1: Apply knowledge of computing, mathematics, science, and foundational engineering concepts to solve the computer engineering problems.

PO2: Identify, formulate and analyze complex engineering problems.

PO3: Plan, implement and evaluate a computer-based system to meet desired societal needs such as economic, environmental, political, healthcare and safety within realistic constraints.

PO4: Incorporate research methods to design and conduct experiments to investigate real- time problems, to analyze, interpret and provide feasible conclusion.

PO5: Propose innovative ideas and solutions using modern tools.

PO6: Apply computing knowledge to assess societal, health, safety, legal and cultural issues and the consequent responsibilities relevant to professional engineering practice.

PO7: Analyze the local and global impact of computing on individuals and organizations for sustainable development.

PO8: Adopt ethical principles and uphold the responsibilities and norms of computer engineering practice.

PO9: Work effectively as an individual and as a member or leader in diverse teams and in multidisciplinary domains.

PO10: Effectively communicate and comprehend.

PO11: Demonstrate and apply engineering knowledge and management principles to manage projects in multidisciplinary environments.

PO12: Recognize contemporary issues and adapt to technological changes for lifelong learning.

Program Specific Outcomes:

PSO1: Problem Solving Skills: Ability to apply standard practices and mathematical methodologies to solve computational tasks, model real world problems in the areas of database systems, system software, web technologies and Networking solutions with an appropriate knowledge of Data structures and Algorithms.

PSO2: Knowledge of Computer Systems: An understanding of the structure and working of the computer systems with performance study of various computing architectures.

PSO3: Successful Career and Entrepreneurship: The ability to get acquaintance with the state-of-the-art software technologies leading to entrepreneurship and higher studies.

PSO4: Computing and Research Ability: Ability to use knowledge in various domains to identify research gaps and to provide solution to new ideas leading to innovations.

Justification for CO-PO and PSO mapping

In our Final Year Project (FYP), we have addressed several Course Outcomes (COs), Program Outcomes (POs), and Program Specific Outcomes (PSOs) through the development and implementation of our automated identification of medicine prescription system.

CO1: Formulate the problem definition, conduct literature review, and apply requirements analysis.

We have successfully formulated the problem definition by identifying the need for an efficient and accurate system to extract relevant information from prescription texts. Conducting a thorough literature review helped us understand existing solutions and techniques in the field of natural language processing (NLP) and named entity recognition (NER). Applying requirements analysis allowed us to define the functionalities and features required for our system.

CO2: Develop and implement algorithms for solving the problem formulated.

We developed and implemented algorithms for various components of the natural language processing (NLP) pipeline, including text preprocessing, tokenization, part-of-speech tagging, named entity recognition (NER), and entity normalization.

These algorithms were designed to accurately identify and extract medication names, dosages, instructions, and patient details from prescription texts.

CO3: Comprehend, present, and defend the results of exhaustive testing and explain the major findings.

We comprehended the results of exhaustive testing by evaluating the performance of our system in accurately identifying and extracting prescription entities. Through thorough testing, we were able to identify strengths and weaknesses in our system and explain the major findings to stakeholders. We presented and defended our results in project presentations and reports.

PO3: Plan, implement and evaluate a computer-based system to meet desired societal needs such as economic, environmental, pharmaceutical, healthcare, and safety within realistic constraints.

Our project aimed to meet the societal need for an efficient and accurate system for medication prescription management in healthcare settings. We planned, implemented, and evaluated our computer-based system to address this need, considering constraints such as accuracy, usability, and scalability.

PSO1: Problem Solving Skills: Ability to apply standard practices and mathematical methodologies to solve computational tasks, model real-world problems in the areas of database systems, system software, web technologies, and Networking solutions with appropriate knowledge of Data structures and Algorithms.

Through the development of our system, we applied standard practices and mathematical methodologies to solve computational tasks related to natural language processing (NLP) and named entity recognition (NER). We modeled real-world problems in the healthcare domain, specifically in medication prescription management, by implementing algorithms and techniques related to data structures and algorithms.

PSO2: Knowledge of Computer Systems: An understanding of the structure and working of computer systems with the performance study of various computing architectures.

We gained an understanding of computer systems by implementing and testing our system on different computing architectures, including desktop computers and mobile devices. We studied the performance of our system in terms of efficiency, scalability, and resource utilization.

PSO4: Computing and Research Ability: Ability to use knowledge in various domains to identify research gaps and provide solutions to new ideas leading to innovations.

Throughout the project, we utilized our computing knowledge to identify research gaps in medication prescription management and provide innovative solutions through the development of our automated identification system. We conducted research to explore state-of-the-art techniques and technologies in natural language processing (NLP) and named entity recognition (NER), leading to the innovation of our system.

Table of Mapping of CO, PO and PSO:

SUBJECT	CODE	CO	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12	PSO1	PSO2	PSO3	PSO4
Project Work	20CS83P	CO1																
		CO2																
		CO3																

Note:

Scale

0 –Not Applicable

1 – Low relevance Scale

2 – Medium relevance Scale

3 – High relevance Scale

REFERENCES

[1] Vincenza Carchiolo, Alessandro Longheu, Giuseppa Reitano, Luca Zagarella "Medical prescription classification: a NLP-based approach", 2019 Federated Conference on Computer Science and Information Systems (FedCSIS)

[2] Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean, "Efficient Estimation of Word Representations in Vector Space", 2013

[3] Arya Roy, "Recent Trends in Named Entity Recognition", 2021.

[4] Alexander M. Rush, Sumit Chopra, Jason Weston, "A Neural Attention Model for Abstractive Sentence Summarization", 2015.

[5] Ashish Vaswani, "Attention Is All You Need", 2017.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", 2019.

[7] Andrey Kormilitzin, Nemanja Vaci, Qiang Liu, Alejo Nevado-Holgado, "Med7: a transferable clinical natural language processing model for electronic health records", 2020.