Semiconductor Optoelectronics, *Vol. 42 No. 1 (2023), 1450-1467*
https://bdtgd.cn/

DIABETES DRUG ONTOLOGY MAPPING IN EHRS USING WORD EMBEDDING BERT AND A BIDIRECTIONAL LSTM MODEL

K. Pushpavathi
Associate professor, Department of Computer Science, School of Arts and Science, AV Campus, Vinayaka Mission's Research Foundation (DU) Paiyanoor, Tamil Nādu, India

Dr. K.L. Shunmuganathan
Director (Academic), Jaya Engineering College, Thiruninravur, Chennai

Abstract
Electronic health records (EHR) store patient data in an unstructured format, making it challenging for healthcare providers to extract meaningful insights from the data. The use of natural language processing (NLP) and machine learning algorithms can transform EHRs into useful tools for clinical decision making and population health management. By analyzing unstructured data such as physician notes and patient feedback, natural language processing (NLP) can identify patterns and trends that can improve treatment plans and patient outcomes. In addition, machine learning algorithms can aid in predicting disease progression and identifying patients at risk. Patient data should be maintained as structured data in order to effectively extract the information and maintain data consistency to reduce the risk of errors. In terms of Type 1 and Type 2 ontologies, numerous ontologies exist for diabetic patients. To manage the medication details of multiple patients, we've developed a diabetic drug ontology for analyzing the dosage, side effects, and chemical components of both diabetic patients and the prescribed medications. This ontology enables healthcare professionals to make more informed decisions regarding the medication management of diabetic patients, resulting in improved patient outcomes and quality of life. In addition, the use of ontologies in healthcare can facilitate interoperability between different electronic health record systems, allowing for the seamless sharing of patient data and enhancing the overall delivery of healthcare. Then, we performed ontology mapping on various ontologies to examine the semantic relationships between them. We have utilized various word embedding models, including word2vec, Bidirectional Encoder Representation From Transformers (BERT), and the bidirectional long short-term memory (Bilstm) model, to examine the similarity of word embeddings across various ontologies. The Bilstm model yields superior precision, recall, F1 score, and precision.
Keywords: Ontology mapping, Word2vec, skip n-gram, BERT, Bilstm, Word Embeddings

1. Introduction

Healthcare records detail a patient's medical history, diagnosis, and follow-ups. Healthcare records are kept by carers, hospitals, clinics, and individuals. Healthcare records are used to keep track of a patient's medical history, monitor their health, improve communication, and

conduct public health

research. Healthcare records, such as photographs and electronic health records, are kept in the form of instructions (EHR). The EHR stores patients' demographic information, medical history, prescriptions, laboratory test results, immunizations, diagnostic imaging, and treatment plans. An unstructured EHR, on the other hand, has several disadvantages, including difficulty in extracting data, inconsistencies in data, a lack of interoperability, a higher risk of errors, and a security risk for patients' data.

To prevent the aforementioned issues, healthcare providers are increasingly adopting structured EHRs. Structured EHRs provide a standardized format for data entry, making it easier to extract and analyse data. They also reduce inconsistencies in data by enforcing standardization across all entries. Interoperability is improved through the use of standardized formats, allowing different systems to communicate with each other seamlessly. The risk of errors is minimized as structured EHRs often include built-in error-checking mechanisms. Security risks are also reduced as structured EHRs allow for more granular access controls and audit trails. Overall, structured EHRs offer numerous benefits over unstructured ones and are becoming increasingly popular in the healthcare industry as a result. As technology continues to advance, we can expect to see even more improvements in the field of electronic health records, ultimately leading to better patient care and outcomes. EHR records are converted to structured documents using an ontology. Ontology is concerned with the relationship between concepts.

Ontology is a very effective method for organizing data and ensuring consistency in EHR records. By using ontology, healthcare providers can create a standardised vocabulary that is used across all systems, reducing the risk of inconsistencies in data. This also helps with interoperability, as different systems can communicate with each other using the same language. Additionally, converting EHR records to structured documents through ontology reduces the chance of errors as the data is organised in a logical and consistent manner. Finally, ontologies can improve security by allowing for fine-grained access control to patient data. Overall, implementing ontologies in EHR systems can address many of the issues that currently plague healthcare providers and lead to better patient outcomes. All medical documents are written in the same language. As a result, it is used for simple communication and collaboration between healthcare providers, and it also enables the creation of decision support systems that can help improve patient outcomes.

EHRs have unique features, but they all share the same goal of improving patient outcomes. The use of EHRs has revolutionised the healthcare industry by providing a more efficient and accurate way of storing and sharing patient information. This technology has allowed healthcare providers to access patient records from anywhere at any time, which is particularly important in emergency situations. Additionally, EHRs can help reduce medical errors by providing real-time alerts for potential drug interactions or allergies. They also allow for easier tracking of patient progress and can help identify patterns that may indicate a need for a change in treatment plan.

1.1 Problem Statement

While there are still challenges to overcome in terms of interoperability between different EHR systems, the benefits they provide for better patient outcomes cannot be ignored. Hence, the integration of data, analysis of data, and exchange of data are difficult to understand. The above issues are solved through standard protocols and interfaces that ensure interoperability between EHR systems, allowing for the exchange of patient data and clinical information in a secure and efficient manner. The integration, analysis, and exchange of healthcare data require a multidisciplinary approach involving technology, policy, and governance considerations. It involves mapping terminologies to a common language, using advanced analytics tools, and developing robust data governance frameworks. Diabetes is a global health concern that affects millions of people worldwide. By leveraging secure and efficient healthcare data exchange methods, we can better understand this disease and develop effective treatments to improve the lives of those affected by it.

Each diabetic patient has different complications and medications. Frameworks such as the Chronic Care Model and Patient-Centered Medical Home provide comprehensive care to patients with diabetes, and digital health technologies can be integrated to help monitor blood glucose levels and receive personalized coaching. Hence, the doctors cannot maintain a separate EHR record. EHRs and telemedicine are valuable tools for diabetes management, allowing patients to consult with their healthcare provider remotely, reducing the need for in-person visits, and improving access to care. They should easily extract the information through ontology mapping. Personalized coaching is a powerful tool for improving diabetes management and helping patients achieve better health outcomes. It involves working with a healthcare provider to develop an individualized plan, using EHRs and telemedicine, and using ontology mapping to manage EHR records. Healthcare providers should use an ontology mapping technique to ensure seamless data transfer between EHRs and telemedicine platforms, improving patient care and reducing administrative burden. EHRs and telemedicine have the potential to revolutionise healthcare delivery by increasing access to care and reducing costs. Hence, we have developed the Drug Ontology Mapping Tool (DOM). The DOM enables seamless data transfer between EHRs and telemedicine platforms, reducing administrative burden and streamlining patient care. It also increases access to care and reduces costs, leading to better health outcomes for patients. to extract the implicit information from different ontologies by applying ontology mapping techniques.

1.2 Contribution

1. The DOM model creates a connection between EHRs and telemedicine platforms, reducing administrative burden and streamlining patient care. It also reduces costs by eliminating redundant data entry and minimizing errors in patient records.
2. Ontology mapping techniques are used in healthcare to integrate data from different sources, such as the DOM model, which reduces administrative burden and improves patient care by enabling clinicians to access relevant information in real-time. from the DOM model
3. To create a standardized vocabulary for each medication, making it easier for clinicians to compare and contrast different drugs.
4. To ensure accuracy and consistency, we evaluate the mapping of ontologies by

calculating pre- and post-mapping similarity scores.

2. Related Work

Standardizing medical vocabularies is essential to ensuring accuracy and consistency in the mapping process, and other efforts are underway to create a more consistent and accurate EHR ontology that benefits both clinicians and patients. In addition, standardizing information among healthcare professionals is essential for ensuring consistency in the EHR ontology. Diabetes, being a universal disease, requires a standardized ontology for efficient management. The use of a consistent ontology can improve the quality of patient care by enabling better communication between healthcare professionals and reducing errors in diagnosis and treatment. It also facilitates research by allowing for more accurate analysis of data across different institutions. Therefore, it is important to continue efforts towards standardizing EHR ontology to improve healthcare outcomes for patients with diabetes and other diseases. Each medication has its own unique properties and potential side effects, making it crucial for clinicians to be able to compare and contrast different drugs. By mapping ontologies, we can simplify this process and make it easier for healthcare professionals to access accurate and consistent information.

The authors [1] outlined a neural word embedding translation model to convert standard medical jargon to Human Phenotype Ontology (HPO) terms. The term "HPO" refers to a systematic vocabulary of phenotypic anomalies used to define various diseases. A knowledge-based care-based reasoning system (KI-CBR) based on SNOMED-CT was developed [2] in order to assess 60 real cases of diabetes patients and locate essential documents. The author [3] suggested employing term frequency and inverse term frequency to increase the accuracy of doctor-patient matching for online medical therapy in order to avert the loss of specialists. The author [4] suggested a clinical decision support system to help diabetic patients in Sri Lanka. They presented a number of ontology-based models based on Type 1, Type 2, and gestational diabetes patients using the Jene Framework. A model based on the OMDP ontology was produced by [5]. The author[6] developed a model for predicting diabetes based on machine learning and an ontology. A fuzzy-based ontology framework was established by[7] in order to comprehend confusing medical terminology through the semantic knowledge of medical concepts. The Clinical Pathway Management System (CPMS) was developed by [8] in order to standardize and digitize clinical data in healthcare management information systems. Ontological concepts were used and expanded to produce smart healthcare integration services [9]. The author [10] developed a dynamic ontology-based healthcare monitoring system for managing inpatient and outpatient chronic disease patients. They used the semantic web rule language (SWRL) to divide the patients into normal and pathological groups. Based on patient choices, the author [11] established an interdisciplinary health care team (IHT). They establish a team leader and members and delegate responsibilities to them based on an ontology and a first-order logical framework. The author [12] developed the interestingness of semantic data using the BERT model and the Apriori method. To illustrate the relationship between various medications and their commonalities based on the symptoms they produce,The author[13] created a drug-based knowledge graph. The model is employed to produce fresh medicine suggestions. For diabetic patients in China, The author

[14] developed the Diabetes Care Pathway Ontology (DCPO). In Mexico, an ontology network for diabetes mellitus was built [15].

2.1 Research Gap

Data extraction from electronic health records presents several challenges, including data inconsistencies, a lack of interoperability, a higher risk of errors, and a security risk for patients. The diabetic ontology mapping tool is used to map various ontologies with various drugs available to diabetics. The proposed model lowers administrative risk and simplifies patient care, lowering costs and improving patient health outcomes.

3. Methodology

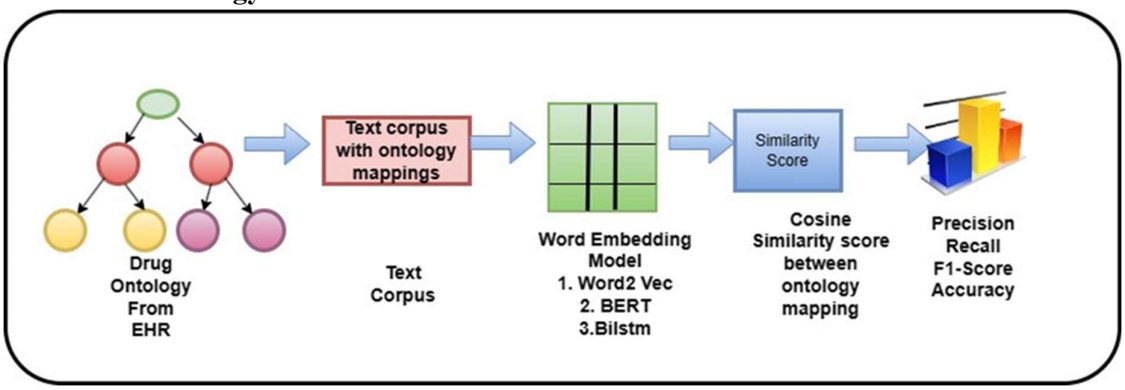


Figure 1:Architecture of Drug Ontology Mapping tool

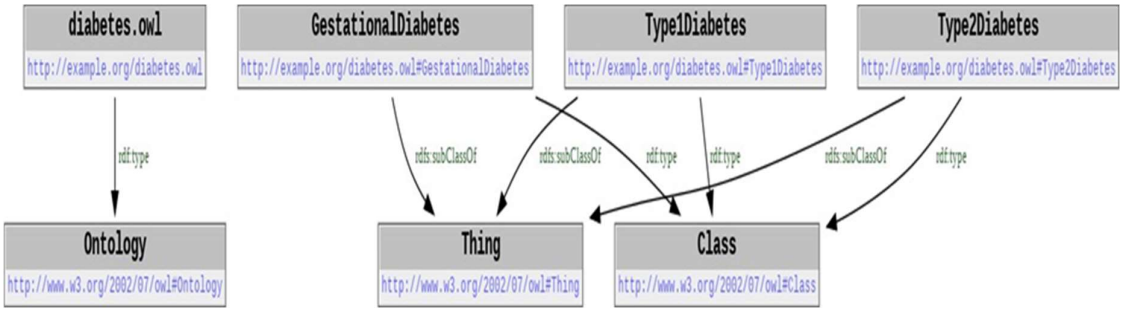
Information about diabetic patients, including their blood sugar analysis and type of diabetes, is included in the electronic health record. CSV-formatted electronic records could be at risk for privacy and security issues. All healthcare policy providers must have access to patient health information. As a result, all EHR records have been converted to structured documents. As a result, three types of ontologies containing the concepts of type of diabetes, medicines used for diabetes, and blood glucose level analysis are generated from the CSV file. The drug ontology is then translated into the associated ontology in a CSV file. The ontology's implicit information is then extracted, including its classes, specific properties of each class, and relationships among them. Next, three methods—Word2Vec, BERT and Bilstm word embedding models are used to analyze the ontology. Finally, the precision, recall, and accuracy of ontologies are assessed, and an F1 score is computed for each technique.Figure 1shows the architecture of ontology mapping tool.

3.1 Ontology Creation

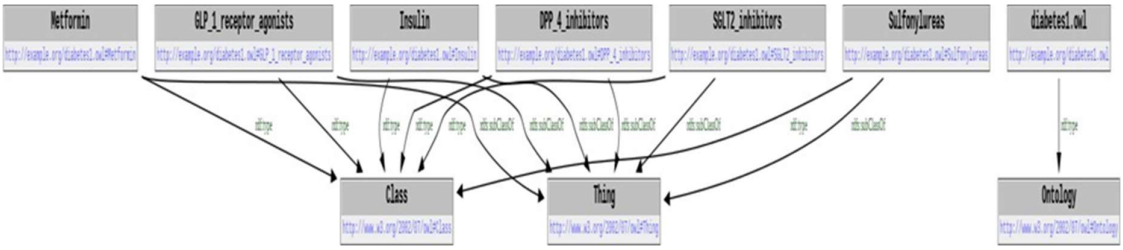
An ontology is a formal description of knowledge as a set of concepts within a domain and relationships between them. It involves incorporating data into the OWL ontology with relevant classes and properties. Creating an ontology from a CSV file helps standardize the data, make it interoperable, facilitate data integration and sharing, improve the quality of the data, and provide a common vocabulary and structured data. The creation of an ontology helps users reduce their involvement in mapping files. There are three types of ontologies constructed for the diabetic ontology model, such as types of diabetes, medicines used, and blood glucose

analysis, as shown in Figure 2.

(a). Type of Diabetes



(b). Blood Sugar Level Analysis



(c). Diabetes Medicines

Figure 2. Ontology in Type of Diabetes, Blood sugar Analysis and Diabetes Medicines

3.2 Ontology Mapping in Diabetes Medicine

High-quality data on diabetes complications, treatments, and medications can be generated via a top-down approach by designing an ontology. This method facilitates the development of a universal structure applicable in a variety of healthcare settings and scientific investigations. A diabetes ontology can also be used to spot research gaps and uncover new insights into this disease. uses evidence to extract key problems from diabetes medication. Domain experts should use ontology to distinguish between ontology rule types and medication classes. Improved semantic interoperability and treatment for diabetes are two outcomes of EHRs that incorporate ontologies. By enhancing the precision with which disease is predicted, diabetes is diagnosed, appropriate treatments are recommended, and comprehensive, consistent treatment plans are provided, the ontology of diabetic medicine enhances patient outcomes. It also offers a strategy based on solid evidence. Individualized care for diabetic patients is made possible by ontology mapping.Figure 3shows the sample drug ontology mapping.

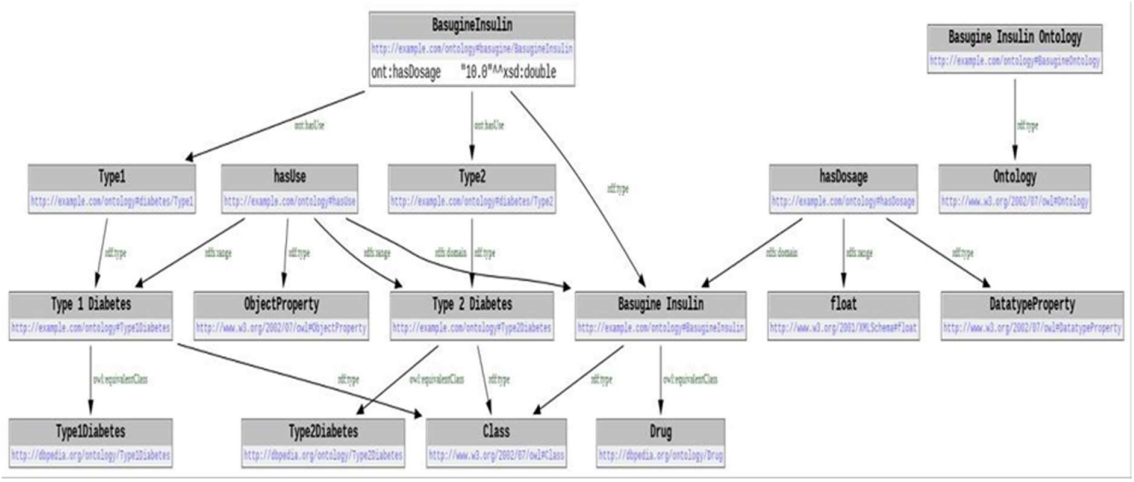


Figure 3: Drug ontology mapping

3.3 Extracting implicit information

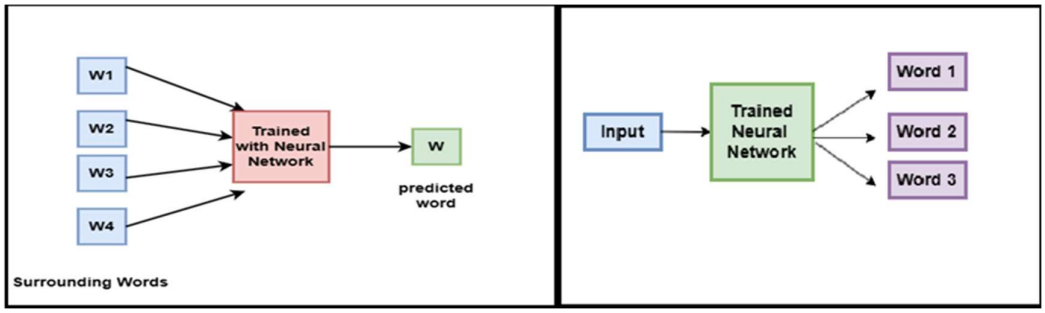
Extracting implicit information from an ontology includes augmentation and inferences to extract information from an ontology. Augmentation involves adding new concepts or relationships to the ontology, while inferences involve deriving new knowledge from existing information in the ontology. These processes can help improve the accuracy and completeness of the information contained within an ontology, making it a valuable resource for data analysis and decision-making. The extraction of information from text is used to identify appropriate information from unstructured text and add structured information from the ontology. Natural language processing techniques are often employed in text extraction to automatically identify and extract relevant information. Once extracted, this information can be integrated into the ontology to enhance its usefulness and relevance for various applications. The process includes multiple texts and the preprocessing of data into machine processable formats with word embedding models such as word

2 vector mode,Bidirectional Encoder Representation from Transformers(BERT), Bidirectional LSTM model

3.4 Word Embedding

Word embeddings (WEs) are a way of describing how each word is mapped onto a vector (a collection of real values). Word-and-context-meaning vectors (WEs) are quite real. A word's vector represents its meaning, its associations with other words, and the context in which those words are used. Every word has its own special vector. The dimensions of each vector range between 50 and 500.One-hot encoding recognizes the word "embedding." The N-dimensional vector that makes up a one-hot encoding is exactly the same length as the vocabulary size. The values 1 and 0 are encoded in each word. The number 1 represents their matching location. However, if there are more than 500 words inserted, the work becomes quite challenging. Custom word embedding is utilized to keep dimensions down. We set the vector size in our model to 10. Ten-dimensional vectors ranging from 0 to 1 represent each individual character in a dictionary. The vector sequences of closely related words are virtually identical. The continuous bag of words model and the skip n- gram model are two examples of training

algorithms used to create word embedding models. The CBOW[16]model predicts the central word from a series of input vectors. In direct opposition to the CBOW model is the skip n-gram (Jiho Noh) model. The focus word is used to determine the cluster of words.Figure 4 shows the architecture of Word 2 vector model.



CBOW Skip-ngram

Figure 4 : Architecture of Word 2 Vector CBOW and skip n-gram model

3.4.1 Algorithm

Initialize an empty vocabulary set (V)={} N varies from 50 -500
For each corpus:
Tokenize the individual words in the corpus Update the set V with unique words
End corpus
Assign unique index to each word in the document V[word]=index
Initialize the binary vector matrix with the dimensions N×V, N is the word and V is a vocabulary
Set the binary index for the corresponding word as 1 in the specified location

3.4.2 Algorithm: Skip n-gram Input:Corpus,number of words in ngram, skip count For each sentence:

Split sentence into array of Words Initialize result=[]
For I = 1 to len(words)-skip count: ngram=[]
For j=0 to n:
If j%(skipcount+1)==0: Ngram.push(word)
End For j End For i
result.push(ngram) End Sentence

3.5 Bidirectional Encoder Representation from transformers(BERT)

The Bidirectional Encoder Representations from Transformers (BERT)[17], [12] architecture paradigm describes transformers. BERT consists of an encoder and a decoder, both of which are trained on massive quantities of text data to comprehend language and generate predictions. The encoder is used for tasks such as language comprehension and text classification, whereas the decoder is used for language generation and machine translation. BERT comprises 12 bidirectional encoders with 12 bidirectional self-attention heads and 24 encoders with 16 bidirectional self attention networks, making it a potent natural language processing tool. Its ability to comprehend the context of words in a sentence has resulted in significant enhancements to duties such as question answering and sentiment analysis. In addition, BERT

has been optimized for specific domains, such as biomedical text and legal documents, thereby expanding its applicability across industries. The BERT model can read data either from left to right or from right to left. BERT performs two tasks, including language modelling and predicting the next sentence. Using statistical probabilities and language modelling, we have predicted the next word in a text. The BERT model generates text that is coherent and pertinent to its context. The relationships between each word were generated using attention weights, which generated semantically related words.Figure 5shows the architecture of BERT model.

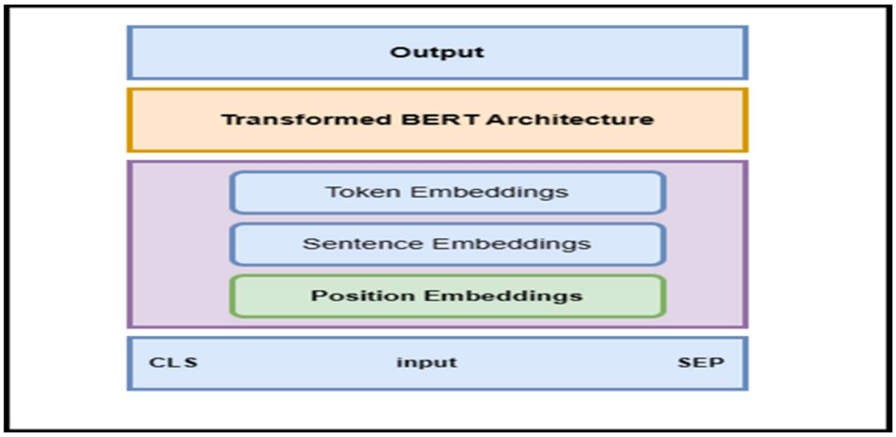


Figure 5: Architecture of BERT Model

3.5.1 Algorithm for BERT

function mapOntologies(sourceOntology, targetOntology, bertModel): mappings = []
for conceptA in sourceOntology: bestMatch = None
maxScore = -Infinity
for conceptB in targetOntology:
score = calculateSimilarity(conceptA, conceptB, bertModel) if score >maxScore:
maxScore = score bestMatch = conceptB
if bestMatch is not None: mappings.append((conceptA, bestMatch))
return mappings
function calculateSimilarity(conceptA, conceptB, bertModel):
textA = preprocessText(conceptA) textB = preprocessText(conceptB)
embeddingA = bertModel.encode(textA) embeddingB = bertModel.encode(textB)
return calculateCosineSimilarity(embeddingA, embeddingB)
function preprocessText(text):
Apply any necessary preprocessing steps such as tokenization, lowercasing, etc. return preprocessedText
function calculateCosineSimilarity(embeddingA, embeddingB):
dotProduct = dotProduct(embeddingA, embeddingB) magnitudeA = sqrt(dotProduct(embeddingA, embeddingA)) magnitudeB = sqrt(dotProduct(embeddingB, embeddingB))
return dotProduct / (magnitudeA * magnitudeB)

3.6 Bi-LSTM Model

Character-level convolutional neural networks are utilized by a bidirectional LSTM network. Sequence input is fed to the first BiLSTM layer. The BiLSTM layer is composed of two passes of LSTM cells, one for processing the input sequence in the forward direction and the other for processing it in the reverse direction. This enables the network to determine the past and future context of each character in the input sequence. In addition, the character level CNNs aid in the extraction of essential features from each character prior to passing it to the BiLSTM layer. The forward pass establishes context prior to the word. The backword pass builds the context following the word. Both stages generate intermediate word vectors, which are then concatenated and processed by a fully connected layer for classification. Several natural language processing tasks, such as sentiment analysis, named entity recognition, and text classification, have exhibited promising results with this method. It excels at processing lengthy text sequences and extracting the contextual information contained within them. However, it requires a considerable quantity of computational resources and may not be appropriate for environments with limited resources. The intermediate word vectors are transmitted to the next BiLSTM layer, layer r, which further refines contextual information and improves classification accuracy. In addition, pre-training the model on a large corpus of text can reduce the quantity of computational resources required for training and improve performance in environments with limited resources. The final word is composed of a weighted sum of raw vectors and two intermediate vectors. The model accounts for polysemy, in which a single word has multiple meanings and senses.Figure 6shows the architecture of Bilstm Model.

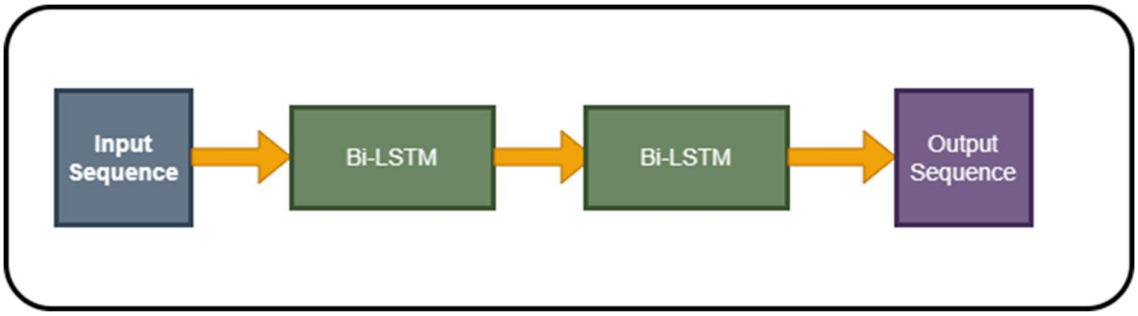


Figure 6: Architecture of Bilstm Model

3.6.1 Algorithm

function ELMoModel(inputSentences, embeddings, lstmHiddenSize): sentenceEmbeddings = []
for sentence in inputSentences: tokenizedSentence = tokenize(sentence)
embeddingIndices = getEmbeddingIndices(tokenizedSentence) lstmOutput = BiLSTM(embeddingIndices, embeddings, lstmHiddenSize) sentenceEmbedding = averagePooling(lstmOutput) sentenceEmbeddings.append(sentenceEmbedding)
return sentenceEmbeddings
function BiLSTM(embeddingIndices, embeddings, lstmHiddenSize): forwardHiddenStates = []
backwardHiddenStates = []
lstmInput = embeddings[embeddingIndices]

```
forwardHiddenState = zeros(lstmHiddenSize) backwardHiddenState = zeros(lstmHiddenSize)
for i from 0 to length(lstmInput) - 1:
forwardHiddenState = LSTMCell(lstmInput[i], forwardHiddenState)
forwardHiddenStates.append(forwardHiddenState)
for i from length(lstmInput) - 1 to 0:
backwardHiddenState = LSTMCell(lstmInput[i], backwardHiddenState)
backwardHiddenStates.prepend(backwardHiddenState)
lstmOutput = concatenate(forwardHiddenStates, backwardHiddenStates) return lstmOutput
function LSTMCell(input, hiddenState): # LSTM cell implementation
# Update gates (input, forget, output)
 inputGate = sigmoid(input * W_i + hiddenState * U_i + b_i) forgetGate = sigmoid(input *
W_f + hiddenState * U_f + b_f) outputGate = sigmoid(input * W_o + hiddenState * U_o +
b_o)
# Candidate hidden state
candidateState = tanh(input * W_c + hiddenState * U_c + b_c)
# Update hidden state
updatedHiddenState = forgetGate * hiddenState + inputGate * candidateState
updatedHiddenState = outputGate * tanh(updatedHiddenState)
return updatedHiddenState function tokenize(sentence):
# Tokenization logic
# Split the sentence into individual tokens or words return tokenizedSentence
function getEmbeddingIndices(tokenizedSentence):
# Get the indices of word embeddings for each token in the sentence
# This could involve looking up the indices in a pre-trained word embedding matrix or
vocabulary
return embeddingIndices
function averagePooling(lstmOutput):
# Perform average pooling over the LSTM output sequence to obtain a fixed-length sentence
embedding
sentenceEmbedding = average(lstmOutput) return sentenceEmbedding
```

4. Results and Discussions

The electronic health record is a CSV file. The EHR contains details regarding Type 1, Type 2, gestational diabetes, medications used, and blood sugar level analysis. This information can be used to analyze trends and patterns in diabetes management and inform patients' individualized treatment plans. In addition, EHR records can be securely shared between healthcare providers to ensure continuity of care and enhance patient outcomes. Therefore, we have organized the data in a structured format to protect and manage the data's security while also facilitating easy access to and retrieval of patient data. In addition, it has been demonstrated that the use of EHRs reduces medical errors and improves the overall quality of healthcare by providing real-time updates on patient status and treatment progress. The structured documents are created using the OWL ontology language, which enables standardized terminology and coding, allowing for improved communication between healthcare providers and enhancing patient outcomes. In addition, EHRs can improve

efficiency and productivity in healthcare settings by eliminating the need for manual data entry and paperwork, thereby freeing up more time for patient care. We have performed three levels of analysis to ensure that the data stored in our EHR system is accurate and reliable. This has resulted in improved decision-making and better patient care, as healthcare providers now have access to comprehensive and current information regarding their patients' medical history, medications, and treatment plans, such as text mapping, concept mapping, and structured mapping, for Type 1 and Type 2 diabetics, with precision, recall, F1-score, and accuracy.

$$Precision = Number\ of\ Concepts\ correctly\ predicted / Total\ number\ of\ positive\ predictions$$

$$Recall = Number\ of\ correctly\ predicted\ positive\ class / Total\ number\ of\ positive\ instances\ in\ the$$

$$F1 - Score = 2 * precision * Recall/(precision + recall)$$

$$Accuracy = Number\ of\ correct\ predictions / Number\ of\ all\ predictions$$

4.1 Text Corpus

A text corpus is a language resource consisting of a large, structured corpus of text that has been collected and analysed for linguistic purposes. Text corpora can be utilised to investigate various linguistic aspects, including syntax, semantics, and discourse patterns. Text corpora can also be used to develop natural language processing algorithms and improve machine learning models for classification and sentiment analysis of text. It typically consists of millions of words or more and may include news articles, novels, and academic papers, among others. Linguists, computer scientists, and other researchers frequently use these corpora to study language usage and develop natural language processing algorithms. They are utilised for statistical analysis and testing of hypotheses, as well as for training machine learning models and enhancing language technologies. In addition, corpora can shed light on social and cultural trends as well as changes in language usage throughout history.Figure 7shows the sample of text corpus for ontological mapping.

"Basugine is a long-acting insulin used for controlling blood sugar levels in people with diabetes.", "Type 1 diabetes, also known as juvenile diabetes, is a chronic condition in which the pancreas produces little or no insulin.", "Basugine injection is administered once daily to provide a steady release of insulin throughout the day.", "Type 2 diabetes is a metabolic disorder characterized by insulin resistance and high blood sugar levels.", "Basugine dosage may vary depending on individual needs and should be determined by a healthcare professional.", "Type 2 diabetes can often be managed through lifestyle changes, medication, and insulin therapy if necessary.", "Basugine is an effective treatment option for both Type 1 and Type 2 diabetes patients.", "People with Type 1 diabetes require lifelong insulin therapy to manage their condition.", "Basugine is a brand of insulin glargine, which is a recombinant human insulin analogue.", "Type 2 diabetes is typically associated with obesity, sedentary lifestyle, and genetic factors."

Figure 7:Sample text corpus of Basugine drug

4.2 Word Embeddings using tSNE

stochastic t-distributed distribution Word embedding (t-SNE) is employed to visualise word embedding in a two-dimensional work space. This method permits the examination of the relationship between words and their proximity in vector space. Additionally, t-SNE can be used to identify clusters of related words within a corpus, revealing their semantic similarities. The illustration depicts word embedding in the diabetic medication basugine. Five words, including basugine, insulin, glucose, injection, and pancreas, are used to analyse the word embeddings: basugine, insulin, glucose, injection, and pancreas. Based on their semantic similarities, the t-SNE algorithm has successfully clustered these words into distinct groups. This visualization can be helpful for comprehending the relationships between words in a corpus and can aid in information retrieval and natural language processing. Each word is analyzed based on its word embeddings, and their cosine similarity y scores are used to group them. This technique is frequently employed in machine learning algorithms and has been demonstrated to be effective in a variety of applications, including sentiment analysis and recommendation systems. Additionally, the resulting clusters can provide insights into the corpus's underlying themes and topics. Outliers are used to eliminate dissimilar words, which increases the clustering's accuracy. In addition, the use of different distance metrics and clustering algorithms can affect the quality of the resulting clusters, so it is essential to select these parameters with care based on the application and data being analyzed.Figure 8 shows the visualization of k-means clustering of word embedding model and Figure 9 shows the results obtained in proximity, semantic relationships of Word Embedding.

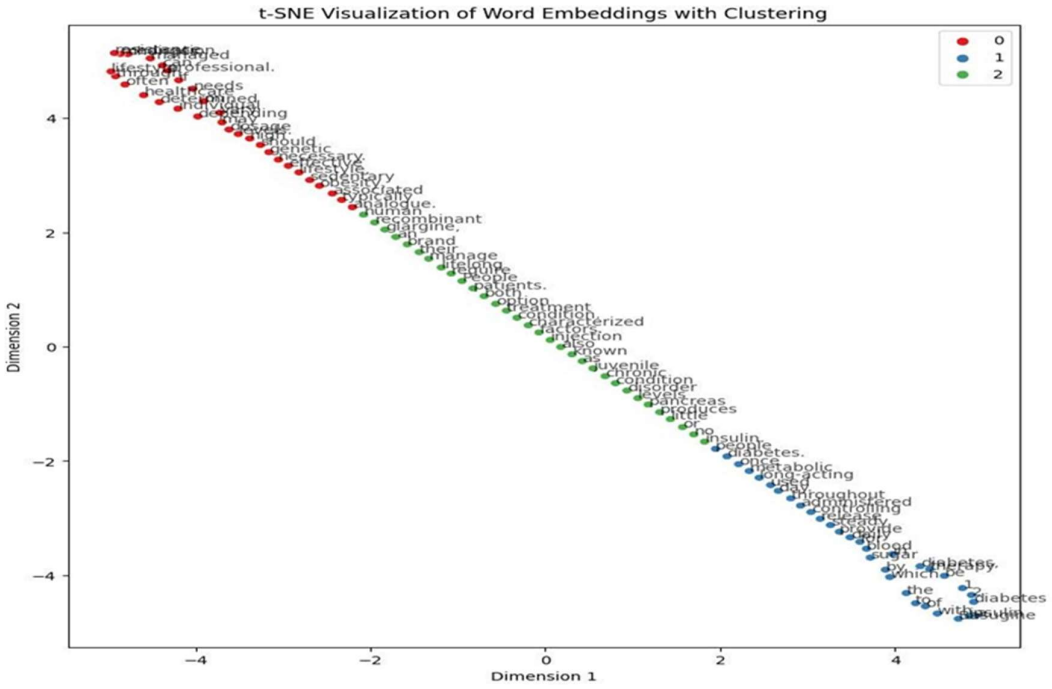


Figure 8:Clustering Analysis of Word Embeddings using K-means Clustering

4.3 Proximity, Semantic relationship

Closest words to daily: ['provide', 'for', 'steady', 'blood', 'release'] Closest words to provide: ['steady', 'daily', 'release', 'for', 'controlling']
Closest words to steady: ['release', 'provide', 'controlling', 'daily', 'administered']

Closest words to release: ['controlling', 'steady', 'administered', 'provide', 'throughout'] Closest words to controlling: ['administered', 'release', 'throughout', 'steady', 'day.'] Closest words to administered: ['throughout', 'controlling', 'day.', 'release', 'used'] Closest words to throughout: ['day.', 'administered', 'used', 'controlling', 'long-acting'] Closest words to day.: ['used', 'throughout', 'long-acting', 'administered', 'metabolic'] Closest words to used: ['long-acting', 'day.', 'metabolic', 'throughout', 'once']
Figure 9Proximity , Semantic Relationships

4.4 Cosine Similarity Measure

Table 1demonstrates the cosine similarity measure between various ontology mappings of various word embedding models, including the word2vector model, CBOW, skip ngram, BERT, and Bilstm models. Ontology mapping employs word embedding models because they map similar words to similar ontology concepts in order to extract implicit meanings from the ontology. Word embedding models extract the semantic relationships between various ontologies. The CBOW model identifies a word from a set of words with a specified windowsize, whereas the skip n-gram model identifies words from the main word. We must predict rare combinations of ontology mappings in our drug ontology, such as Type 1 diabetes, Juvenile diabetes, and basugine dosage. Therefore, the CBOW skip n gram yields a higher degree of similarity between word mappings and ontology. Because it is pretrained on a large

corpus, the BERT model produces the highest similarity of mappings between two words. Bilstm manages sequences of variable length and takes into account the polysemy of words. Thus, Bilstm provides a greater degree of similarity between various ontologies based on the semantic relationships between their various polysemy.

Table 1 Cosine similarity between ontology mapping

Mappings	**Word2Vec CBOW**	**Word2Vec Skip ngram**	**BERT**	**Bilstm**
Basugine, Insulin	0.234	0.456	0.789	0.889
Type 1 Diabetes,Juvenile Diabetes	0.324	0.567	0.687	0.987
Type 2 Diabetes, Metabolic disorder	0.423	0.732	0.678	0.967
Insulin,Blood Sugar Levels	0.543	0.743	0.788	0.887
Basugine, injection	0.423	0.732	0.678	0.967
Insulinresistance,High blood sugar levels	0.567	0.455	0.567	0.768
Basugine,Dosage	0.342	0.654	0.823	0.923
Type 2 diabetes, Lifestyle Changes	0.543	0.743	0.788	0.887
Insulin Therapy,Medication	0.423	0.732	0.678	0.967
Type 1 Diabetes,Lifelong insulin theraphy	0.234	0.456	0.789	0.889

Type 1 Diabetes,Lifelong insulin theraphy 0.234 0.456 0.789 0.889

4.6 Precision, Recall, f1score, and Accuracy

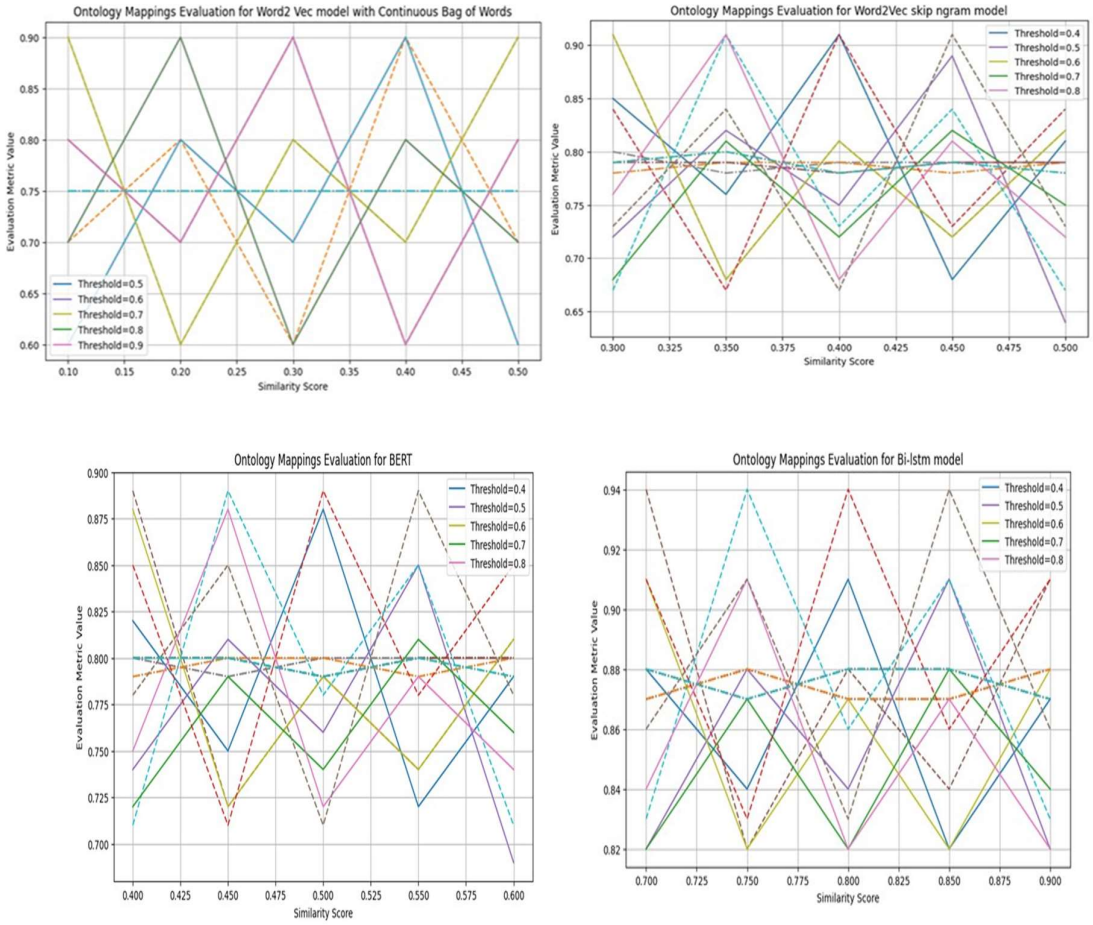


Figure 10. Precision, Recall, F1 score and Accuracy of Ontology mapping

The figure 10illustrates the precision, recall, f1 score, and precision of different ontology mappings with word2vec of continuous bags of words, skip n-gram, BERT, and bilstm models. According to the results, the BERT model outperformed the other models in terms of precision, recall, and f1 score. Among all tested models, the bilstm model achieved the highest level of precision. It is crucial to consider both precision and accuracy when choosing a model for ontology mapping tasks. The bilstm word embedding model achieves the highest precision, recall, f1 score, and accuracy in comparison to other models because it employs the polysemy technique to deconstruct textual words. In addition, the bilstm model has a low computational cost and can manage large datasets efficiently, making it a viable option for ontology mapping tasks.

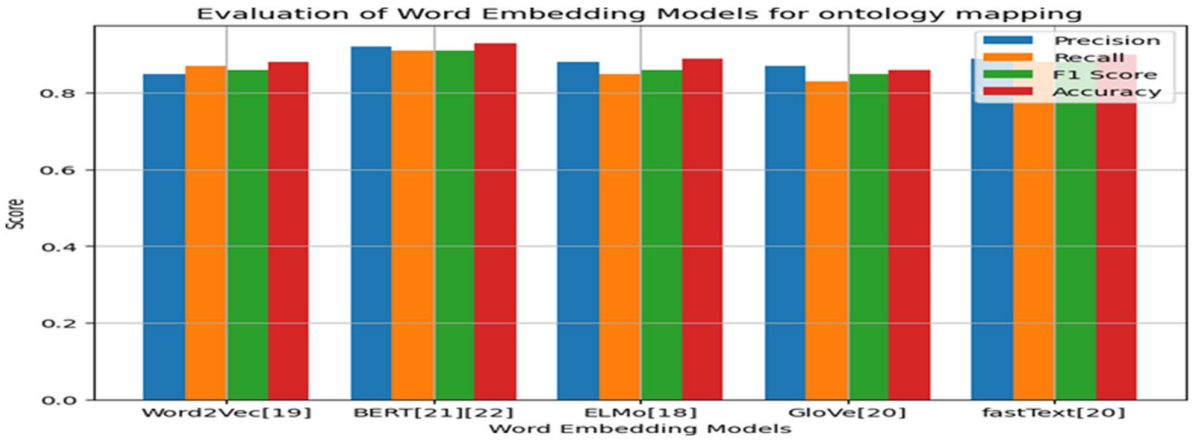


Figure 11.Average Precision, Recall, F1 score and accuracy of Word Embedding Models

The figure 11depicts the precision, recall, f1-score, and accuracy of various word embedding models. When compared to other models such as ELMo[18], Word2vec[19], Glove[20], and Fasttext[20], the BERT[21], [22] model produces higher precision, recall, and accuracy scores.

However, it takes longer to train and requires more computational resources. As a result, the word embedding model selected should be based on the task's specific needs and resources.

5. Conclusion

Drug-based ontology mapping in electronic health records performs well on Word2Vec, BERT, and the Bilstm model, among three other word embedding models. This shows that the ontology mapping method is effective and can be used with different kinds of healthcare data. Additionally, employing multiple word embedding models can boost the ontology mapping system's overall performance and yield more accurate results. In drug ontology mapping, the Bilstm model outperforms the word2vec and BERT models thanks to its polysemy technique, which can handle words with multiple meanings. Ontologies and domain-specific knowledge can also be incorporated to improve the precision and applicability of the ontology mapping system for healthcare data. The model's precision, recall, F1 score, and accuracy all increased. The sequential and contextual information from the input data was used to train the Bilstm model. The model efficiently gathers data on drug associations and semantic relationships to enhance ontology mapping performance. To improve the domain specific context of the specified model, we will eventually integrate domain specific ontologies like medical ontologies, drug databases, and semantic relationships.

References

[1] Y. He, J. Chen, D. Antonyrajah, and I. Horrocks, “BERTMap: A BERT-Based Ontology Alignment System,” 2022. [Online]. Available: www.aaai.org

[2] A. CB, K. Mahesh, and N. Sanda, “Ontology-based semantic data interestingness using BERT models,” Conn Sci, vol. 35, no. 1, 2023, doi: 10.1080/09540091.2023.2190499.

[3] D. S. Asudani, N. K. Nagwani, and P. Singh, “Impact of word embedding models on text analytics in deep learning environment: a review,” ArtifIntell Rev, 2023, doi: 10.1007/s10462- 023-10419-1.

[4] X. Xue, H. Wang, J. Zhang, Y. Huang, M. Li, and H. Zhu, "Matching Transportation Ontologies with Word2Vec and Alignment Extraction Algorithm," J Adv Transp, vol. 2021, 2021, doi: 10.1155/2021/4439861.

[5] M. Poetsch, U. Brisolara Correa, and L. Astrogildo De Freitas, "A Word Embedding Analysis towards Ontology Enrichment." [Online]. Available: http://nilc.icmc.usp.br/embeddings.

[6] J. Youn, T. Naravane, and I. Tagkopoulos, "Using Word Embeddings to Learn a Better Food Ontology," Front ArtifIntell, vol. 3, Nov. 2020, doi: 10.3389/frai.2020.584784.

[7] Y. He, J. Chen, D. Antonyrajah, and I. Horrocks, "Biomedical Ontology Alignment with BERT," 2021. [Online]. Available: http://www.cs.ox.ac.uk/isg/projects/SEALS/oaei/.