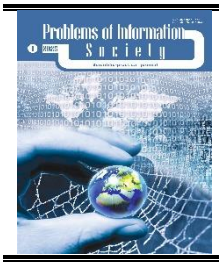

www.jpis.az

Application of the word2vec algorithm for clinical diagnosis determination

Uzeyir Gurbanli

Institute of Control Systems, B. Vahabzadeh str., 68, AZ1141, Baku, Azerbaijan

uzeyir@sinam.net

ARTICLE INFO

http://doi.org/10.25045/jpis.v16.i1.03

Article history:
Received 09 september 2024
Received in revised form
10 November 2024
Accepted 17 January 2025

Keywords:

Natural Language Processing
Diagnosis prediction International Classification of Diseases
Hospital İnformation System,
Artificial intelligence
Health Level Seven
Electronic Healthcare Records

ABSTRACT

The article examines the use of Natural Language Processing technologies in modern medicine for knowledge acquisition and the implementation of decision-making methods. The application of information technology in healthcare has become one of the main requirements of the modern era. Enhancing the quality and accessibility of medical services necessitates the utilization of modern technologies, mathematical methods, and the capabilities of artificial intelligence, alongside the development of comprehensive information systems. The paper proposes methods for analyzing and applying tools to assist physicians in diagnosing conditions and determining treatment plans with the help of artificial intelligence. It also focuses on evaluating the quality of diagnostic and treatment processes through applıcation of different methods and practical application. The analysis of large volumes of medical data using Natural Language Processing technology enables the extraction of valuable insights. A significant portion of medical data is stored and exchanged in text form compliant with the Health Level Seven standard, making semantic similarity methods that operate on textual data highly effective in this domain. By designing and implementing rules for applying and integrating different algorithms, it is possible to transform medical data into valuable knowledge, contributing significantly to advancements in the medical field. The article presents a Word2Vec algorithm-based approach for detecting diagnoses of cardiovascular diseases from collected patient histories, as well as refining existing diagnoses. The development of an algorithm capable of assigning new diagnoses based on historical patient records constitutes one of the key outcomes of this research.

1. Introduction

The rapid growth in the volume of digital data due to the application of technologies in modern medicine necessitates the effective analysis of this data, the extraction of knowledge from collected information, and the making of accurate decisions. Medical data is primarily recorded in text, image, and video formats(Masuma H. Mammadova, 2016), and the individual analysis of these types of data offers new perspectives in the field of e-health (fig.1.). 60-70% of medical data is in a Health Level Seven (HL7) text format. In particular, Natural Language Processing(NLP) algorithms hold significant importance in analyzing textual data, as medical records, patient histories, diagnostic results, and clinical notes are predominantly stored and processed in text format. The application of Artificial Intelligence(AI) and machine learning methods is critical for the analysis, processing, and extraction of

knowledge from such data. In the medicine, the accuracy and efficiency of clinical diagnostics, the proper assessment of a patient's health status, and the optimization of treatment processes are of vital importance. The volume, structure, and nature of medical data require the application of new analytical methods for their processing and analysis. Medical information, such as patient records, laboratory results, clinical reports, and treatment plans, is often in an unstructured text format, forming extensive clinical databases. To maximize the benefits of this data, AI and NLP technologies are being extensively studied and applied (M.D. Devika et al., 2016).

NLP algorithms like Word2Vec have the capability to transform unstructured data into meaningful formats by learning semantic relationships between terms in medical texts. By representing medical terminology in a multidimensional vector space, this method allows for a deeper understanding of semantic relationships between words. Beyond learning connections among diseases, symptoms, medical procedures, and medications, the Word2Vec model also facilitates its application in predicting and diagnosing diseases at an early stage. The use of NLP methods alongside statistical and machine learning algorithms in medical data processing plays a crucial role in clinical decision-making and the prevention of diagnostic errors. These technologies enable the automated processing of patient records, analysis of relationships between symptoms and diseases, and the preparation of personalized diagnostic predictions.

Word2vec is a predictive algorithm that operates through two primary models: the continuous bag-of-words (CBOW) and the skip-gram (SG) approaches (Antonio, 2022). Both rely on shallow neural networks to project words into a vector space, as described in prior studies. The distinction lies in their objective: CBOW predicts a target word based on its surrounding context, while SG predicts the context from a given word.

Key parameters in training word2vec embeddings include the number of dimensions in the embedding space, generally ranging from 50 to 500 and determined through experimental tuning, and the size of the context window, which defines the span of words considered around the target word (commonly set to 5 or 10). Additional hyperparameters, discussed in supporting materials, can also influence performance. Larger embeddings require more training data but can capture nuanced semantic features, as each dimension encodes specific aspects of meaning (Faiza et al., 2019).

The advantage of the Word2Vec model lies in its ability to easily identify meaningful relationships within unstructured text, thereby accelerating the analysis of complex medical data. This article explores the application of the Word2Vec algorithm in medical diagnostics, demonstrating how artificial intelligence is revolutionizing the medical field. It highlights the synergies created by combining medical data processing, statistical analysis, and AI technologies in medical decision-making. Based on advanced global research and case studies, the article evaluates the key results and benefits of using the Word2Vec algorithm.

Data types in e – Health is shown in fig. 1.

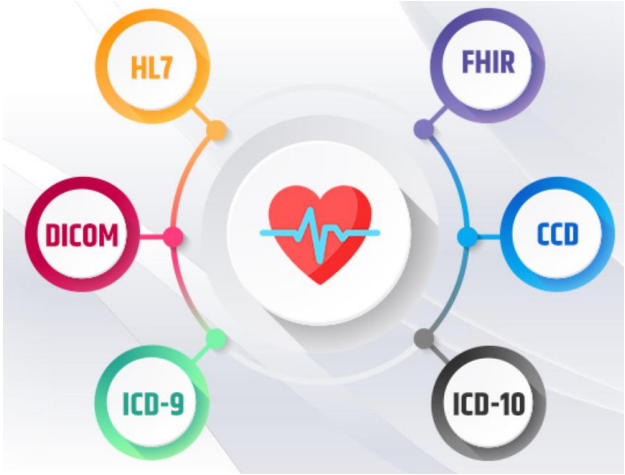

Fig. 1. Data types in e – Health

2. Related works

The integration of AI and NLP algorithms into clinical diagnostic systems has led to significant advancements in extracting knowledge from large volumes of medical data and improving diagnostic accuracy(Young, 2018). The use of the Word2Vec model in medical research has opened new avenues for analyzing relationships between diseases and symptoms and automating clinical decision-making processes. Key global research directions on the application of this model are outlined below:

Analysis of Electronic Health Records (EHRs):

Numerous studies have utilized the Word2Vec algorithm to extract relevant clinical information from EHRs and understand relationships between diseases. For instance, a study conducted by the Mayo Clinic analyzed text data from patients medical records using the Word2Vec model (Faiza et al., 2019). The algorithm learned connections between medical terms, providing precise results

in identifying rare diseases and determining symptom clusters. This support helped clinicians in decision-making, reducing diagnostic time and minimizing errors.

Enhancing Cancer Diagnosis and Prognosis:

The application of Word2Vec and other NLP algorithms has demonstrated high efficiency in the early diagnosis of complex diseases such as cancer. In collaborative studies conducted by Massachusetts General Hospital and Harvard Medical School, the Word2Vec algorithm was applied to radiological reports and patient history records. The algorithm analyzed disease descriptions and tumor markers, aiding in identifying cancer risk factors and, in some cases, reducing the need for invasive diagnostic procedures. Its risk analysis based on patients medical history improved the accuracy and reliability of clinical decisions (Mikolov et al., 2013).

Detection of Symptom-Diagnosis Pairs Using NLP: Advanced AI platforms like IBM Watson Health have implemented Word2Vec and other NLP algorithms for automating clinical diagnoses. The IBM Watson platform analyzes symptoms in patients' medical records using the Word2Vec model and identifies symptom-diagnosis pairs (Jiho et al., 2021). Such NLP models enable a more precise understanding of text-based medical terminologies, generating detailed and useful insights about the patient.

Disease Risk Modeling and Prediction: Extensive research by Renmin University of China applied the Word2Vec model to medical databases to model and predict disease risks. These studies analyzed potential risk factors based on patients' medical histories and symptoms. The Word2Vec algorithm explored connections among genetic, lifestyle, and symptomatic factors of diseases, aiding in the early detection of high-risk patients and enabling timely and effective preventive measures(Bofang et al., 2019).

Automating Diagnoses and Improving Patient Care Practices: Studies at Stanford University School of Medicine explored the use of Word2Vec and other NLP models to analyze patient data and create integrated decision-support systems for diagnostics. These systems, developed using the Word2Vec model, enable automated diagnosis based on symptom analysis. Such technologies ensure quicker and more accurate analysis of patient data, facilitating higher-quality service for both clinical staff and patients.

The results of these studies indicate that applying NLP algorithms like Word2Vec in clinical diagnostics significantly improves early detection of diseases, minimizes diagnostic errors, and optimizes the decision-making process for clinicians. The semantic analysis capabilities of these algorithms enable the automated processing of clinical data and the structuring of medical records, thereby expanding the applications of AI in medicine.

The analysis of the aforementioned research approaches and literature suggests a vast potential for further studies utilizing NLP algorithms, particularly Word2Vec, to derive valuable insights from existing medical data. Current studies primarily focus on identifying relationships between medical terms and predicting specific diseases. The analysis and structuring of patient anamneses—initial clinical complaints—play a crucial role in establishing preliminary diagnoses across all medical fields. Early diagnosis is one of the most critical elements in healthcare, typically performed by general practitioners in developed countries.

By applying NLP algorithms, preliminary diagnoses could be immediately identified through information systems or applications. Leveraging historical diagnostic data based on patient anamneses enables the discovery of new diagnoses from recent anamneses and the refinement of previously established diagnoses. This article focuses on developing scientific methods and validating results for diagnosing clinical conditions using the Word2Vec algorithm. It presents research findings and their implementation based on statistically validated data. The processing of medical histories collected in the Azerbaijani language and the proposal of new diagnoses based on their analysis hold significant importance. The use of the mentioned methods and algorithms for processing unstructured medical histories presented in different dialects and obtaining results is one of the main objectives of the article.

3. Materials and methods

To demonstrate the effectiveness of the applied algorithm, a small example was prepared, and corresponding software was developed using the C# programming language. In the example, two diagnoses based on patient anamneses were used as statistical data. The determination of diagnoses was based on the ICD-10 international classification system. The methodology for predicting diagnoses for newly received

anamneses was described based on the diagnoses determined from the two anamneses. The software was developed in accordance with this example. The same software is capable of conducting trials with a larger dataset. It is planned to obtain real data from hospitals and conduct trials on a large volume of data.

3.1. Medical data analysis

Nowadays, the application of information technologies in modern medicine has led to the generation of various types of data in different formats. Diverse systems and medical devices produce and process data in multiple formats. Each of these data types contains valuable knowledge, which can be extracted through proper processing. The primary types of data used for information exchange in medicine are as follows:

HL7 : HL7 provides frameworks and standards for the exchange, integration, sharing, and retrieval of electronic health information. Key standards include:

- HL7 v2: Widely used for messaging between healthcare applications.
- HL7 v3: Focuses on XML-based data exchange.

Fast Healthcare Interoperability Resources (FHIR): A modern standard designed to enable quick and flexible data exchange via RESTful APIs.

Digital Imaging and Communications in Medicine(DICOM): Standard for storing, transmitting, and sharing medical imaging information (e.g., X-rays, CT scans, MRIs).

International Classification of Diseases(ICD): Maintained by the World Health Organization (WHO), ICD is used globally for coding diagnoses and health conditions.

- ICD-10: Current widely used version.
- ICD-11: Latest version with enhanced digital capabilities.

However, anamneses are unstructured data (table 2.) represented solely in text form. Processing, structuring, and extracting valuable knowledge from these data require analyzing information from various sources. Anamneses are typically generated in Hospital Information Systems (HIS) and digital twin systems (Ahmadova, 2024) based on data entered by physicians regarding patient complaints. These anamneses are entered and stored in text form within the system. To detect diagnoses in accordance with the ICD classification (table 2.) based on anamneses, the operational principles of the Word2Vec method, one of the NLP algorithms, are examined.

Table 1. ICD 10 diagnosis data structure

S50.0	AZ: Dirsəyin əzilməsi EN: Contusion of elbow
S57.0	AZ: Dirsək oynağının bərk əzilməsi EN: Crushing injury of elbow joint
Q37.1	AZ: Sərt damağın yarığı dodağın birtərəfli yarığı ilə EN: Cleft hard palate with unilateral cleft lip
S70.1	AZ: Budun əzilməsi EN: Contusion of thigh
S77.0	AZ: Bud-çanaq oynağının bərk əzilməsi EN: Crushing injury of hip and thigh
S81	AZ: Baldırın açıq yarası EN: Open wound of lower leg
S84.0	AZ: Qamış sinirinin baldır səviyyəsində travması EN: Injury of tibial nerve at lower leg level
S91	AZ: Aşıq-baldır oynağı və ayaq nahiyəsinin açıq yarası EN: Open wound of ankle and foot
S92.3	AZ: Ayaq daraq sümüklərinin sınığı EN: Fracture of metatarsal bones

Table 2. Unstructured anamnesis database

AZ: Yüksək təzyiq və baş ağrısı EN: High blood pressure and headache
AZ: Nəfəs darlığı, yeriyəndə təngnəfəs olma EN: Shortness of breath, breathlessness while walking
AZ: Pilləkən qalxanda dizlərdə ağrı EN: Knee pain while climbing stairs
AZ: Gözdə iynə batma effekti, gözün quruluğu və yandırması EN: Needle-piercing sensation in the eye, eye dryness, and burning
AZ: Yeməkdən sonra mədədə küt ağrılar, iştahsızlıq EN: Dull stomach pain after meals, loss of appetite
AZ: Baş gicəllənmə, gözlərin qaralması, huşunu itirmə EN: Dizziness, blurred vision, fainting
AZ: Əl və ayaqların üşüməsi, davamlı yüksək temperaturun müşahidə olunması EN: Coldness in hands and feet, persistent high temperature
AZ: Mədə bulanması, qusma və halsızlıq EN: Nausea, vomiting, and weakness

A key component of NLP algorithms, Word2vec is used for numerical computations to identify similarities and semantic relationships between texts based on textual data. The word2vec algorithm operates with two main approaches (Ruder et al., 2019):

CBOW: This model analyzes the context around a given word and attempts to predict that word (Zhang, et al, 2010). For example, in the sentence "My name is Uzeyir," the words "My" and "is" are used as context to predict the word "name."

SG: In contrast, this model tries to predict the context around a given word. For example, in the sentence "My name is Uzeyir," the words "My" and "is" are predicted as context for the word "name."

Both of these models work by representing words in vector form in textual data (Li et al., 2018). Vectors help to learn the semantic relationships and similarities between words in texts. The Word2vec model is used in text processing, machine translation, text classification, and other natural language processing applications. The main purpose of the algorithm is to better understand the meanings of words in textual data and to be more effective in text processing. The main stages of the algorithm are as follows:

Data Collection: As a first step, medical information about patients needs to be collected from medical information systems. This information should include diseases, symptoms, treatments, prescriptions, and other medical records.

Data Preparation: The collected texts need to be cleaned and prepared. This involves removing unnecessary symbols and punctuation marks from the texts, converting texts to lowercase, and tokenizing them, i.e., splitting them into words.

Model Training: The prepared text data is used to train the word2vec model. The word2vec model converts words into vectors, where these vectors represent the meanings of words in numerical form. The model learns relationships between words and represents similar-meaning words with close vectors.

Diagnosis Giving: The symptoms recorded in the patient's history are represented with vectors learned by the word2vec model (Abdelhakim et al., 2020). For example, symptoms like "headache" and "fatigue" are converted into certain vectors. Similarity is calculated between the symptom vectors of the patient and the existing disease vectors. This can be done using methods like cosine similarity. The disease vectors showing the closest similarity represent potential diagnoses.

Diagnosis Recommendation: Diseases showing the highest similarity are suggested to the physician as a diagnosis. Considering this information, the physician further refines the diagnosis with additional tests and examinations.

3.2. Application of Word2vec method

Let's now look at the process of diagnosing based on histories using the word2vec algorithm through an example.

Let's examine a sample anamnesis written in Azerbaijan language taken from the HIS system:

Anamnesis1(AZ): “Xəstə baş ağrısı, yüksək təzyiq və yorğunluqdan şikayətlənir”.

Anamnsis1(EN):"The patient complains of headache, high blood pressure, and fatigue."

Anamnesis2(AZ): “Xəstə quru öskürək, sinə ağrısı və nəfəs darlığından şikayətlənir”.

Anamnesis2(EN): "The patient complains of dry cough, chest pain, and shortness of breath."

To apply this history using the word2vec CBOW method, it is necessary to first separate the words in the sentence and remove punctuation:

Anamnesis 1(AZ): ['Xəstə', 'baş', 'ağrısı', 'yüksək', 'təzyiq', 'yorğunluq', 'şikayətlənir']

Anamnesis 1(EN): ['Patient', 'head', 'pain', 'high', 'pressure', 'fatigue', 'complains']

Anamnesis 2(AZ): ['Xəstə', 'quru', 'öskürək', 'sinə', 'ağrısı', 'nəfəs', 'darlığı', 'şikayətlənir']

Anamnesis 2(EN): ['Patient', 'dry', 'cough', 'chest', 'pain', 'shortness', 'of', 'breath', 'complains']

Next, vectors need to be created for both anamnesis. For example, a vector consisting of 3 random numbers is given for illustration.

Vectorization of sample anamneses is shown in table 3.

Table 3. Vectorization of sample anamneses

Vector created for Anamnesis1	Vector created for Anamnesis2
In Azerbaijani Language	
"Xəstə" -> [0.1, 0.2, 0.3] "baş" -> [0.2, 0.1, 0.4] "ağrısı" -> [0.3, 0.1, 0.2] "yüksək" -> [0.4, 0.5, 0.2] "təzyiq" -> [0.5, 0.4, 0.1] "yorğunluq" -> [0.2, 0.3, 0.5] "şikayətlənir" -> [0.1, 0.3, 0.4]	"Xəstə"-> [0.1, 0.2, 0.3] "quru" -> [0.4, 0.1, 0.5] "ağrısı" -> [0.3, 0.1, 0.2] "öskürək"->[0.3, 0.5, 0.4] "sinə" -> [0.5, 0.4, 0.2] "darlığı" -> [0.2, 0.1, 0.5] "şikayətlənir"-> [0.1, 0.3, 0.4]
In English	
"Patient" -> [0.1, 0.2, 0.3] "head" -> [0.2, 0.1, 0.4] "Pain" -> [0.3, 0.1, 0.2] "high" -> [0.4, 0.5, 0.2]	"Patient"-> [0.1, 0.2, 0.3] "dry" -> [0.4, 0.1, 0.5] "pain" -> [0.3, 0.1, 0.2] "cough"->[0.3, 0.5, 0.4]

"pressure" -> [0.5, 0.4, 0.1] "fatigue" -> [0.2, 0.3, 0.5] "complains" -> [0.1, 0.3, 0.4]	"chest" -> [0.5, 0.4, 0.2] "tightness" -> [0.2, 0.1, 0.5] "complains"-> [0.1, 0.3, 0.4]

Calculation of the mean vector for Anamnesis1:
Mean vector 1(A) = ([0.1, 0.2, 0.3] + [0.2, 0.1, 0.4] + [0.3, 0.1, 0.2] + [0.4, 0.5, 0.2] + [0.5, 0.4, 0.1] + [0.2, 0.3, 0.5] + [0.1, 0.3, 0.4]) / 7 = [0.257, 0.271, 0.3]
Calculation of the mean vector for Anamnesis2:
Mean vector 2(B) = ([0.1, 0.2, 0.3] + [0.2, 0.1, 0.4] + [0.3, 0.1, 0.2] + [0.4, 0.5, 0.2] + [0.5, 0.4, 0.1] + [0.2, 0.3, 0.5] + [0.1, 0.3, 0.4]) / 7 = [0.271, 0.243, 0.357]
Establishing a neural network for predicting diagnoses involves associating the vector corresponding to the given history with previously assigned ICD diagnoses:
X (Input) = [[0.257, 0.271, 0.3], [0.271, 0.243, 0.357] ...]
Y (Output)(AZ) = ["Hipertoniya", Pnevmaniya"...]
Y (Output)(EN) = ["Hypertension", Pneumonia"...]
Once diagnoses corresponding to any history are learned by the models, it becomes possible to predict the diagnosis for newly entered histories. The following example demonstrates the determination of the diagnosis corresponding to a newly entered anamnesis.
Anamnesis3(AZ) = Xəstə yüksək təzyiq və baş ağrısı yaşayır.
Anamnesis3(EN) = The patient experiences high blood pressure and headache.

New anamnesis(AZ): ['Xəstə', 'yüksək', 'təzyiq', 'baş', 'ağrısı', 'yaşayır']
New anamnesis(EN): ['Patient', 'high', 'blood', 'pressure', 'headache', 'experiences']

Vector = [0.1, 0.2, 0.3], [0.4, 0.5, 0.2], [0.5, 0.4, 0.1], [0.2, 0.1, 0.4], [0.3, 0.1, 0.2], [0.2, 0.3, 0.4]
Mean vector 3(C) = ([0.1, 0.2, 0.3] + [0.4, 0.5, 0.2] + [0.5, 0.4, 0.1] + [0.2, 0.1, 0.4] + [0.3, 0.1, 0.2] + [0.2, 0.3, 0.4]) / 6 = [0.283, 0.267, 0.25]
Comparison of the new history's average vector with the vectors in the trained model is conducted using cosine similarity. The comparison based on cosine similarity is performed as follows:

$$\cos\_sim(C,A) = \frac{CA}{||C||||A||} = \frac{\sum_{i=1}^{n} C_i A_i}{\sqrt{\sum_{i=1}^{n} C_i^2}\sqrt{\sum_{i=1}^{n} A_i^2}} =$$

$$\frac{(0.283*0.257+0.267*0.271+0.25*0.3}{\sqrt{0.283^2+0.267^2+0.25^2}*\sqrt{0.257^2+0.271^2+0.3^2}} \frac{0.220}{0.214*0.223} =$$

$$\frac{0.220}{0.232} = 0.845$$

$$\cos\_sim(C,B) = \frac{CB}{||C||||B||} = \frac{\sum_{i=1}^{n} C_i B_i}{\sqrt{\sum_{i=1}^{n} C_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

$$\frac{(0.283*0.271+0.267*0.243+0.25*0.357}{\sqrt{0.283^2+0.267^2+0.25^2}*\sqrt{0.271^2+0.243^2+0.357^2}} \frac{0.220}{0.214*0.223} =$$

$$\frac{0.231}{0.235} = 0.351$$

As a result, the calculated average vector for Anamnesis1 is closer to the average vector of the new history. Therefore, the diagnosis corresponding to Anamnesis1 will be assigned to the new anamnesis.

Below is the program code written in the C# programming language for the given example (fig. 2):

```
// Statistics anamnesis samples
var data = new[]
{
    new AnamnezData { Text = "Xəstə baş ağrısı, yüksək təzyiq və yorğunluqdan şikayətlanir" },
    new AnamnezData { Text = "Xəstə quru öskürək, sinə ağrısı və nəfəs darlığından şikayətlanir" }
};
//Statistical diagnoses according to anamneses
var diaqnozlar = new List<string>
{
"Hipertoniya",
"Astma"
};
// Data loading
var dataView = context.Data.LoadFromEnumerable(data);
// Data converting (text -> vector)
var pipeline = context.Transforms.Text.FeaturizeText("Features", "Text");
var model = pipeline.Fit(dataView);
var transformedData = model.Transform(dataView);
// Model training
var vectors = context.Data.CreateEnumerable<AnamnezVector>(transformedData,
    reuseRowObject: false).ToArray();
// Creation of a sample vector for a new anamnesis
var newAnamnez = new AnamnezData { Text = "Xəstə yüksək təzyiq, və baş ağrısı yaşayır" };
var newDataView = context.Data.LoadFromEnumerable(new[] { newAnamnez });
var newTransformedData = model.Transform(newDataView);
var newVector = context.Data.CreateEnumerable<AnamnezVector>(newTransformedData,
    reuseRowObject: false).First().Features;
// Comparison of vectors using cosine similarity
var similarity1 = CosineSimilarity(newVector, vectors[0].Features);
var similarity2 = CosineSimilarity(newVector, vectors[1].Features);
//Console.WriteLine($"{newAnamnez} {data[1]}: {similarity1}");
Console.WriteLine($"Kosinus oxşarlığı (Yeni anamnez, Anamnez1): {similarity1}");
Console.WriteLine($"Kosinus oxşarlığı (Yeni anamnez, Anamnez2): {similarity2}");
if (similarity1 > similarity2)
{
    Console.WriteLine("Yeni anamnez Anamnez1-ə daha çox uyğundur. Diaqnoz:" + diaqnozlar[0]);
}
else
{
    Console.WriteLine("Yeni anamnez Anamnez2-ya daha çox uyğundur. Diaqnoz:" + diaqnozlar[1]);
}
```

Fig 2. C# code sample for word2vec method

The comparison of anamnesis using cosine similarity is performed with the following cosine similarity algorithm (fig. 3):

```
// cosine similarity method
2 references
static double CosineSimilarity(float[] vectorA, float[] vectorB)
{
    double dotProduct = 0;
    double normA = 0;
    double normB = 0;
    for (int i = 0; i < vectorA.Length; i++)
    {
        dotProduct += vectorA[i] * vectorB[i];
        normA += Math.Pow(vectorA[i], 2);
        normB += Math.Pow(vectorB[i], 2);
    }
    return dotProduct / (Math.Sqrt(normA) * Math.Sqrt(normB));
}
```

Fig 3. Cosine similarity C# code sample

Given the history *"Xəstə yüksək təzyiq,quru öskürək, sinə ağrısı və baş ağrısı yaşayır(The patient experiences high blood pressure, dry cough, chest pain, and headache)"* the result would be as follows (Table 4):

Table 4. Diagnosis determination sample 1

Kosinus oxsarligi (Yeni anamnez, Anamnez1): 0,425023893534168 Kosinus oxsarligi (Yeni anamnez, Anamnez2): 0,591599575713413 Yeni anamnez Anamnez2-ə daha çox uygundur. Diaqnoz:Astma

Given the history *"Xəstə yüksək təzyiq, və baş ağrısı yaşayır (The patient experiences high blood pressure and headache)"* the result would be as follows (Table 5):

Table 5. Diagnosis determination sample 2

Kosinus oxsarligi (Yeni anamnez, Anamnez1): 0,578744274179744 Kosinus oxsarligi (Yeni anamnez, Anamnez2): 0,305322666936385 Yeni anamnez Anamnez1-ə daha çox uygundur. Diaqnoz:Hipertoniya

It is possible to develop such programs in any programming language of choice (Oubenali, N., Messaoud, S., Filiot, A. et al., 2022).

4. Discussion

The application of the Word2Vec method based on patient anamneses and diagnosed conditions to detect new diagnoses demonstrates that the algorithm is generally result-oriented and suitable for this purpose. Different vector dimensions and values for each character can be determined. For the example provided, a 3-dimensional vector was utilized, but larger datasets may benefit from higher-dimensional vectors (e.g., 50-100 dimensions) to detect similarities in anamneses effectively. Smaller dimensions may suffice for limited semantic datasets, though computational resources and time increase with vector size.

One limitation of Word2Vec is its computational speed, particularly with semantically broad datasets requiring high-dimensional vectors. Optimization may involve testing with specific-domain databases and adapting vector sizes based on technical resource constraints. If satisfactory results are achieved, reducing vector dimensions is advisable.

Compared to other NLP algorithms like Naive Bayes and Support Vector Machines, which excel in text classification, Word2Vec is better suited for identifying semantic text similarities. Models like BERT and GPT-3 are also viable for similarity detection, but Word2Vec remains more effective for simple, focused similarity analyses and knowledge extraction from large, unstructured datasets.

5. Conclusion

The research findings indicate the effectiveness of using NLP algorithms for detecting clinical diagnoses and refining previously identified diagnoses. The experiments on sample data demonstrate that it is possible to identify diagnoses based on clinical anamneses. To achieve this, statistical data from various patients' medical histories—typically stored in EHR systems—should be utilized. The same method can also be applied to assign prescriptions based on diagnoses.

In the examples provided in the article, the model trained on medical histories given in Azerbaijani and their corresponding diagnoses is later used to identify diagnoses for newly entered medical histories. Here, the language of the database does not matter. Since the Word2Vec algorithm converts words in a sentence into vectors and checks their proximity through mathematical calculations, the language of the words does not affect the results. This demonstrates that the proposed algorithm can be applied to a database in any language.

Future studies will focus on broadening research into clinical diagnosis identification and exploring methods for prescription assignment. The results highlight that applying NLP algorithms plays a significant role in advancing modern medicine and improving efficiency through technology integration. Developing specialized software and continuing investigations in this area can greatly contribute to improving public health and accessibility to medical services globally and nationally.

References

Abdelhakim, A. E., Elhoseny, M., & Farouk, A. (2020). Hybrid Intelligent Framework for Word2Vec-Based Sentiment Analysis Using Gated Recurrent Unit Network and Particle Swarm Optimization. IEEE Access, 8:152385–152397.

Antonio Desai, Aurora Zumbo, Mauro Giordano (2022). Word2vec Word Embedding-Based Artificial Intelligence Model in the Triage of Patients with Suspected Diagnosis of Major Ischemic Stroke: A Feasibility Study. International Journal of Environmental Research and Public Health. https://doi.org/10.3390/ijerph192215295

Aytan Ahmadova (2024). Applications of digital twins in medicine and the ontological model of medical digital twins. Problems of Information Society, 15(1):98-105. http://doi.org/10.25045/jpis.v15.i1.10

Bofang Li, Aleksandr Drozd, Yuhe Guo, Tao Liu, Satoshi Matsuoka & Xiaoyong Du (2019). Scaling Word2Vec on Big Corpus, Data Science and Engineering, 4:157-175. https://link.springer.com/article/10.1007/s41019-019-0096-6

Faiza Khan Khattak, Serena Jeblee, Chloé Pou-Prom, Mohamed Abdalla, Christopher Meaney, Frank Rudzicz (2019). A survey of word embeddings for clinical text. Journal of Biomedical Informatics, 100:1-18. https://doi.org/10.1016/j.yjbinx.2019.100057

Jiho Noh, Ramakanth Kavuluru (2021). Improved biomedical word embeddings in the transformer era. Journal of Biomedical Informatics, 120:1-11. https://doi.org/10.1016/j.jbi.2021.103867

Li, Y., & Yang, T. (2018). Word embedding for understanding natural language: a survey. Guide to big data applications, 83-104.

Devika M.D., Sunitha C., Ganesh Amal (2016). Sentiment Analysis: A Comparative Study on Different Approaches. Procedia Computer Science, 87: 44-49. https://doi.org/10.1016/j.procs.2016.05.124

Mammadova Masuma H. (2016). Bıg data ın electronıc medıcıne: opportunıtıes, challenges and perspectıves. Problems of Information Technology, 7(2):8/24 https://doi.org/10.25045/jpit.v07.i2.02

Mammadova, M.H., & Jabrayilova, Z.G. (2019). Electronic medicine: formation and scientific-theoretical problems. Baku: "Information Technologies" publishing house, 1-318. https://ict.az/uploads/files/E-medicine-monograph-IIT-ANAS.pdf

Mikolov T., Sutskever I., Chen K., Corrado G., Dean J. (2013). Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems, 1-9. https://doi.org/10.48550/arXiv.1310.4546

Oubenali, N., Messaoud, S., Filiot, A. et al. (2022). Visualization of medical concepts represented using word embeddings: a scoping review. BMC Med Inform Decis Mak 22(83):1-14. https://doi.org/10.1186/s12911-022-01822-9

Ruder, S., Vulić, I., & Sogaard, A. (2019). A survey of cross-lingual word embeddingmodels. Journal of Artificial Intelligence Research, 65:569-631.

Young, T., Hazarika, D., Poria, S., & Cambria, E. (2018). Recent Trends in Deep Learning Based Natural Language Processing. IEEE Computational Intelligence Magazine, 13(3):55–75. https://doi.org/10.1109/MCI.2018.2840738

Zhang, Y., Jin, R., & Zhou, Z.-H. (2010). Understanding Bag-of-Words Model: A Statistical Framework. IEEE Transactions on Image Processing, 19(4):944–963.