Development of public health decision-making based on semantic evaluation of EHR data

Dr. Priya Vij[1], Patil Manisha Prashant[2]

[1]*Assistant Professor, Department of CS & IT, Kalinga University, Raipur, India*
[2]*Research Scholar, Department of CS & IT, Kalinga University, Raipur, India*

KEYWORDS

Public health, electronic health record (EHR), semantic, named entity, decision-making, cubic support vector machine with dipper throated optimization (CSVM-DTO

ABSTRACT

The study examines the way semantic analysis of data from electronic health records (EHRs) is performed. Implementing machine learning (ML) is challenging, EHRs are essential for acquiring medical data because they include textual information from physicians about patients' ailments and treatment plans, supporting well-informed public health decision-making. The goal of the study was to allow computers to semantically comprehend assessments, physical parts, indicators, and therapies in addition to implicitly identifying medical terms in EHRs. To efficiently identify medical phrases in electronic health records, a novel cubic support vector machine with a dipper-throated optimization (CSVM-DTO) strategy is suggested. This study uses tokenization as a pre-processing step for the data gathered from public sources. The semantic data of medical terminology is then extracted using the word-2-vector technique. Next, the named entities are found using the CSVM, and their efficiency is improved by implementing the DTO technique. Based on the experimental results, we can conclude that our suggested approach outperformed other current approaches in locating the identified entities inside the EHRs.

1. Introduction

Long-term health is a consequence of a complex interplay of genetics, lifestyle, environment, and diet, as well as persistent rejection of behaviors that compromise one's health [3]. In recent years, hospital or outpatient electronic health records (EHRs) have contained information about risk factors in addition to other health and/or disease-related data [1]. Patients are becoming more and more eager to provide healthcare providers with better and more data [6]. The objective is to compromise an arrangement for assessing natural language processing (NLP) methods in the background of public health based on past executions [2]. A key objective in natural language processing (NLP) is NER. In artificial intelligence (AI), the field of NLP aims to create computations and framework that can utilize information in the same way individuals interact [12]. NLP allows the inspection and withdrawal of data from unorganized foundations, as well as the mechanization of question-answering, sentiment examination, and text summarization [4]. Modern machine learning (ML) has generated the use of named entity recognition approaches; moreover, a diversity of deep learning (DL) techniques is used for entity recognition tasks [5]. The main objective of this study is to improve NER and healthcare phrase retrieval reliability by developing an effective semantic assessment system for EHR data. It offers to employ innovative methods for data analysis to motivate more accurate healthcare choices.

1. Related work

Kotechaet al. [13] proposed the CODE-EHR structure to enable an accurate and efficient use of healthcare data for research. Researchers and clinicians should adopt the CODE-EHR fundamental requirements architecture to enhance methodology accessibility and research design. The health digital state (HDS) and the smart EHR system, infrastructure architecture for EHRs intended to help medical practitioners make decisions based on the HDS, were both introduced by Serbanati [7]. The HDS influenced the body of information that guides medical practitioners' judgments about diagnosis and course of therapy. 488 patients with congenital heart disease who were dependent on a shunt for ventilation and had a single ventricle were taken [11]. A single-center prospective cohort research conducted between 2014 and 2019 found that patients were admitted to the cardiovascular intensive care unit before the second phase of palliative care. Ruiz et al. [8] routinely gathered EHR data for determined individuals at high risk of clinical deterioration by developing and assessing a multidimensional, information-driven framework. A retrospective, single-center study was conducted by Liu et al. [14] using a significant, longitudinal data set that was gathered from the largest tertiary hospital's HER [9]. The creation of a novel, dynamic, and easily interpreted System for Emergency Risk Triage (SERT) that uses ML and extensive EHRs to facilitate risk assessment in the ED. Alexander et al. [10] identified Alzheimer's disease (AD) patients using a previously validated rule-based phenotyping method in Clinical Practice Research Data (CPRD) link primary care HER [15]. A variety of comorbidities, symptoms, and demographic characteristics were retrieved and included as patient attributes. Four distinct clustering techniques were evaluated to group AD patients and each approach used metrics of predicting outcome effectiveness, stability, replication in outside information sets, and creation of clusters. Clustering based on significant clinical consequences was contrasted.

2. Methodology

Data from the open-source platform are gathered and presented in this part. Tokenization is used for preprocessing the collected data, and word 2 vector methods are used to extract the properties. The development and use of a unique Cubic Support Vector Machine with Dipper Throated Optimization (CSVM-DTO) technique has shown significant gains in the recognition and extraction of healthcare terms and entities from EHRs. Figure 1 depicts the framework methodology.

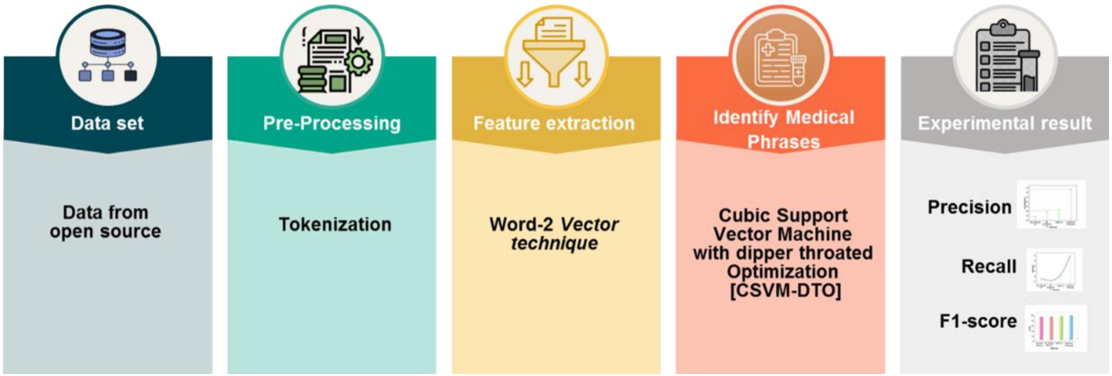

Figure 1 Framework methodology

Dataset

The data collection is in CSV format and includes 1447 records with unique patient IDs, demographics, hospital admission, and stay duration, illness conditions the patient is suffering from, discharge status, and the patient's destination [10].

Pre-processing using Tokenization

Tokenization is the most important step in NLP, which is dividing text into discrete units like words, symbols, phrases, or tokens. DL models require this procedure to manage and interpret textual material efficiently. Tokenization greatly helps medical records, which include information about the patient's health. It enables organized analysis and the extraction of pertinent medical data that is embedded in the text.

Feature extraction

In word2vector, word illustrations are produced using word embeddings. The vectors with the capacity to encode words closer together in the vector space should have the same meaning. One type of model is a continuous skip-gram (CSG) and the other is a continuous bag of words (BoW) made up of the word2vector. Using the words to predict their neighboring words is the fundamental concept of the CSG. The continuous BoW quantified intelligence predicts words by using context words from a neighboring booth. The benefit of continually condensing the dispersed information in the data is logically implied by the continuous BoW model structure.

Cubic support vector machine with dipper throated optimization (CSVM-DTO)

Cubic support vector machine (CSVM)

Selecting a new classifier that works for the extremely compact collection of medical phrases is a critical and significant job after preprocessing and extracting features. To effectively recognize medical terms in electronic health records, the classifier makes distinctions. CSVM classifier is a modified version of the typical SVM algorithm that utilizes cubic kernels to improve its capabilities. Supervisory learning models are especially valuable in high-dimensional circumstances and are employed in issues involving regression and classification. CSVM mostly utilizes a cubic kernel, whereas standard SVMs implement linear, polynomial kernels. The SVM's potential to recognize complicated connections in the information that could not be separated by linearity is enhanced by the CSVM kernel function. The kernel function is defined in equation (1).

$$l(\overline{w}) = \begin{cases} 1 & if|\overline{w}| \ \leq 1 \\ 0 & otherwise \end{cases} \tag{1}$$

These functions provide the advantageous characteristic space's innermost combination of two adjacent locations. Thus, it provides a concept of resemblance that requires low computational expenses and high three-dimensional area. A polynomial kernel function of order three is employed as indicated by the equation (2).

$$L\left(w_j, w_i\right) = (w_j^S . w_i + 1) \tag{2}$$

γ is a scaling variable.

r is a coefficient that has translational potential.

d is the degree of polynomial.

$$L\left(w_j, w_i\right) = (w_j^S . w_i + 1)^3 \tag{3}$$

Dipper throated optimization

Dipper the distinctive hunting style of the throated bird is characterized by its quick bending motions, which are accentuated by its flawless white breasts. When the prey sees anything, it plunges headfirst into the water, even if the water is tumultuous and moving quickly. The Dipper Throated Optimization (DTO) method operates under the assumption that the birds are flying and swimming in search of food sources that are accessible to n birds. The following matrices can be used to describe the positions, AO and velocities, AU of the birds.

The following equation (4) serves as the foundation for the optimizer's initial DTO process, which updates the swimming bird's position.

$$AO_{me}(s+1) = AO_{best}(s) - D_1 . |D_2 . AO_{best} - AO_{me}(s) \tag{4}$$

$AO_{me}(s)$ Is a typical flying posture during the repetition$s$, $AO_{best}$is the ideal place for a bird, "." is the multiplying of pairs, $AO_{me}(s+1)$ the outcome's modified bird location.

The following equation (5) is used to update the speed and location of the flying bird in the second DTO mechanism. The locations of the birds that flap have been revised.

$$AO_{me}(s+1) = AO_{me}(s) + AU(s+1) \tag{5}$$

$AO_{me}(s+1)$ Is the standard bird's altered position, with each bird's revised speed $AU(s+1)$ is calculated in equation (6).

$$AO(s+1) = D_3 AO(s) + D_4 q_2\left(AO_{best}(s) - AO_{me}(s)\right) + D_5 q_2(AO_{Hbest} - AO_{me}(s)) \tag{6}$$

This equation (7) describes the DTO algorithm.

$$AO_{me}(s+1) = \begin{cases} AO_{best}(s) - D_1.|N| \ \ if \ Q < 0.5 \\ AO_{me}(s) + AU(s+1) \qquad otherwise \end{cases} \tag{7}$$

$N = D_2;\ AO_{best}(s), AO_{me}(s)$ And Q is arbitrary range of [0, 1]

The CSVM-DTO methodology outperforms in identifying and retrieving recognized items from electronic health records.

3. Results and discussion

We evaluate the effectiveness of the CSVM-DTO strategies under studies regarding determining medical phrases in EHR in terms of recall, f1 score, and precision in this investigation using a variety of established methods, including class-weighted long-short-term memory conditional random fields (CW-LSTM-CRF) [11], concept-enhanced named entity recognition model (CNER) [11], and CW-(Bidirectional LSTM-CRF) [11]. The following accuracy and F1-Score numerical results are displayed in Table 1.

Table 1 Numerical outcomes of precision, recall, F1-score

Methods	**Recall (%)**	**Precision (%)**	**F1-Score (%)**
CNER	88.29	88.23	88.26
CW-BLSTM-CRF	87.23	87.92	87.68
CW-LSTM-CRF	87.56	86.87	87.12
CSVM-DTO [Proposed]	91.89	92.73	91.64

Precision: The precision metric quantifies the percentage of accurately recognized medical terms among all terms detected by the algorithm. It shows that the favorable predictions made by the model are accurate, as defined in equation (8). Figure 2 compares the outcomes of precision. The precision was 92.73% when the recommended CSVM-DTO was used in contrast, lesser scores were obtained using other conventional techniques, including CNER (88.23%), CW-BLSTM-CRF (87.92%), and CW-LSTM-CRF (86.87%).

$$precision = \frac{True\ postive}{True\ postive + False\ positive} \tag{8}$$

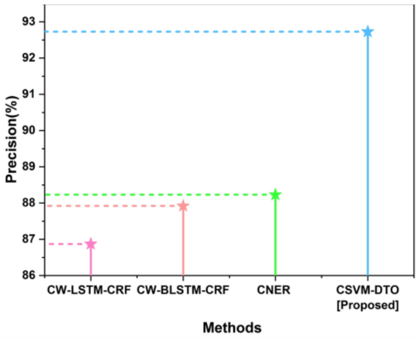

Figure 2 Precision

Recall: Recall is the process of confirming that the suggested model produces beneficial results. Equation (9) defines the proportion of recall. Figure 3 displays the comparison of recall. When the suggested CSVM-DTO was utilized, the recall was 91.89%, which is higher than traditional approaches such as CW-BLSTM-CRF (87.23%), CNER (88.29%), and CW-LSTM-CRF (87.56%).

$$Recall = \frac{True\ postive}{True\ postive + False\ negative} \tag{9}$$

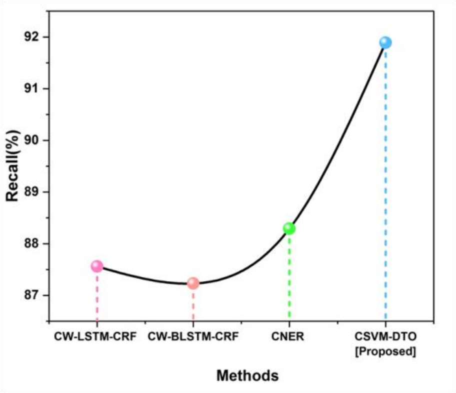

Figure 3 Recall

F1 Score: The F1-score becomes essential for assessing phrase identification in light of the unequal distribution of classes. To ensure a trustworthy and significant EHR in the context of public health, this measure is crucial for balancing false positives and false negatives, which are defined by Equation (10). Figure 4 compares the outcomes and F1-score. Using the suggested CSVM-DTO, the F1-score was 91.64%, whereas other standard approaches, such as CW-LSTM-CRF (87.12%), CW-BLSTM-CRF (87.68%), and CNER (88.26%), produced lower scores.

$$F1 - score \ = \frac{2 \times recall \times Precision}{recall + Precision} \qquad (10)$$

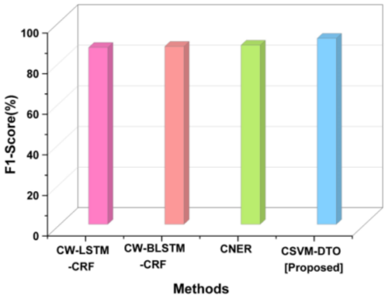

Figure 4 F1-score

4. Conclusion and future scope

This study was determined by emphasizing the potential use of semantic methods for assessment for improving healthcare decision-making through the examination of EHR data. Data are collected from the open-source platform. The gathered data are preprocessed using tokenization and the attributes are extracted by word 2 vector techniques. Major improvements in the detection and extraction of healthcare records from EHRs have been demonstrated by establishing and implementing a distinctive CSVM-DTO approach. The CSVM-DTO methodology performed more effectively than prior methods for recognizing and obtaining recognized objects from EHRs. This innovation can substantially enhance public health decision-making by providing greater certainty on patient status and treatment results. CSVM-DTO achieves precision (92.73%), recall (91.89%), and f1-score (91.64%). It may be contingent heavily on quality and reliability of HER data. The model's future scope is evaluating interaction with advanced technologies like NLP and deep learning, which could improve the EHR analysis's semantic assessment skills and enhance patient outcomes and healthcare delivery.

Reference

[1] J.K. Taitsman, A. VanLandingham, and C.A.Grimm,“Commercial influences on electronic health records and adverse effects on clinical decision-making”, JAMA Internal Medicine, 180(7), pp.925-

926.2020.https://doi.org/10.1001/jamainternmed.2020.1318

[2] O. Baclic, M. Tunis, K.Young, C. Doan, H.Swerdfeger, and J. Schonfeld, "Artificial intelligence in public health: Challenges and opportunities for public health made possible by advances in natural language processing", Canada Communicable Disease Report, 46(6), p.161. 2020.https://doi.org/10.14745%2Fccdr.v46i06a02

[3] S. Neelima, Manoj Govindaraj, Dr.K. Subramani, Ahmed ALkhayyat, & Dr. Chippy Mohan. (2024). Factors Influencing Data Utilization and Performance of Health Management Information Systems: A Case Study. Indian Journal of Information Sources and Services, 14(2), 146–152. https://doi.org/10.51983/ijiss-2024.14.2.21

[4] C.Giordano, M. Brennan, B. Mohamed, P. Rashidi, F.Modave, and P. Tighe, "Accessing artificial intelligence for clinical decision-making". Frontiers in digital health, 3, p.645232.2021.https://doi.org/10.3389/fdgth.2021.645232

[5] D. Kotecha,F.W. Asselbergs,S. Achenbach,S.D. Anker,D. Atar, C. Baigent,A. Banerjee,B. Beger,G. Brobert,B. Casadei, and C. Ceccarelli, "CODE-EHR best-practice framework for the use of structured electronic health-care records in clinical research", The Lancet Digital Health, 4(10), pp.e757-e764. 2022.https://doi.org/10.1016/S2589-7500(22)00151-0

[6] Sonya, A., & Kavitha, G. (2022). A Data Integrity and Security Approach for Health Care Data in Cloud Environment. Journal of Internet Services and Information Security, 12(4), 246-256.

[7] V.M. Ruiz, M.P. Goldsmith, L. Shi,A.F. Simpao, J.A. Gálvez, M.Y. Naim, V. Nadkarni, J.W. Gaynor, andF.R. Tsui, "Early prediction of clinical deterioration using data-driven machine-learning modeling of electronic health records", The Journal of Thoracic and Cardiovascular Surgery, 164(1), pp.211-222.2022.https://doi.org/10.1016/j.jtcvs.2021.10.060

[8] N. Liu,F. Xie, F.J. Siddiqui, A.F.W. Ho, B. Chakraborty, G.D. Nadarajan, K.B.K. Tan, and M.E.H. Ong, "Leveraging large-scale electronic health records and interpretable machine learning for clinical decision making at the emergency department: protocol for system development and validation", JMIR Research Protocols, 11(3), p.e34201.2022.https://doi.org/10.2196/34201

[9] Mohamed, K.N.R., Nijaguna, G.S., Pushpa, Dayanand, L.N., Naga, R.M., & Zameer, AA. (2024). A Comprehensive Approach to a Hybrid Blockchain Framework for Multimedia Data Processing and Analysis in IoT-Healthcare. Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications (JoWUA), 15(2), 94-108. https://doi.org/10.58346/JOWUA.2024.I2.007

[10] https://www.kaggle.com/datasets/anuchhetry/electronic-health-record

[11] Q. Zhao, D. Wang, J. Li, and F. Akhtar, "Exploiting the concept level feature for enhanced name entity recognition in Chinese EMRs", The Journal of Supercomputing, 76, pp.6399-6420. 2020. https://doi.org/10.1007/s11227-019-02917-3.

[12] W. Leeson, A. Resnick, D. Alexander, and J. Rovers, "Natural language processing (NLP) in qualitative public health research: a proof of concept study", International Journal of Qualitative Methods, 18, p.1609406919887021. 2019.https://doi.org/10.1177/1609406919887021

[13] L.D. Serbanati."Health digital state and Smart EHR systems", Informatics in Medicine Unlocked, 21, p.100494.2020.https://doi.org/10.1016/j.imu.2020.100494

[14] N.Alexander, D.C. Alexander, F.Barkhof, and S. Denaxas, "Identifying and evaluating clinical subtypes of Alzheimer's disease in care electronic health records using unsupervised machine learning", BMC Medical Informatics and Decision Making, 21, pp.1-13.2021.https://doi.org/10.1186/s12911-021-01693-6

[15] Bobir, A.O., Askariy, M., Otabek, Y.Y., Nodir, R.K., Rakhima, A., Zukhra, Z.Y., Sherzod, A.A. (2024). Utilizing Deep Learning and the Internet of Things to Monitor the Health of Aquatic Ecosystems to Conserve Biodiversity. Natural and Engineering Sciences, 9(1), 72-83.