Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Unsupervised Semantic Mapping for Healthcare Data Storage Schema

FAHAD AHMED SATTI[1,2], MUSARRAT HUSSAIN[1], JAMIL HUSSAIN[3], SYED IMRAN ALI[1], TAQDIR ALI[1], HAFIZ SYED MUHAMMAD BILAL[1,2], TAECHOONG CHUNG[1],and SUNGYOUNG LEE[1] (Member, IEEE)

[1]Department of Computer Science and Engineering, Kyung Hee University, Seocheon-dong, Giheung-gu South Korea (e-mail: {fahad.satti, musarrat.hussain, imran.ali, bilalrizvi, sylee}@oslab.khu.ac.kr, tcchung@khu.ac.kr)
[2]National Univeristy of Sciences and Technology, School of Electrical Engineering and Computer Science NUST-SEECS, Islamabad, Pakistan (e-mail: {fahad.satti, bilal.ali}@seecs.edu.pk)
[3]Department of Data Science, Sejong University, Seoul, South Korea (e-mail: jamil@sejong.ac.kr)

Corresponding authors: Sungyoung Lee (e-mail: sylee@oslab.khu.ac.kr) and TaeChoog Chung (e-mail: tcchung@@khu.ac.kr)

ABSTRACT Data, information, and knowledge processing systems, in the domain of healthcare are currently plagued by the heterogeneity at various levels. Current solutions, have focused on developing a standard based, manual intervention mechanism, which requires a large amount of human resource and necessitate realignment of existing systems. State-of-the-art methodologies in the field of natural language processing and machine learning, can help to partially automate this process, reducing the resource requirements and providing a relatively good multi-class based classification algorithm. We present a novel methodology for bridging the gap between various healthcare data management solutions by leveraging the strength of transformer based machine learning models, to create mappings between the data elements. Additionally, the annotated data, collected against five medical schemas and labeled by 4 annotators is made available for helping future researchers. Our results indicate, that for biased, dependent multi-class text classification, RoBERTa shows the best performance, achieving a cohen's kappa score of 0.47 and MCC score of 0.48, with human annotated data.

INDEX TERMS Electronic medical records, Health information management, Medical information systems, Text processing, Unsupervised learning

I. INTRODUCTION

DATA and Information modeling in the healthcare domain have witnessed significant improvements in the last decade owing to advances in the development of state-of-the-art Information and Communication Technologies (ICT) and formalization of storage and messaging standards. Subsequently, the scope of Healthcare Management Information Systems (HMIS), medical ontologies, and Clinical Decision Support Systems (CDSS) has broadened, beyond the operational capabilities of traditional rule based systems. One of the major reasons behind this limitation is due to the numerous heterogeneities in healthcare at data, knowledge, and process level. Thus, healthcare interoperability which aims to provide a solution to this problem, can be compartmentalized into data interoperability, process interoperability, and knowledge interoperability. Various abbreviations, used in this manuscript are listed in Table 1.

Data interoperability resolves the heterogeneity between

TABLE 1. Abbreviations

Abbreviations	Full Form
ICT	Information and Communication Technologies
HMIS	Healthcare Management Information Systems
CDSS	Clinical Decision Support Systems
HL7	Health Level Seven International
HIMSS	Healthcare Information and Management Systems Society
FHIR	Fast Healthcare Interoperability Resources
Snomed-CT	Systematized Nomenclature of Medicine—Clinical Terms
LOINC	Logical Observation Identifiers Names and Codes
MLM	Medical Logic Module
CIMI	Clinical Information Modeling Initiative
UHP	Ubiquitous Health Platform
NLP	Natural Language Processing
BERT	Bidirectional Encoder Representations from Transformers
UMLS	Unified Medical Language System
EMR	Electronic Medical Records
EHR	Electronic Health Records
MCC	Matthews Correlation coefficient
NLI	Natural Language Inference

data artificats, to enable, seamless and interpretable communcation among source and target organizations, while preserving the data's original intention during storage, communication, and usage (as defined by IEEE 610.12 [1], Health Level Seven International (HL7), and Healthcare Information and Management Systems Society HIMSS [2]). On the other hand, process interoperability regulates the communication among organizational processes to provide compatability between process artifacts within and seamless transformations across different organizations [3]. Lastly, knowledge interoperability provides a sharing mechanism for reusing interpretable medical knowledge, acquired through expert intervention and other mechanisms, across decision support systems [4]. In more tangible terms, healthcare interoperability at data, process, and knowledge level can be exemplified within the healthcare constraints experienced due to the emergence of Covid 19. The operational capabilities of the current healthcare service delivery infrastructure has gone under tremendous stress due to Covid 19. World over, large primary healthcare units have managed to create separate units for managing patients, suffering from extreme cases of the novel coronavirus. For secondary and tertiary care units, government involvement has become necessary to filter coronavirus patients and adhering to a national pandemic response policy. These complex circumstances have enhanced the need for sharing patient data and state-of-the-art medical knowledge in real-time, to provide the medical experts with a tool to make accurate and timely decisions. Data interoperability can enable the front line medical workers to fetch, understand, and use patient data, especially comorbidities, across organizational and physical boundaries, without suffering from societal taboos that may prevent the patient from sharing their complete and accurate medical histories. Knowledge interoperability can improve the knowledge acquisition and sharing protocols to provide the medical experts such as epidieomologists and vaccinologist, with latest information on affected population trends, disease diagnosis, treatment, and followup procedures, and interpretable decisions leading to positive or negative outcomes. Process interoperability can help reduce and in some cases remove the operational redundancies between health centers. In this way, successive healthcare treatments can take benefit from earlier diagnosis, treatment, and followup procedures, thereby reducing the stress on healthcare experts and systems. Standards such as HL7 - Fast Healthcare Interoperability Resources (FHIR), and openEHR provide the foundations for storing and communicating medical data, through the use of well defined protocols. While Systematized Nomenclature of Medicine—Clinical Terms (Snomed-CT) [5] and Logical Observation Identifiers Names and Codes (LOINC) [6] provide a standard definition for clinical terminologies and laboratory tests, respectively. Similarly Medical Logic Module (MLM) provides a standardized way for expressing medical knowledge. However, the plethora of standards, necessitates the creation of bridging standards, that can resolve the heterogeniety between the medical standards. Substantial effort has gone into this endeavor with the Clinical Information Modeling Initiative (CIMI) [7] taking the lead in bridging the gap between HL7v3 and openEHR. Similarly, SNOMED CT and LOINC are working to resolve the redundancies between the two terminological standards since 2013. This healthcare interoperability solution follows a formal, albiet long process, which is greatly dependent on the human factor. However, the current healthcare scenario, requires a quick solution to create a scaffolding of an interoperable bridge between various healthcare providers. It is also important to ensure that this scaffolding should be able to support the formal standardization processes of the future. In [8]we have presented the Ubiquitous Health Platform (UHP), which provides semantic reconciliation-on-read based data curation for resolving data interoperability between various schema. This methodology is based on the creation and management of schema maps, that can provide the framework for transforming a source schema into a target schema. In the current manuscript, we will present our research work to build and manage the schema map knowledge base. Overall, our methodology is based on the creation, evaluation, and application of a novel schema matching technique to identify the relationships between attributes of the participating medical data schema. Since the terms used to identify attributes in the data schemas are not defined in any standard way, it is important to first identify the component words of the attribute term and then to append semantic concepts with these to create a meaningful sentence. This process is based on word expansion using concept lookup from Unified Medical Language System (UMLS). Once the sentence has been created, it is then trivial to create its embedded vectors representation using transformer based pretrained models. The cosine similarity of any two embedded vectors can then indicate the degree of similarity between the original attributes. The motivation behind this methodology along with its development and evaluation details will be presented in the following flow:

- Section II presents the related work
- Section III contains the details of our methodology
- Section IV provides the experimental setup
- Section V presents the results
- Section 6 concludes the paper.

II. RELATED WORK

Althubait et al. [9] proposed an ontology expansion methodology that identifies and extracts new class from text articles using word embedding and machine learning techniques. The authors identified the similarity of tokens and phrases of the text articles with the exiting classes of the ontology. The target ontology is expanded with classes from text articles having greater similarity with that of already added classes. A similar word embedding technique was also used by Nozaki et al. [10], where the authors used instance based schema matching technique to identify the semantic similarity between two instances. The results of the study showed the possibility of detecting similar string attributes of differ-

ent schemas. Yousfi et al. [11] also utilized semantic base techniques and proposed xMatcher XML schemas matching approach. xMatcher transforms schemas into a set of words, followed by measuring words context, and relatedness score using WordNet. The terms from different schemas having similarities greater or equal to 0.8 are considered similar. Bylygin et al. [12] devised an ontology and schema matching approach by combining lexical and semantic similarity with machine learning approaches. The authors used lexical and semantic measures as features and trained various machine learning algorithms including Naive Bayes, logistic regression, and gradient boosted tree. The result achieved showed that the combination of algorithms outperformed the single modal.

Martono et al. [13] provided overview of previous studies related to linguistic approaches used for schema matching. Linguistic methods focused on finding strings and evaluate there similarity in different schemas. The string are normally normalized before to align both the strings before similarity comparison. The normalized strings are categories based on the information relatedness and element with similar category are compared using various similarity measure including Jaro-distance, Lavenstein (edit-distance), and many more. Alwan et al. [14] have summarized the techniques used in the literature for schemas and instances based schema matching. The information used for schema matching is categories into schema information, instance and auxiliary information. Most of the searchers have used syntactic techniques (including n-gram, and regular expression), semantic techniques (including Latent Semantic Analysis, WordNet/Thesaurus and Google Similarity) for schema level and instance level matching to achieve the final goal of data/information interoperability. Kersloot et al. [15] performed a comprehensive systematic review to evaluate Natural Language Processing (NLP) algorithms used for clinical text mapping onto ontological concepts. The findings of the studies were evaluated with respect to five categories; use of NLP algorithms, data used, validation and evaluation performed, result presentation, and generalization of results. The authors revealed that over one-fourth of the NLP algorithms used were not evaluated and have no validation. The systems that claimed generalization, were self evaluated and having no external validation.

Xu et al. [16] presented a framework for discovering indirect links besides direct links among schema elements. The indirect matches were detected for relations such as union, composition, decomposition, selection, and boolean. The indirect links are useful to handle concepts merge, split, generalization, and specialization. The matching techniques utilized in the study considered terminological relationships (word synonym and hypernym), structural characteristics, data-value characteristics, and expected data values. The experimental results revealed framework effectiveness by achieving more than 90% precision and recall for direct and indirect link matching.

A comprehensive survey from 176 experts including physicians and nurses was conducted by Moll et al. [17] to check their perspectives regarding Patient Accessible Electronic Health Record (PAEHR). The authors discovered that the PAEHR positively effect after six years of operations despite negative expectations. The primary concerns revealed was logger meeting time, change in documentation practices, and increasing varies of patients regarding their health conditions. However, attitude of both healthcare service providers and patients are changing positively with respect to PAEHR and its benefits.

III. METHODOLOGY

Healthcare interoperability, with a focus on non-standard compliant medical schema, is dependent on the generation and validation of schema maps, as discussed above. To this end, the creation of a cohesive workflow is of utmost importance. In our earlier work [8] we used maximum sequence identification and suffix tress based matching for syntactic matching of two distinct data schemas. This was followed by semantic concept enrichment and subsequently concept matching, for creating rules in the form of schema maps. The simplified mapping functions, thereby created, provided a simple methodology for converting semi-structured medical data, into an interpretable, model form. In our current methodology we have utilized state-of-the-art natural language processing (NLP) techniques to extract the schema mapping rules from semi-structured data schemas. This methodology is based on identifying similarity between vector representations of two attributes, belonging to different medical schemas. Traditional NLP techniques such as WordNet are able to convert a word into an embedded vector, while Bidirectional Encoder Representations from Transformers (BERT) extracts an embedded vector from a sentence [18]. However, the terms forming the attribute names are bigger than a word (combination of multiple words) and smaller than a sentence. In order to resolve this problem, we extracted the set of suffixes from the terms forming the attribute names. The bidirectional nature of BERT, allows the creation of contextual embedded vectors, where each target word is affected by its neighboring words. Hence to convert the set of suffixes into a sentence, we collected the set of concepts corresponding to each suffix, from UMLS. This operation has two effects, firstly it is used to remove any suffix, which does not have a corresponding concept and secondly the extracted concepts are used to add context to each suffix and produce a contextual sentence. The following subsections provide the practical details for our methodology, from schema acquisition, to attribute name expansion, and finally schema map generation.

A. SCHEMA ACQUISITION

In the first step of our semantic reconciliation methodology, we simulate medical data acquisition from five distinct Electronic Medical Records (EMR) storage systems (S). These include patient reports from OpenEMR (s_1), 100,000 patient records from EMRBOTS (s_2) [19], custom database design

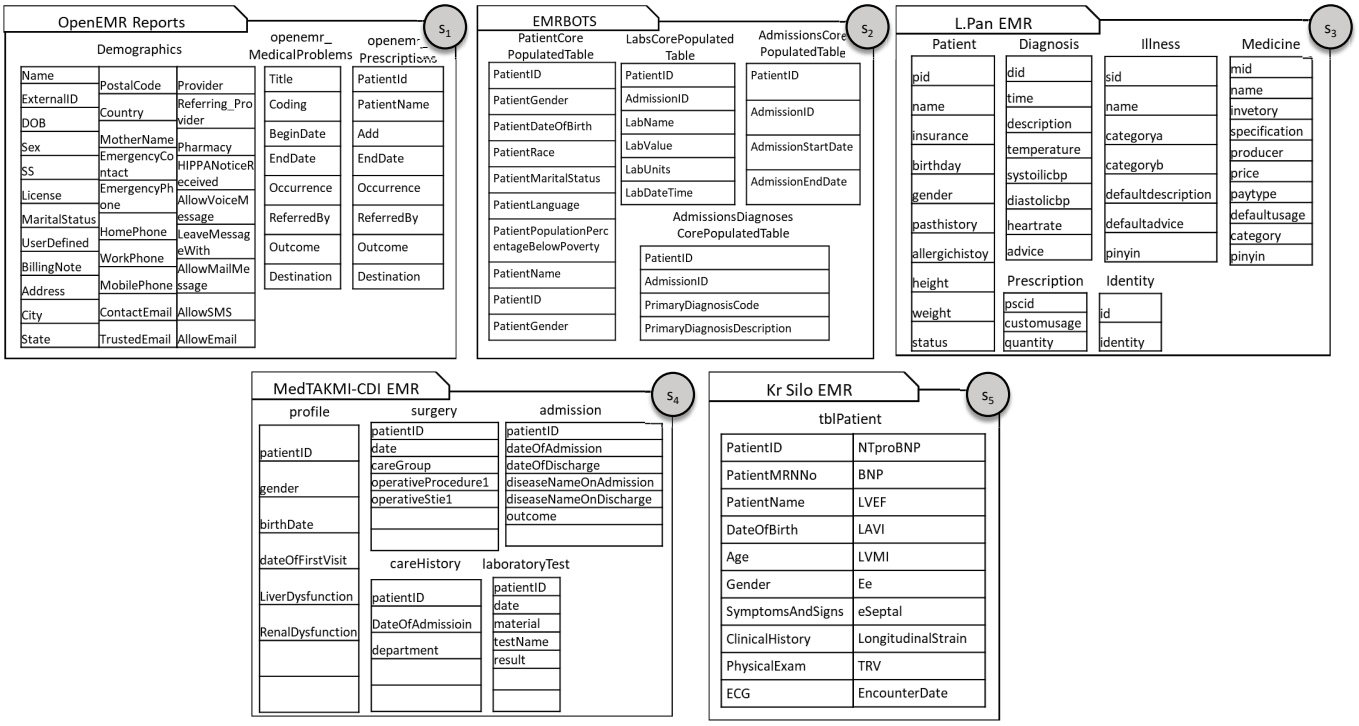

FIGURE 1. The five medical schemas used for achieving data interoperability.

by Pan et. al(s_3) for supporting regional clinics and health care centers in China [20], clinical knowledge discovery tool MedTAKMI-CDI (s_4) [21], and our custom implementation (s_5). Each of these medical systems as shown in 1, follows the relational database design, with logical entities, such as demographics, diagnosis, medicine or others, placed into tables which can be further linked to one or more tables. While the database design implemented by each of these systems, fulfills the need of their respective information processing applications, the lack of interoperability, in terms of identifying similar attributes or exchanging the medical data is very much evident here.

A similar notion of data heterogeneity, in terms of medical data schema, is evident across the healthcare domain. This is caused by various factors, including the lack of one all-encompassing, and universally applicable terminological standard and different normalization level for representing attributes.

In the former case, while SNOMED-CT provides a mechanism for identifying the standard codes for clinical terms and LOINC can be used for laboratory related terms, most attribute names are created based on the gut feeling of the database designer. Additionally, while these codes can be used to represent data instances, the data schema, achieves no benefit from the same. Consider the terms “name" and “patientName", which refer to the same attribute of the patient entity. However, since there is no standard way to represent this attribute, both are considered correct (s_1 and s_3 use the former representation, while s_2 and s_5 use the latter).

In the later case, differences in normalization cause semantic differences, due to which some data could be available in one schema but absent in others, such as OpenEMR demographics identifying the patient’s residential location using specific attributes like “Address", “City", “State", “Postal Code", “Country", and others. Similarly, “EncounterDate" from s_5 is semantically similar to “BeginDate" of “openemr_MedicalProblems" table in s_1, “AdmissionStartDate" of “LabsCorePopulatedTable" in s_2, “time" in “Diagnosis" table of s_3, and “dateOfAdmission" in “Diagnosis" and “CareHistory" tables of s_4. Finally, s_1 and s_3 have separate tables containing the medicinal prescription details, however the same details are unavailable in s_2, s_4, s_5. Once again, this is not an incorrect behavior since this information, might not be a part of the context or the requirements for the EMR/EHR storage systems. In fact, the change in context of the medical data storage system from the initial time of development to a later stage of collaborative processing systems, is the main cause of heterogeneity. In order to provide an interoperable solution, it is therefore necessary to enhance the semantics of each data attribute by its contextually equivalent sentence.

B. FROM ATTRIBUTE TO SENTENCE

In order to process the EMR/EHR schema set S and produce a set of corresponding semantically enriched sentences, we use the data representation s_i, generated through the process explained in sequence acquisition to collect the various medical fragments in memory. We then iterate over these fragments, building a set of attributes, distinguished by

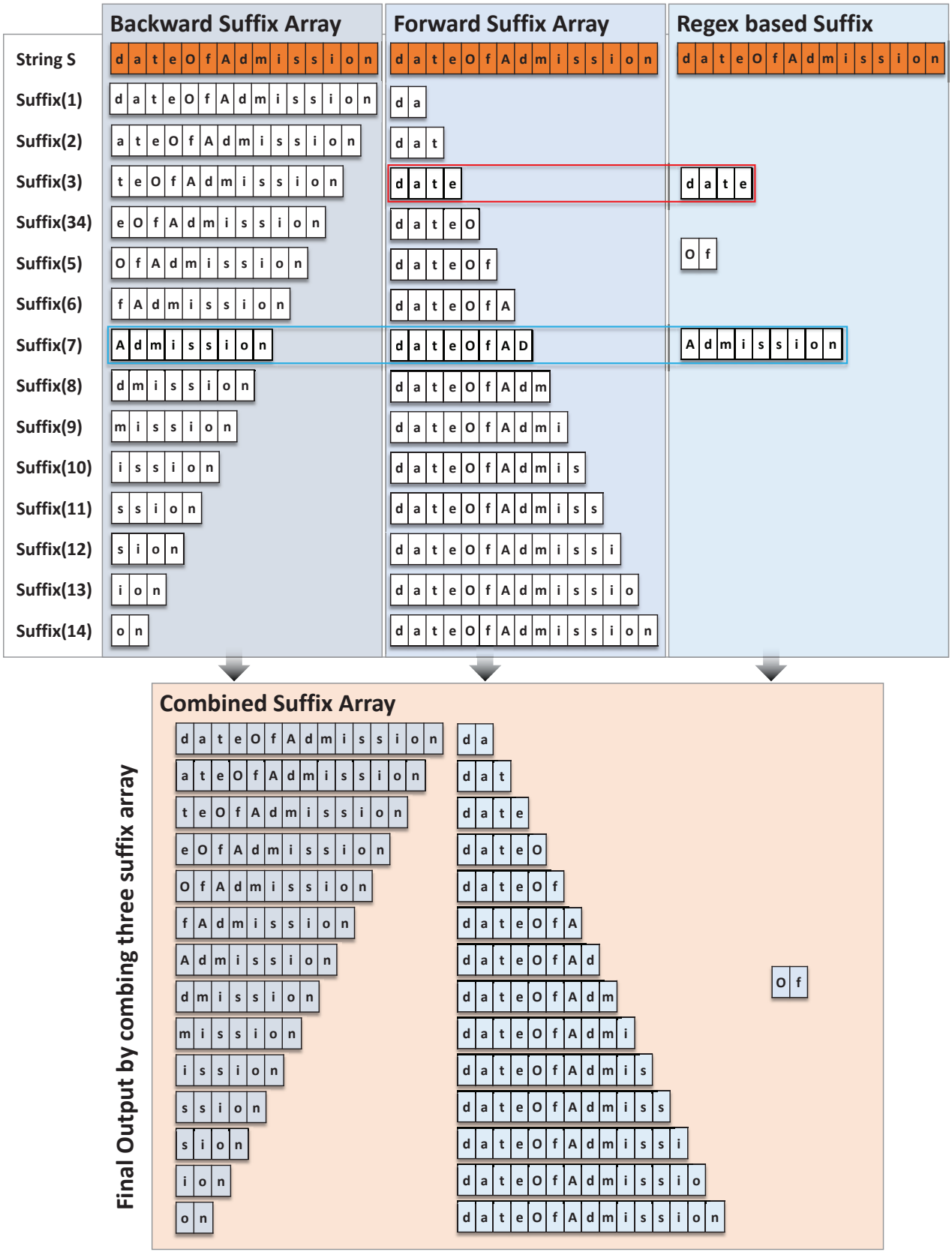

FIGURE 2. Suffix attributes.

their name, schema's name, table's name, schema's version, source, and recorded data. This entails that "PatientID" from each of the four tables in s_2, and "patientID" from five tables in s_4, would result into nine attributes (assuming, as in the current case, of no differences in versions of these systems). For each attribute, we then generate the suffix array, which provides all possible substring representations contained within the attribute name. In order to generate the set of suffixes, we employ three strategies, forward suffix generation, whereby for a word w of length n, $n-1$ suffixes of size 2 to $n-1$ are produced, backward suffix generation, to produce $n-1$ suffixes in reverse order with size $n-1$ to 2, and regular expression based suffix generation, which splits each word on, change of case, special characters(such as -, _, !, and others), and numbers. In this way a large list of suffixes is generated, which is combined using a "TreeSet" data structure of Java, which internally sorts this list as well. An example of this suffix generation process, using the attribute name "dateOfAdmission" as it appears in s_4 is shown in 2.

Suffix strings for similar attributes such as "AdmissionStartDate". "diseaseNameOnAdmission", and "AdmissionEndDate" appear in s_2, produce many, synctactically similar suffixes, to the presented example. This process, is only able to generate syntactic suffixes, producing many incoherent and unrelated suffixes. In order to counter this problem, and to limit the list of suffixes within the domain, we then query UMLS, with exact search strategy, looking for the existence of any concepts, against each suffix. In case, no semantic concept is found for a particular suffix, it is removed from the final Suffix Array. On the other hand, if atleast one semantic concept is found against the queried suffix, it is retained. Meanwhile the process continues for the next attribute, then the next table, and finally the next system, till no further processing is possible. The set of suffixes and their corresponding concepts are then used to build the sentence, where by each concept, corresponding to a suffix is appended next to the suffix. An example of the resultant sentence for the attribute "DateOfAdmission" is shown as follows:

> "Date Value type - Date date allergenic extract Date in time Data types - Date Date Fruit;Of SPI1 wt Allele SPI1 gene TAF1 wt Allele BRIP1 gene Within Degrees fahrenheit Oral contraception BRIP1 wt Allele;Da Displacement of abomasum dalton Anterior descending branch of left coronary artery deca units cytarabine/daunorubicin protocol Dai Chinese Asymptomatic diagnosis of Drug Accountability Domain;ion Iontophoresis Route of Drug Administration Ions;on SPARC wt Allele Osteonectin SPARC gene On (qualifier value) Upon - dosing instruction fragment;Admission Admission activity Hospital admission;Dat SLC6A3 gene SLC6A3 wt Allele dopamine transporter Direct Coombs test SLC6A3 protein, human Test Date cytarabine/daunorubicin/thioguanine Alzheimer's Disease;mission Religious Missions;"

Here the various suffixes and their concepts are seperated by the symbol ";", however together they form one sentence, for which an embedded vector is generated.

C. SCHEMA MAP GENERATION

Schema Maps, provide an interoperable bridge between two medical systems ($s_i \wedge s_j$), by identifying the semantic relationship between their participating attributes. This identification is based on the similarity between the embedded vectors, of the semantically enriched sentences, corresponding to each data attribute. While the embedded vectors can be generated using any methodology (we tested 11 methodologies, with WordNet and 10 models based on BERT (including BERT base, BERT large, distilBert, and RoBERTa) further detailed in section IV), the large/STSb version of RoBERTa [22],provides the best results. The pair of embedded vectors thus produced are then used to calculate cosine similarity, which is based on the inverse cosine distance between them. For our classification, we used the raw results (unnormalized) of cosine similarity, which produces a score between -1, and 1. Cosine similarity score of 0 indicates orthogonal relationship between the two vectors, which in our scenario indicates that the two sentences, and by extension their attributes are not related to each other. -1 indicates, inverse relationship between the attributes, while 1 indicates the two attributes are very much the same. For producing our schema maps, we are interested in three types of relationships, "equal" (the two attributes are same), "related" (the two attributes are related to each other), and "unrelated" (no relationship between the attributes). In order to classify the similarity results, into one of these three classes, we then calculated the best thresholds, using Matthews Correlation coefficient (MCC) [23] for classifying each instance as "equal", "related", and "unrelated".

$$MCC = \frac{(TP.FN) - (FP.FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \longrightarrow \quad (1)$$

MCC provides a fair measure of the ability with which a classifier can correctly predict both positive and negative instances [23]. MCC score of 0 represents random classification, however, with an increase in the number of true positives and true negatives, MCC moves closer to 1. It also takes into account the false positives and false negatives, which shift the mcc score towards -1. This measurement is markedly different from accuracy, that fails to account for imbalanced datasets and F_1 measure which is not affected by the true negative scores. As a result, MCC provides an acceptable alternate, in our current scenario comprising of imbalanced dataset (largely in favour of class "unrelated") to measuring the true performance of the models, used for threshold selection and model evaluation. Finally on a test dataset we evaluated our multi-class classification approach, using MCC and Cohen's Kappa coefficient (κ) [25] to identify the relationships between each pair of attributes.

IV. EXPERIMENTAL SETUP

In our earlier work [8] s_1, s_2, and s_5 were used to generate over 115 million patient records, which are converted into a semi-structured form and stored in Hadoop Distributed File System (HDFS). We extended the same setup to create an additional 100,000 records, for 1000 patients with 3 medical fragments for s_1, s_2, and s_4, and 97 randomly selected and generated medical fragments amongst s_1, s_2, s_3, s_4 and s_5. These fragments, follow various design elements, supporting a variety of valid relational storage architectures. Such as, s_1, s_2 and s_4 are represented by creating a separate medical fragment for each participating table, s_3 utilizes its medical fragment to generate a linked record (from a linked object graph), where by the attributes can refer to other objects besides the elements of t, mimicking the application of explicit foreign keys, and s_5 is a flat table structure. The code to generate this data set is available at "uhp_map_generation"[1]. This application produces three custom formatted files, containing an index for patients, an index for their medical fragments, and the medical fragment, corresponding to the EMR data. Using the medical fragments file, we then generate the semantically enriched attribute [2], which contains the suffixes and their concepts corresponding to each EMR data attribute. The resulting set of enriched attributes are temporarily stored in a "json" file, which is then read by the same application to partially generate the schema maps. This process, is used to create 20,349 distinct pairs of attributes, across s. Each pair also contains the "relationshipList", which stores the results of fuzzy string matching [24] [3] between the attribute names. The "json" file thus produced, is then used by a python script to generate the semantically enriched sentences and their embedded vectors using Word2Vec, and 10 pre-trained BERT NLI models [22]. The BERT models include 'bert-base-nli-stsb-mean-tokens', 'bert-large-nli-stsb-mean-tokens', 'roberta-base-nli-stsb-mean-tokens', 'roberta-large-nli-stsb-mean-tokens', 'distilbert-base-nli-stsb-mean-tokens', 'bert-base-nli-mean-tokens', 'bert-large-nli-mean-tokens', 'roberta-base-nli-mean-tokens', 'roberta-large-nli-mean-tokens', and 'distilbert-base-nli-mean-tokens'. The embedding vectors are then compared using cosine similarity, which produces a score between -1 and 1. The rationale behind switching the applications at various stages, is to cache the results and create checkpoints for restarting any failed stages, easily. Additionally, since python provides better support for easy generation of embedding vectors, it was thus preferred over the Java based implementation, which is otherwise very beneficial for other tools. These applications were executed on a workstation running Ubuntu 20.04.2 on top of AMD Ryzen 3 2200G, and 32GB ram.

V. RESULTS

The validity of our proposed approach has been evaluated using several techniques, including comparison of the proposed semantic matching process with fuzzy string matching, embedded vector generation and comparison using Word2Vec, and 10 BERT nli models.

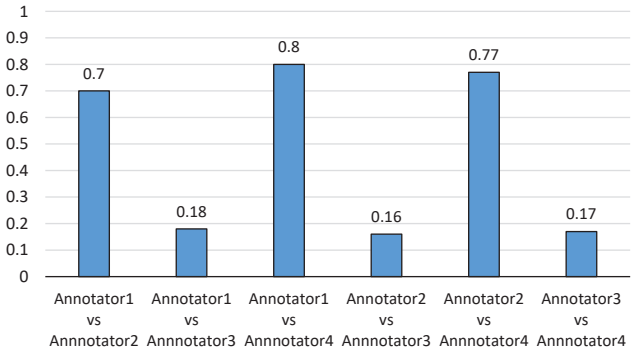

FIGURE 3. Cohen's Kappa (κ) score between the four annotators

A. DATASET ANNOTATION

In order to compare our computed models with ground truth, and to identify the best thresholds for classifying each instance as "equal", "related", or "unrelated" 4 human annotators were utilized, to anonymously, score the similarity of each pair of attribute names. In order to support this process, we first repurposed one of our generated data matrix, by marking all attribute pairs belonging to the same schema with the symbol "-". Following this, the annotators, marked each cell corresponding to a pair of attributes (conversely, each attribute pair corresponds to two cells, with the positioning of the pair-participants swapped; which is used for clarity and identify correct relationships between the attribute on left and attribute on right), by determining the similarity in terms of dissimilar as "0", exactly similar as "1", row attribute as child of column attribute as "<", row attribute as a parent of the column attribute as ">", and finally, unknown as " ". The data sheets generated after this extensive human effort, have been made available for other researchers[4].

These sheets, additionally contain some missing values, which were left out by the annotators, but in order to maintain their originality, these values were not filled; instead during our evaluation for these datasets, the missing values were considered as having the score "0". Using κ, we evaluated the inter-rater agreement of these annotations, which have been visualized in Figure 3. It can be seen in this plot, that "Annotator3" has very small correlation with the other 3 annotators. This difference can be traced back to the number and type of annotations performed by each annotator, which is shown in Table 2. The "Annotator3" has marked 2103 cells as related (one of >, < , or) and left 153 as empty. However, even in the presence of these difference, it is pertinent to include the data for all annotators in order to avoid any biasedness.

[1]https://github.com/desertzebra/UHP_v4/tree/main/uhpr_storage
[2]https://github.com/desertzebra/UHP_v4/tree/main/uhp_map_generation
[3]Java Library: https://github.com/xdrop/fuzzywuzzy
[4]https://github.com/desertzebra/EMR-Interoperability/tree/master/Implemenation/Da

TABLE 2. Annotations performed by the four annotators on five medical schema

Method	Total Matches	Marked as Equal	Marked as Related	Marked as Unrelated	Not Marked
Annotator1	40698	238	109	40351	0
Annotator2	40698	241	116	40341	0
Annotator3	40698	260	2103	38182	153
Annotator4	40698	225	62	40400	11

This annotated data was then processed, to replace all related entries with “0.5" and all“-" with “0", while the values for similar at “1" and “0" for dissimilar were kept the same. This conversion was then used to produce a consolidated dataset of 40,698 attribute pairs, using mode scores of all annotators, for each cell. We also tested average scores between the annotators, but that would produce scores between “0", “0.5", and “1", greatly increasing the number of classes for classification. Hence the maximum agreement between the annotators, maintains the final label values within these three classes, which become easier to evaluate. Additionally, the original dataset and its mode consolidated form is biased in favour of class "0", since most attribute pairs are not related to each other. This dataset is then split into development and testing partitions with a ratio of 70:30. The development partition is used for threshold selection, based on the best MCC score for identifying class “equal", followed by best scores for class “related" and finally best of class "unrelated". The optimal threshold thus achieved is used to classify the instances of the test dataset, which is finally evaluated on its MCC and F-1 score.

B. THRESHOLD SELECTION

A good text classification methodology is dependent on the correct choice of a threshold, which can maximize the target class participation. In case of independent labels, area under the precision recall curve can provide this optimal measure, however as in our case, for dependent classes on a biased dataset the Matthews Correlation Coefficient (MCC), is better [23]. Since our aim is to apply an optimal text similarity classifier to resolve this multi-class problem (class 0, class 0.5, and class 1), we have to test various threshold scores for separating the instances between 0 and 0.5 (t_1), and then 0.5 and 1(t_2). Additionally, since our aim is to correctly identify the similar attribute instances, it is pertinent to maximize the classification performance of class 1 (similar), followed by class 0.5 (related), and finally class 0 (unrelated). With a step size of 0.05 ($step_size$), and starting from t_1 as 0.0 and t_2 as t_1 + $step_size$, we move the thresholds until t_2 reaches 1, followed by increase in t_1 by step size. Eventually, t_1, reaches 0.95 and t_2 reaches 1, at which point, the process stops. This is to ensure that t_1 remains behind t_2, for all iterations, measuring MCC score, for the 12 models. These models include, “Fuzzy_Wuzzy", “Word2Vec", and 10 BERT based models.

Threshold values for Word2Vec are placed at the lower end of the spectrum, indicating a very large number of instances are classified as similar (above similarity score of 0.1), while a small number of instances (with similarity score 0.05) are classified as dissimilar. Similarly, the class related lies within the similarity threshold of 0.05 similarity points. It can be observed that the threshold for selecting the related class, is within 0.05 points, in all except one case (bert-large-nli-mean-tokens where the difference is 0.15 points). Five BERT based models, trained on the STSb dataset, all have minimum threshold values of 0.85 and maximum of 0.9, while the 4 remaining BERT models, lie between 0.9 and 0.95. These results show a general trend, of how the cosine similarity varies/maintains itself, against embedded vectors generated from various pre-trained models. In absolute terms, however these threshold values provide the mechanism for classifying the test dataset, which is evaluated for performance, in the next subsection.

C. MODEL EVALUATION

On unseen test dataset, with thresholds selected in the previous step and the 12 models, we measured the performance score, using one vs all binarization of the multi-classes.

Finally, we evaluated the overall κ coefficient and MCC score to evaluate the performance of each model on the test dataset. These scores range between [-1,1], providing a measure quantifying the accuracy of the classifier to correctly predict correct and incorrect instances. The models Word2Vec, bert-large-nli-mean-tokens, and roberta-base-nli-mean-tokens, with values between [0,0.20] indicate random classification, while κ score between [0.21,0.39], achieved by fuzzy wuzzy, bert-base-nli-mean-tokens, roberta-base-nli-sts-mean-tokens, distilbert-base-nli-stsb-mean-tokens, bert-base-nli-mean-tokens, roberta-large-nli-mean-tokens, and distilbert-base-nli-mean-tokens, show only minimal agreement with the annotated data [26]. Relative best rates, in this case, are achieved by roberta-large-nli-stsb-mean-tokens, which substantially surpasses the other models, however in absolute terms, it shows only weak agreement. Similarly, MCC scores show a similar result, with roberta-large-nli-stsb-mean-tokens, achieving the relative best performance amongst all classifiers.

VI. DISCUSSION

F_1 score, which provides the harmonic mean of precision and recall, ranges between 0 and 1, while MCC provides a fair measure of the ability with which a classifier can correctly predict both positive and negative instances [23].

Normalized MCC -> linear interpolation x := (c - a) / (b - a)

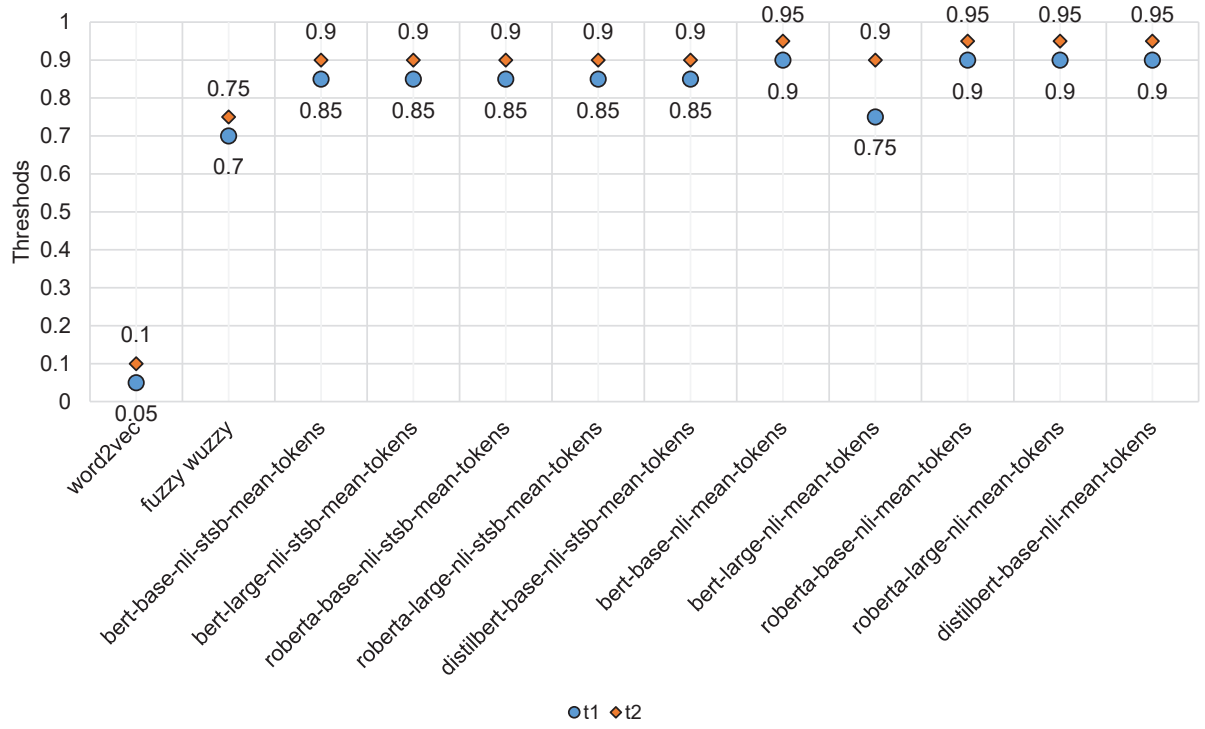

FIGURE 4. Thresholds selection.

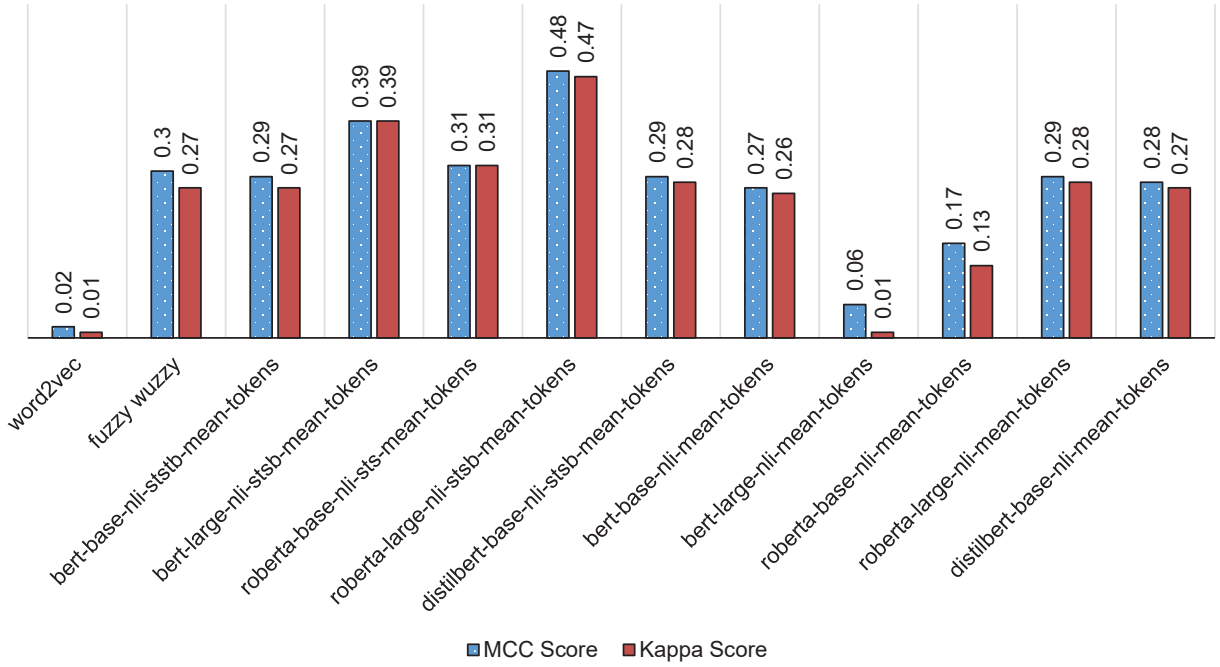

FIGURE 5. Models evaluation.

VII. CONCLUSION

Conclusion goes here.

ACKNOWLEDGMENT

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2017-0-01629)

TABLE 3. One vs all binary classifications results

Model	Class Positive	Class Negative	accuracy	precision	recall	f-1	mcc
FUZZY_MATCH	0.0	0.5,1	0.98	1.00	0.99	0.99	0.37
	0.5	0.0,1	1.00	0.00	0.00	0.00	0.00
	1	0.0,0.5	0.98	0.20	0.59	0.29	0.33
Word2Vec	0.0	0.5,1	0.24	0.99	0.24	0.38	0.01
	0.5	0.0,1	1.00	0.15	0.10	0.12	0.12
	1	0.0,0.5	0.35	0.01	**0.84**	0.01	0.03
bert-base-nli-stsb-mean-tokens	0.0	0.5,1	0.99	1.00	0.99	0.99	0.37
	0.5	0.0,1	1.00	0.00	0.00	0.00	0.00
	1	0.0,0.5	0.99	0.22	0.50	0.30	0.32
bert-large-nli-stsb-mean-tokens	0.0	0.5,1	0.99	1.00	1.00	1.00	0.51
	0.5	0.0,1	1.00	0.00	0.00	0.00	0.00
	1	0.0,0.5	0.99	0.41	0.50	0.45	0.45
roberta-base-nli-stsb-mean-tokens	0.0	0.5,1	0.99	1.00	0.99	1.00	0.40
	0.5	0.0,1	1.00	0.00	0.00	0.00	0.00
	1	0.0,0.5	0.99	0.26	0.50	0.34	0.36
roberta-large-nli-stsb-mean-tokens	0.0	0.5,1	1.00	1.00	1.00	1.00	0.61
	0.5	0.0,1	1.00	0.00	0.00	0.00	0.00
	1	0.0,0.5	**1.00**	**0.59**	0.50	**0.54**	**0.54**
distilbert-base-nli-stsb-mean-tokens	0.0	0.5,1	0.99	1.00	0.99	0.99	0.38
	0.5	0.0,1	1.00	0.00	0.00	0.00	0.00
	1	0.0,0.5	0.99	0.23	0.50	0.32	0.33
bert-base-nli-mean-tokens	0.0	0.5,1	0.99	1.00	0.99	0.99	0.35
	0.5	0.0,1	1.00	0.00	0.00	0.00	0.00
	1	0.0,0.5	0.99	0.20	0.50	0.29	0.31
bert-large-nli-mean-tokens	0.0	0.5,1	0.28	1.00	0.27	0.43	0.05
	0.5	0.0,1	0.99	0.34	0.90	0.49	0.55
	1	0.0,0.5	0.57	0.00	0.50	0.01	0.01
roberta-base-nli-mean-tokens	0.0	0.5,1	0.96	1.00	0.97	0.98	0.23
	0.5	0.0,1	1.00	0.00	0.00	0.00	0.00
	1	0.0,0.5	0.96	0.08	0.50	0.13	0.19
roberta-large-nli-mean-tokens	0.0	0.5,1	0.99	1.00	0.99	0.99	0.39
	0.5	0.0,1	1.00	0.00	0.00	0.00	0.00
	1	0.0,0.5	0.99	0.20	0.50	0.28	0.31
distilbert-base-nli-mean-tokens	0.0	0.5,1	0.99	1.00	0.99	0.99	0.36
	0.5	0.0,1	1.00	0.00	0.00	0.00	0.00
	1	0.0,0.5	0.99	0.21	0.50	0.30	0.32

supervised by the IITP(Institute for Information & communications Technology Promotion), IITP-2017-0-00655, NRF-2016K1A3A7A03951968, and NRF-2019R1A2C2090504.

REFERENCES

[1] Anne Geraci, Freny Katki, Louise McMonegal, Bennett Meyer, John Lane, Paul Wilson, Jane Radatz, Mary Yee, Hugh Porteous, and Fredrick Springsteel. IEEE standard computer dictionary: Compilation of IEEE standard computer glossaries. IEEE Press, 1991.

[2] Healthcare Information and Management Systems Society. Definition of Interoperability, 2013.

[3] Wajahat Ali Khan, Maqbool Hussain, Khalid Latif, Muhammad Afzal, Farooq Ahmad, and Sungyoung Lee. Process interoperability in healthcare systems with dynamic semantic web services. Computing, 95(9):837–862, 2013.

[4] Taqdir Ali, Maqbool Hussain, Wajahat Ali Khan, Muhammad Afzal, Jamil Hussain, Rahman Ali, Waseem Hassan, Arif Jamshed, Byeong Ho Kang, and Sungyoung Lee. Multi-model-based interactive authoring environment for creating shareable medical knowledge. Computer Methods and Programs in Biomedicine, 150:41–72, 2017.

[5] SNOMED Clinical Terminologies.

[6] LOINC.

[7] CIMI. Clinical Information Modeling Initiative (CIMI), 2015.

[8] Fahad Ahmed Satti, Taqdir Ali, Jamil Hussain, Wajahat Ali Khan, Asad Masood Khattak, and Sungyoung Lee. Ubiquitous Health Profile (UHPr): a big data curation platform for supporting health data interoperability. Computing, 2020.

[9] Sara Althubaiti, Şenay Kafkas, Marwa Abdelhakim, and Robert Hoehndorf. Combining lexical and context features for automatic ontology extension. Journal of biomedical semantics, 11(1):1–13, 2020.

[10] Kenji Nozaki, Teruhisa Hochin, and Hiroki Nomiya. Semantic schema matching for string attribute with word vectors. In 2019 6th International Conference on Computational Science/Intelligence and Applied Informatics (CSII), pages 25–30. IEEE, 2019.

[11] Aola Yousfi, Moulay Hafid El Yazidi, and Ahmed Zellou. xmatcher: Matching extensible markup language schemas using semantic-based techniques. International Journal of Advanced Computer Science and Applications, 11(8):655–665, 2020.

[12] Lev Bulygin. Combining lexical and semantic similarity measures with machine learning approach for ontology and schema matching problem. In Proceedings of the XX International Conference "Data Analytics and Management in Data Intensive Domains"(DAMDID/RCDL'2018), pages 245–249, 2018.

[13] Galih Hendro Martono and SN Azhari. Review implementation of linguistic approach in schema matching. International Journal of Advances in Intelligent Informatics, 3(1):1–9, 2017.

[14] Ali A Alwan, Azlin Nordin, Mogahed Alzeber, and Abedallah Zaid Abualkishik. A survey of schema matching research using database schemas and instances. International Journal of Advanced Computer Science and Applications, 8(10), 2017.

[15] Martijn G Kersloot, Florentien JP van Putten, Ameen Abu-Hanna, Ronald Cornet, and Derk L Arts. Natural language processing algorithms for mapping clinical text fragments onto ontology concepts: a systematic review and recommendations for future studies. Journal of biomedical semantics, 11(1):1–21, 2020.

[16] Li Xu and David W Embley. Discovering direct and indirect matches for schema elements. In Eighth International Conference on Database Systems for Advanced Applications, 2003.(DASFAA 2003). Proceedings., pages 39–46. IEEE, 2003.

[17] Jonas Moll and Åsa Cajander. Oncology health-care professionals' perceived effects of patient accessible electronic health records 6 years after

launch: a survey study at a major university hospital in sweden. Health informatics journal, 26(2):1392–1403, 2020.
[18] Xunjie Zhu, Tingfeng Li, and Gerard De Melo. Exploring semantic properties of sentence embeddings. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 632–637, 2018.
[19] Uri Kartoun. A methodology to generate virtual patient repositories. arXiv preprint arXiv:1608.00570, 2016.
[20] Lijun Pan, Xiaoting Fu, Fangfang Cai, Yu Meng, and Changjiang Zhang. Design a novel electronic medical record system for regional clinics and health centers in china. In 2016 2nd IEEE International Conference on Computer and Communications (ICCC), pages 38–41. IEEE, 2016.
[21] Akihiro Inokuchi, Koichi Takeda, Noriko Inaoka, and Fumihiko Wakao. Medtakmi-cdi: interactive knowledge discovery for clinical decision intelligence. IBM Systems Journal, 46(1):115–133, 2007.
[22] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 11 2019.
[23] Davide Chicco and Giuseppe Jurman. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. BMC genomics, 21(1):1–13, 2020.
[24] Adam Cohen. Fuzzywuzzy: Fuzzy string matching in python, Jul 8th, 2011.
[25] Kilem L Gwet. Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters. Advanced Analytics, LLC, 2014.
[26] Mary L McHugh. Interrater reliability: the kappa statistic. Biochemia medica, 22(3):276–282, 2012.

FAHAD AHMED SATTI is currently pursuing Ph.D. in Computer Engineering from Kyung Hee University, South Korea. He completed his MS in Computer Science from University of Trento, Italy in 2014. His primary research interest is directed towards providing a technical solution to resolve the lack of interoperability in information systems. In general, he is interested in domains of semantic matching, stream reasoning, knowledge extraction, and machine learning.

MUSARRAT HUSSAIN is a Ph.D. candidate in the Ubiquitous Computing Laboratory, Kyung Hee University, South Korea. He has completed his MS software Engineering in 2015 from National University of Science and Technology (NUST) Pakistan. His research interests include clinical text mining, knowledge extraction and representation, and machine learning.

JAMIL HUSSAIN received the Ph.D. degree from the Department of Computer Engineering, Kyung Hee University, South Korea in 2019. He worked as a postdoctoral researcher in the Ubiquitous Computing Laboratory, Kyung Hee University, South Korea, from Sep 2019 to Feb 2021. He is currently an Assistant Professor at Sejong University, South Korea. He has professional experience of over 7 years in the software industry, working on user experience design and development on various projects. His research interest includes user experience design, artificial intelligence, and information extraction from textual data.

SYED IMRAN ALI received the BS degree in Computer Science from IQRA University, Islamabad in 2008 and MS degree in Computer Science From National University of Computer and Emerging Sciences (NUCES) in 2012. He has taught in NUCES as an adjunct faculty. He is currently pursuing Ph.D. degree from Department of Computer Science and Engineering, Kyung Hee University, South Korea. His area of interests include machine learning, health analytics and data-driven systems.

TAQDIR ALI received the Ph.D. degree from the Department of Computer Engineering, Kyung Hee University, South Korea in 2019. He is currently working as a postdoctoral researcher in the Ubiquitous Computing Laboratory, Kyung Hee University, South Korea. From 2006 to 2011, he was a Senior Software Engineer, a System Analyst, and a Researcher in a reputable software house. His current research includes knowledge acquisition for clinical decision support systems, applications of machine learning, text processing, and e-health standardization.

HAFIZ SYED MUHAMMAD BILAL received the M.S. degree in computer science from National University of Sciences and Technology, Pakistan,in 2008. He is currently pursuing the Ph.D. degree in computer science and engineering at Kyung Hee University, South Korea. He has working experience of more than three years in data science and open source development and is actively involved in developing big data ecosystem for academic and health care. His research interests include behavior quantification & assessment, machine learning, and behavior modeling & adaptation.

TAE-CHOONG CHUNG received his B.S. degree in Electronic Engineering from Seoul National University, Republic of Korea, in 1980, and his M.S. and Ph.D. degrees in Computer Science from KAIST, Republic of Korea, in 1982 and 1987, respectively. Since 1988, he has been with the Department of Computer Science and Engineering, Kyung Hee University, Republic of Korea, where he is now a Professor. His research interests include machine learning, meta search, and robotics.

SUNGYOUNG LEE received the B.S.degree from Korea University, Seoul, South Korea,and the M.S. and Ph.D. degrees in computer science from the Illinois Institute of Technology,Chicago, IL, USA, in 1987 and 1991, respectively.He was an Assistant Professor with the Department of Computer Science, Governors State University, University Park, IL, USA, from 1992 to 1993. He has been a Professor with the Department of Computer Engineering, Kyung Hee University,South Korea, since 1993, where he has been the Director of the Neo Medical ubiquitous-Life Care Information Technology Research Center since 2006.He is currently the Founding Director of the Ubiquitous Computing Laboratory. His current research interests include ubiquitous computing and applications, wireless ad hoc and sensor networks, context-aware middle-ware, sensor operating systems, real-time systems and embedded systems,and activity and emotion recognition. He is a member of ACM.