Information Processing and Management 60 (2023) 103213

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

A semantic sequence similarity based approach for extracting medical entities from clinical conversations

Fahad Ahmed Satti [a,b], Musarrat Hussain [a], Syed Imran Ali [a,b], Misha Saleem [c], Husnain Ali [c], Tae Choong Chung [a,*], Sungyoung Lee [a,*]

[a] Department of Computer Science and Engineering, Kyung Hee University, Giheung-gu, Yongin 17104, South Korea
[b] School of Electrical Engineering and Computer Science (SEECS), National University of Sciences and Technology (NUST), Islamabad 44000, Pakistan
[c] Department of Neonatology, Care+ Medical Centre, G-8, Islamabad 44080, Pakistan

ARTICLE INFO

Dataset link: https://github.com/desertzebra/ClinicalConversations

Keywords:
Clinical data mining
Semantic similarity
Natural language processing

ABSTRACT

Clinical conversations between physicians and patients can provide a rich source of data, information, and knowledge. A plethora of tools and technologies have been developed to identify attributes of interest in unstructured text. However, identifying the name and correct value of an attribute, from real world data, in a timely manner is a nontrivial task. In this manuscript we present a novel pipeline using transfer learning, clinical concept dictionaries, and pattern matching to provide an end-to-end solution for identifying attributes and extracting their values from natural clinical text. On real-world data, with 1176 instances, we achieve an accuracy of 56.21%, which is 3% higher than the baseline methodology.

1. Introduction

Technological advancements in Information and Communication Technology (ICT) have provided a boost to the quality and quantity of healthcare services. A plethora of policies, software, and devices have been developed to introduce new and extend the reach of existing best practices (young Jung, Lee, & Hwang, 2022). While the benefits of these initiatives are well documented and acknowledged in the literature, in more practical terms, access to effective digital healthcare services in the developed world versus the developing world is skewed.

World Health organization (2021) has highlighted the most prominent operational factors influencing this difference, including the availability of specialists and equipment, for adapting, developing, and using global standards and technology. In high-income countries, spurred by the effectiveness of technology, variety of sources, and complexity of domain requirements, many novel platforms have been proposed and are being utilized to improve the clinical interactions (Beks et al., 2022; Coppersmith, 2022; Kaplan, 2020). On the contrary, in low-income countries, financial limitations, increased patient load, and the availability and access to healthcare experts, clinical facilities (public and private setups), in-patient care, internet, and electricity, have a very large impact on the healthcare services (Chandra et al., 2022; Furtado, Gygax, Chan, & Bush, 2022). Many commercial solutions are financially not feasible for low-income countries, and many open-source solutions, such as OpenEMR[1] or GNU Health,[2] are difficult to adapt, without substantial intervention by ICT experts.

* Corresponding authors.
E-mail addresses: fahad.satti@oslab.khu.ac.kr, fahad.satti@seecs.edu.pk (F.A. Satti), musarrat.hussain@oslab.khu.ac.kr (M. Hussain), imran@oslab.khu.ac.kr, imran.ali@seecs.edu.pk (S.I. Ali), dr.misha.saleem@gmail.com (M. Saleem), husnainalidr@gmail.com (H. Ali), tcchung@khu.ac.kr (T.C. Chung), sylee@khu.ac.kr (S. Lee).

[1] OpenEMR: https://www.open-emr.org/.
[2] GNU Health: https://gnuhealth.org/.

https://doi.org/10.1016/j.ipm.2022.103213
Received 14 July 2022; Received in revised form 28 October 2022; Accepted 24 November 2022
Available online 7 December 2022

An important requirement of any Healthcare Information Management Systems (HIMS) is the ability to create, store, and share Electronic Medical Record (EMR) for enabling the long-term management of patients and diseases (Dutta & Hwang, 2020). However, due to the challenges faced by healthcare services in the developing world, the encounters are partially recorded via offline methods such as registers and printed forms. Due to recent advances in technologies, particularly, Machine Learning, Optical Character Recognition (OCR) and Natural Language Processing (NLP) (Gasparetto, Marcuzzo, Zangari, & Albarelli, 2022), it is now possible to develop a bridge between the existing HIMS and the challenging requirements of healthcare service delivery in the developing world. Physical interactions in the real world can now be digitized and converted into data, information and knowledge, which can remove the redundancies, associated with the traditional data collection ways of the HIMS (Ismail, Materwala, Karduck, & Adem, 2020).

In this manuscript, we present a novel methodology based on semantic similarity to convert unstructured text obtained from manually transcribed and translated (from Urdu to English) clinical conversations into a semi-structured format. The audio conversations were recorded in a non-intrusive manner to ensure that the clinical content captured in these real clinical encounters was kept natural and unbiased. Thus the unstructured text produced from this data does not follow a preset structure and remains largely informal.[3] Devising a methodology for extracting semi-structured data from this text necessitates the design of a specialized pipeline, which can not only extract the relevant and contextual data but also do so in a timely manner. In particular, in the first phase of our methodology we generate the sequences which form the atomic part of our data. In the second phase, we created a set of medically aligned sequences (MASS) which contain the necessary artifacts for extracting attributes and their values from unseen sequences, and classified the unseen sequences against the MASS using a fine-tuned DistilBERT base uncased model.[4] In the third phase, we extract the attribute-value form of data from classified instances, using syntactic and conceptual semantic matching. This final output forms the semi-structured clinical data which can be aligned with HIMS schema to automatically record the clinical data.

In order to support our methodology and evaluate its performance, we have utilized 167 clinical conversations from two hospitals in Pakistan. Our proposed model achieves an accuracy of 56.21% which is better than the baseline methodology (with accuracy of 53.23%). These results indicate the correctness of our approach, towards identifying semi-structured data from clinical conversations in unstructured text form. However, many other research interventions are still required to complete an End-to-end production ready system, which will benefit the developing world greatly and jump start standard compliance.

Subsequent sections in this manuscript start with Section 2, briefly describes various prior research initiatives. Section 3 then introduces the aim of this study, followed by Section 4 detailing our methodology. Next we present the experimental setup in Section 5 and their results in Section 6. Section 7 discusses some of the problems and limitations of our approach and Section 8 concludes this paper.

2. Related work

Medical information extraction is a challenging task of automatically deriving high-quality structured information from text. Several research initiatives have aimed to solve related problems and achieved very good results. Some of these will be explained in this section.

Automatic keyword extraction has gained a lot of traction in the research community (Nasar, Jaffry, & Malik, 2019) as it pertains to extracting potential information from raw textual data with minimum human intervention. The Named Entity Recognition (NER) task in NLP is also related to this problem, whereby techniques are devised to build models for identifying attributes of interest, according to some preset features (such as identifying all persons, cities, and others), in text (Putthividhya & Hu, 2011; Yan, Gui, et al., 2021). Similarly, the task of identifying sequences in text, pertaining to some input attribute names (similar to preset features) is also related to NER (Wang et al., 2020; Xu, Wang, Mao, Jiang, & Lan, 2019).

In the past, rule-based approaches, such as Chiticariu, Krishnamurthy, Li, Reiss, and Vaithyanathan (2010), Chiu and Nichols (2016), and Vandic, Van Dam, and Frasincar (2012) have been proposed, which typically used regular expressions obtained from domain knowledge. However, these techniques, suffer from the generality problem and are unable to replicate their performance, for any text, syntactically and semantically different from the source, as shown by Chiticariu et al. (2010). Ontologies and semantic web-based solutions can help resolve the semantic matching problem, however, these solutions require a large amount of human effort to build the semantic knowledge graphs. Zheng, Mukherjee, Dong, and Li (2018) proposed a methodology to extract missing attribute values from a free text input such as product profiles. The methodology can leverage open-world assumptions in which case the possible set of values are not known beforehand. Si, Wang, Xu, and Roberts (2019) demonstrated that neural network-based representations e.g. word2vec, Glove, fastText, dramatically improve the performance of natural language processing tasks such as concept extraction.

Recently, various transformer-based solutions have been proposed which utilize semantic similarity to identify the attribute and their values in unstructured text (Chew et al., 2022; Wang et al., 2020). Solutions, such as Adatag, proposed by Yan, Zalmout, et al. (2021) is able to extract multiple attribute values. AdaTag uses adaptive decoding in which the decoder is parameterized with pre-trained attribute embeddings through a hyper network and a Mixture-of-experts module. This allows for separate, but semantically correlated, decoders to be generated on the fly for different attributes. Mehta, Oprea, and Rasiwasia (2021) proposed a high precision

[3] Anonymized data and the code supporting our methodology is available at https://github.com/desertzebra/ClinicalConversations.

[4] https://huggingface.co/distilbert-base-uncased

and scalable framework for extracting numeric attributes from product description text. A distant supervision approach is used for training data generation and removing dependency on manual labels. Moreover, a multi-task learning architecture is proposed to deal with missing labels.

Identification of attributes and their values is of special interest, in the domain of e-commerce, where various solutions have been proposed to identify the implicitly defined characteristics of a product, from its description (Roy, Goyal, & Pandey, 2021; Yang et al., 2022; Zheng et al., 2018). Thus, retailers can continue describing their products in a poignant manner, while businesses can use state-of-the-art tools and techniques to identify the key features, necessary to forecast demand, optimize search, and provide contextual recommendations to the buyers

In the domain of medical NER, machine learning techniques are gaining traction due to their increased accuracy and ability to generalize solutions. Early on, Aramaki et al. (2009) proposed a system that converts medical text from patient discharge summaries into a table structure using medical event recognition and an SVM-based negative event identification. Wang et al. (2018) explored the application of neural network-based models to produce word embeddings for biomedical text. It is demonstrated that these embedding approaches generate vector representations that capture useful semantic properties and linguistic relationships between words. Unstructured text often consists of typographical errors and abbreviations, which act as an impediment to improving the performance of the word embedding-based approaches. Narayanan, Rao, Prasad, and Das (2020) presented an approach based on Bidirectional LSTM (Bi-LSTM) with character level embeddings to avoid this problem and achieve better performance. Similarly, Zeng, Sun, Lin, and Liu (2017), used a Bi-LSTM and CRF-based architecture to identify drug names, using both word-based and character-based representations of each word. Du, Wang, Wang, and Xu (2020) proposed a deep learning-based approach to extract medically relevant attributes from electronic medical records, using ALBERT model, which provides much better results than the traditional LSTM-CRF model. Sun et al. (2021) have formulated a biomedical entity recognition task as a machine reading comprehension problem, which achieves good performance on six BioNER datasets. The proposed formulation can introduce more prior knowledge through well-defined queries.

Collecting the appropriate clinical conversations, which can be used to extract semi-structured data, is a challenging task. Fortunately, with the support of clinical practitioners in this study we were able to collect and evaluate our methodology on real world data. However, the data was collected in Urdu language, where NLP techniques are not able to obtain performance, comparable to other languages, especially English (Andrabi & Wahid, 2022; Kanwal, Malik, Shahzad, Aslam, & Nawaz, 2019). Resources, such as WordNet, terminological dictionaries, corpora, and others are not readily available for Urdu Language (Daud, Khan, & Che, 2017). The current literature points towards numerous initiatives towards automatically translating English to Urdu, however, the inverse case of translating Urdu to English is still comparatively novel (Masroor, Saeed, Feroz, Ahsan, & Islam, 2019). This gap between the spoken language used for clinical interventions and the NLP tools has led to some recent initiatives towards building specialized datasets and preparing appropriate tools for solving real-world problems. Within the clinical domain, Chiu, Villena, Martin, Núñez, and Dunstan (2022) have designed Word2Vec and fastText based embedding vectors, which operate on clinical referrals in Spanish language. In order to validate their methodology, the authors translated English language validation datasets into Spanish, calculated the embedding vectors pertaining to the translated text and compared the accuracy of the word embedding with their original English language counterparts. The authors have highlighted some of the key issues in obtaining accurate translations, which include terminological, grammatical, and functional inadequacies.

In summary, research trends have moved from rule-based syntactic matches to supervised learning, then to unsupervised learning, and to most recently hybrid learning methods leveraging syntactic and semantic matching.

3. Research objectives

Theoretically, our proposed methodology is similar to other methodologies, such as the one presented by Abdullah and Ahmad (2013), which focuses on extraction of metadata, classification and clustering of data, and mapping the data onto the target schema. However, in creating a practical solution, many challenges were encountered, which require novel interventions and assumptions to simplify the problem space. In particular, we are interested in resolving three challenges, which are defined as follows.

- *Challenge 1*; The first challenge is to identify the key domain elements which can relate to an attribute's name. This name can either be a textual identifier from the consuming schema (such as a database or web service) or a generic identifier (such as from a domain adapted concept dictionary), which can be mapped onto the storage schema.
- *Challenge 2*; We should also identify the portion of the input text which corresponds to an identified attribute's name and holds its value.
- *Challenge 3*; Lastly these identifications should take into account, the time required to verify the contents. Essentially, while it is possible to classify each word and group of words as a valid attribute's name and its value, the resultant dataset would be too large, contain many incorrect results and would greatly increase the verification time for the physician.

In order to fulfill our objective, focusing in particular on the above mentioned challenges, we have developed a sequential pipeline, which applies semantic matching and transformative functions on a specialized dataset, to transform unstructured text into a semi-structured form (with pre-determined attribute keys only). The novelty of this applied solution lies in developing an end-to-end methodology (albeit with manual interventions) for solving a real world problem, while utilizing and re-purposing various state-of-the-art tools and technologies.

The main contribution presented in this research work is as follows.

Table 1
Notations used in the manuscript.

Term	Definition
S	Set of Sequences
E	Set of Enriched Sequences
H	Set of Sequences used for threshold selection
T	Set of Test Sequences
P	Set of Probable Medical Sequences
A	Set of Key–Value pairs extracted from Sequences
M	The medical schema used to identify the target attribute names
$\vec{V}$	The embedding vector produced by encoding a sequence.
l	A label indicating the textual key and the expected value
x	A methodology to extract values from a sequence
α	Threshold used for sequence classification
β	Threshold used for filtering similar attributes

- *Data Acquisition*; Firstly, our primary source of data is the interaction between physicians and patients or their guardians. For this we recorded short conversations between physicians and patient/guardians from two hospitals in Pakistan, specializing in pediatric care. This included District Headquarters Hospital, Kotli, Azad Jammu and Kashmir, Pakistan (DHQ-Kotli) and Care+ Medical Center, Islamabad, Pakistan (Care+ MC-Islamabad).
- *Data Pre-Processing*; Secondly, we converted the translated conversations into sequences, which represent a unit of conversation, in the form of a question and its answer, or a statement.
- *Model Development*; Thirdly, using transfer learning methodology and using real data we have created the Medically Aligned Sequence Set (MASS), which contains 322 instances. Each instance holds enough data to classify unseen sequences, identify an appropriate attribute's name, and an extraction methodology to obtain its corresponding value.

Appendix A shows the textual form of a sample conversation.

Once an appropriate attribute's name and value have been identified we can then transform the structured contents of each conversation into a relational schema, designed for an HMIS. The particular data interoperability methodology, for matching the attribute names is based on our previous work presented in Satti et al. (2021). The final key–value, compliant with a consuming platform (such as a database system or a form) can then be presented to a human expert for validation, before it is stored.

4. Methodology

In order to convert the unstructured input text into structured schema elements, we have developed a pipeline, comprising of various transformation, matching, and filter processes. Throughout this manuscript, we have used many terms and notations to simplify the explanation. A brief overview of the notations used in this text are presented in Table 1. Additionally, some of the most important terms are briefly explained here.

- **Sentence**;
  - (within manuscript write-up) Based on the definition by merriam-webster,[5] a sentence is a collection of one or more words, forming a syntactic unit, which can be used to ask a question, provide an answer, and present an assertion or an instruction. In written form, a sentence should end with punctuation (such as question mark, period, semi-colon).
  - (in the context of sentence-similarity) used to describe a famous NLP task of determining the similarity between two texts. While our task is similar to sentence similarity, in order to avoid confusion, we shall call it sequence-similarity, where required.
- **Sequence**; A sequence is a collection of one or more sentences, with at least two words (to support the lookup of key–value pairs). In particular, a sequence can contain, a question and its answer, a question followed by another question, an assertion or instruction, or phrases from the sentence, split on "and" or ","(comma).
- **Medical Sequences**; A sequence containing at least one key–value pair, where the key is a medical concept such as "Finding", "Disease" or others. A probable medical sequence contains computed key–value pairs, while a valid medical sequence is validated by a human expert.

Fig. 1, presents an abstract view of our proposed methodology, which utilizes as input, unstructured text obtained from conversations between physicians and patients (presented in Section 4.1). These conversations are then pre-processed (presented in Section 4.2) to build manageable sequences (S), which are in turn used as the input data for the next steps. In particular, these sequences are used to fine-tune the pre-trained DistilBERT-base-uncased model, creation of MASS, threshold selection, and test set creation. Using semantic similarity in the Sequence Classification phase (presented in Section 4.3), we filter the unseen sequences

[5] https://www.merriam-webster.com/dictionary/sentence

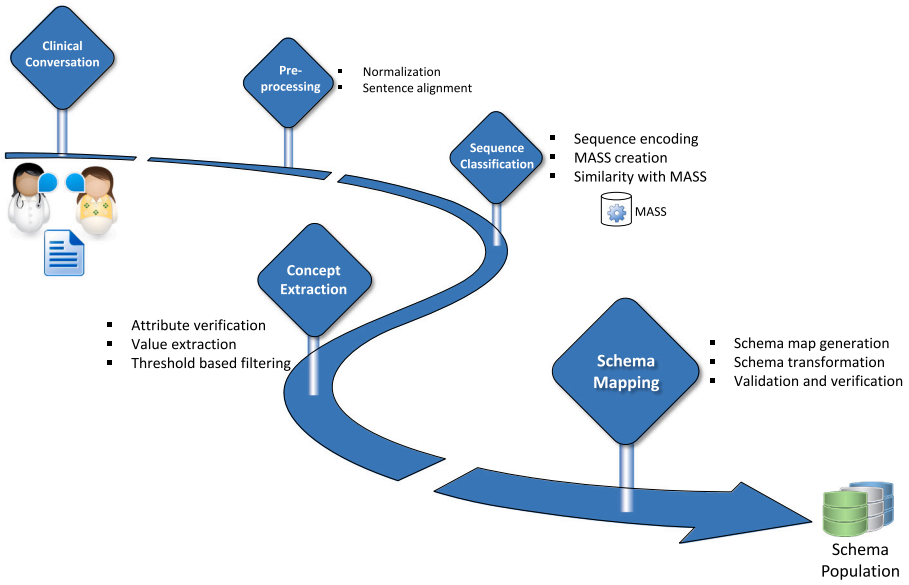

Fig. 1. An overview of the proposed methodology.

producing the probably medical sequences (P). Next, in the Concept Extraction phase (presented in Section 4.4), we verify the identified attributes and extract their values using artifacts from P to apply a syntactic or semantic method. At the end of this phase semi-structured attribute–value pairs(A) are produced which form an input for the Concept Schema Mapping phase (presented in Section 4.5) and be subsequently passed on to a data store for storage.

4.1. Prerequisite: Clinical conversation to unstructured text

The Clinical conversations gathered as a part of this study, represent the verbal interaction between physicians and patients or guardians (for young patients). The audio recordings of these conversations were obtained in a non-intrusive manner after obtaining consent from all participants. Since the primary language of communication between the participants is Urdu (National Language of Pakistan), it would be most beneficial and productive, to use an appropriate and automatic method for transcribing the audio contents. However, the current state of Urdu NLP research, especially in the clinical domain, is not mature enough to appropriately perform the necessary conversions (transcription and translation), automatically. Additionally, our methodology relies on the existence of clinical concept dictionaries and conceptual semantic matching to identify the attributes and their values. Even after extensive literature review, we were unable to find any concept dictionary in Urdu, which works well within the clinical domain. On the other hand, Unified Medical Language System (UMLS)[6] (Bodenreider, 2004) provides a very good service to identify the semantic concepts behind clinical terms in English. Thus, in order to achieve our aim to identify relevant medical data from clinical conversations, we have utilized manual intervention to transcribe and translate the clinical conversations into English.

4.2. Pre-processing

In the first phase, unstructured text representing, the transcribed and translated conversations are converted into sequences that can contain the clinically relevant attributes and their values. Here, we made two assumptions which are described as follows.

- *Assumption 1- Compound Sentences*; The original conversations between the participants include compound sentences, which can provide an answer with multiple clinical artifacts, such as "**** age 3 months and she has some pain dont know where". In these cases even if we are able to partially predict the answers as correct, they are considered completely correct.
- *Assumption 2*; The conversations either contain medical sequences of the form Question–Answer or Statements (Instructional or Assertive), containing both key and its corresponding value. These two types are explained below.
  - In the Question–Answer type sequences the attribute name lies in the question part, while the value lies in the answer part, such as for the sentence, "What is his name?" the attribute name is "name", while its value is found in the statement by the patient/guardian.

[6] https://www.nlm.nih.gov/research/umls/index.html

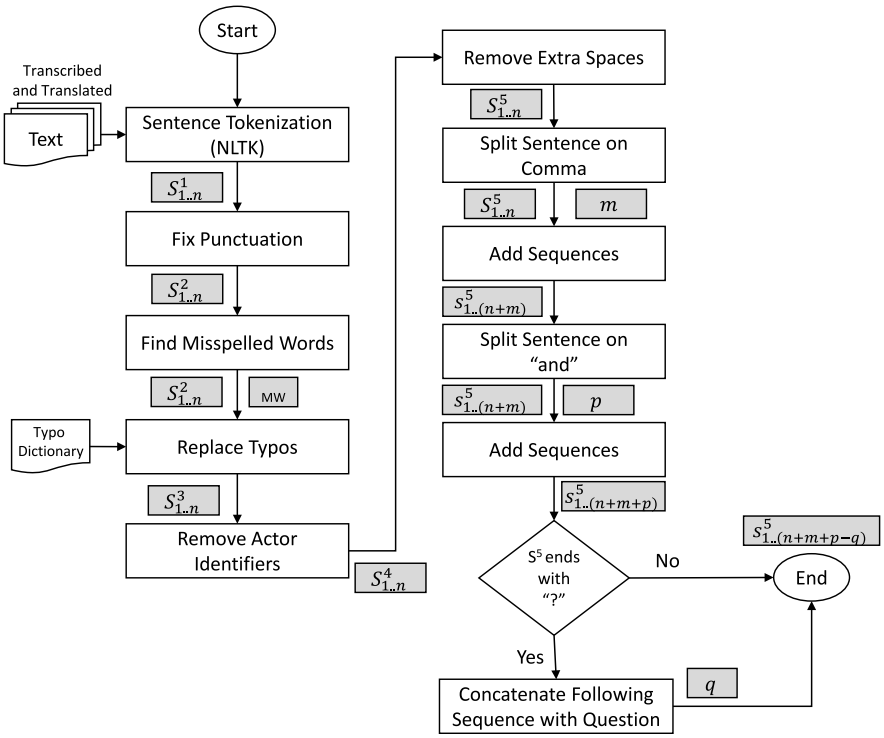

Fig. 2. A flowchart representing the Preprocessing step.

– The statements given by physicians and patients/guardians may contain only values in the text, such as in the sequence, "Fever is a little bit", the key can be "Finding" and its value "fever".

The detailed flow chart for pre-processing is shown in Fig. 2, which process the input clinical conversations to eventually produce a set of sequences S. The transcribed text is first split into sentences ($S^1_{1..n}$) using the Natural Language Toolkit (NLTK)[7] library in python. Next, we fixed the punctuation to fix some human errors in placing the punctuation marks (such as adding a space before punctuation, no space after punctuation, various quote type usage, and others) to align the syntax of the sentences. Here the second transformation is applied, while the number of sentences remains the same, producing $S^2_{1..n}$.

In order to resolve some of the typographical errors (Typos) and incorrect spellings by the transcribers we utilized a spell checker (based on a blog post by Peter Norvig[8]) to identify the misspelled words MW in the training sequences. Since the default spell corrector is well suited for a general domain and to avoid any inaccuracies that may be caused by incorrectly replacing a clinical term, we used a custom typo dictionary. The corrections in this dictionary were validated by a clinical expert before their applications to produce the updated sequence set. Next, we removed the actor identifiers and multiple spaces in each sequence. Reluctantly, we obtain the set of sequences $S^5_{1..(n)}$, of which only content has changed but the total number of sequences still remain the same.

Acting on the *Assumption 1* of fixing some problems with compound sentences, we split the sentences on "," and "and", producing additional sub-sequences. If the length of the sub-sequence was greater than one, we added it to the set of sequences. This would add an additional m and p sentences into the set of sequences, respectively. The original longer sequence still remains a part of this set. The eventual output at the end of this step is the set of sequences $S^5_{1..(n+m+p)}$.

To resolve the *Assumption 2*, we selected the sequences ending with "?", and concatenated the next sequence with this sequence. This additional sequence does not have an independent existence. Thus, the q answers to the sequences ending with "?", are removed from the set, producing $S^5_{1..(n+m+p-q)}$, which we shall simply call S, henceforth.

4.3. Sequence classification

In this phase the aim is to classify unseen data and obtain the probable medical sequences. The overall methodology is shown in Fig. 3. The two steps involved in this phase are discussed in the following sub-sections.

[7] https://www.nltk.org/
[8] https://norvig.com/spell-correct.html

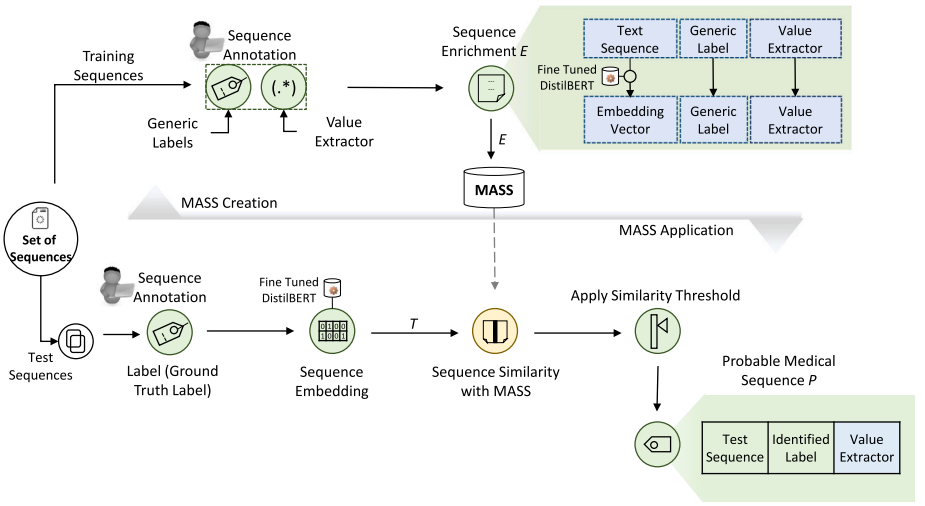

Fig. 3. Workflow for classifying the sequences as medically aligned or not.

4.3.1. MASS creation

MASS represents the set of medically aligned sequences, which have been predetermined as interesting by a clinical experts. Each instance d of MASS contains a generic form of the sequence with special tags (“[CLS]”, “[SEP]”, and “[MASK]”) which is used by our Fine-Tuned DistilBERT-base-uncased model (further explained in Section 5.2) to encode and produce the embedding vector $\vec{V}$. The instance also contains labels l identifying words and phrases within the source sequence. These labels are based on the semantic types, included in UMLS (such as Diagnostic Procedure, Disease or Syndrome, Finding, Sign or Symptoms, and others) and other generic tags (such as name, age, duration, and others). Finally, a value extraction method x is attached to this pattern, which can be used to extract the value from a target, unseen, data instance through the use of UMLS lookup or the application of a regular expression.

When MASS is loaded into memory, each instance d is converted into an enriched sequence e, which is represented in Eq. (1). This includes the embedding vector, produced by encoding the sequence ($encode(d.sequence) \rightarrow \vec{V}^E$), the label l, and the extraction methodology x. The actual text of the sequence is not used.

$$e = \left\langle \vec{V}^E, l, x \right\rangle | e \in E \tag{1}$$

The extraction methodology can be one of “UMLS Lookup” or “Regular Expression”. The “Regular Expression” methodology is used to identify the value for concepts such as “name” and “age” from the test sequence. These patterns are manually built using the sequences found in MASS, during the annotation process. The intuition behind using these patterns is to allow value extraction from sequences that do not contain specialized medical concepts. On the other hand, many sequences contain medical concepts, such as “fever”, “cough”, “flu” and others, which can be found using the UMLS API. The textual part of the test sequence d is split into unigram and bigram tokens, which are sent to the approximate search API of UMLS. The API returns a list of semantic concepts which may be associated with the search term. The clinical expert, then determines, which tokens are valid, within the context of their associated sequence.

Using a threshold selection process, based on Area under Received Operating Characteristics (AuROC), the optimal threshold α for semantic similarity classification of any unseen sequence with MASS was obtained.

4.3.2. MASS application

Unseen data obtained during the execution of the methodology is also pre-processed to produce the set of sequences, first. These sequences are then encoded to produce embedding vectors using the Fine-tuned DistilBERT-base-uncased model ($encode(t) \rightarrow \vec{V}^T$). The resulting instance, however, contains both the raw text t of the sequence and the $\vec{V}^T$. Here the embedding vector is compared with all the embedding vectors from MASS ($\vec{V^E}$) to identify the similarity score between the existing medically aligned sequences and this unseen data. This comparison is performed using cosine similarity, as shown in Eq. (2), which assigns a score between 0 and 1 to the pair.

$$sim_t = \frac{\vec{V}^T \cdot \vec{V}^E}{\sqrt{\vec{V}^T \cdot \vec{V}^T}\sqrt{\vec{V}^E \cdot \vec{V}^E}} \tag{2}$$

The output of the Sequence Classification phase is the Probable Medical Sequence P, which includes the text of the test sequence t, identified label l from the MASS instances, with highest similarity above the threshold α, and their respective value extractors.

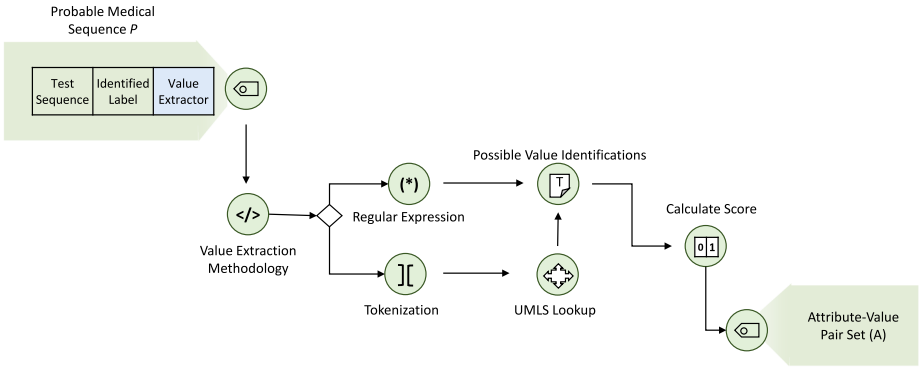

Fig. 4. Workflow for verifying the attributes and extracting the values.

4.4. Concept extraction

For each probable medical sequence $p \in P$, the concept extraction phase verifies the associated label as a valid attribute name, and extracts the value by applying the extractor function. This process is shown in Fig. 4.

If the extraction methodology is based on the semantic method, once again, UMLS is queried with the textual part of the test sequence t to obtain a list of semantic concepts. By matching the semantic type returned by UMLS and the predicted attribute name from p, we can determine the correctness of the label. Additionally, the token used for the query is the value associated with the test sequence.

In case of the syntactic extraction method, regular expressions are used to verify the correctness of the attribute name, if a RegEx group can return a value (such as named group "Age" returning a value). Thereby, the attribute name is statically associated with the text, while the token(s) obtained from named regex group, is its value.

$$a = \left\langle t, l^E, \eta \right\rangle | a \in A \wedge \eta \leftarrow x^E(d) \tag{3}$$

The output of this phase is the set of attribute–value pairs A.

4.5. Schema mapping

Each instance of P, can produce zero or more key–value pairs which, become a part of the set A. This set is a semi-structured representation of the conversations between physicians and patients/guardians. However, this representation is very different from the database or API structure, making it difficult to connect the methodology up to the previous step, with any real application. In our prior work, we have discussed the issues underlying healthcare data interoperability in Satti, Ali, Hussain, Khan, Khattak, and Lee (2020) and introduced our novel semantic reconciliation methodology to map heterogeneous data schemas using BERT-based sequence encoding and semantic similarity measurement in Satti et al. (2021). A visual representation of this process to convert A into one of the target database compliant schemas in M in Fig. 5.

The process starts by first breaking the attribute names into suffixes using suffix array generation from the start of the string to the end (forward suffix generation), from end of the string to the start (backward suffix generation), and regular expression based suffixes (which splits the text into capital letters and special characters). The three arrays are then combined and re-sorted alphabetically. Then for each item in the consolidated suffix array, we query UMLS, to collect the concepts associated with it. In case, no associated concept is found for the substring, it is dropped from the suffix array. On the other hand, the found concepts are all appended to the suffix. As a result, a sequence with the suffixes of the original text and the concept of the suffixes is formed. The newly formed sequence is then encoded to create an embedding vector, which can be compared with other sequences. We then compare each attribute name from $a \in A$ with the attribute name of one of many target schema, using cosine similarity of encoded suffixes + concept sequences. Using a pre-determined threshold value β we can classify a as similar or dissimilar. The similarity matrix then shows the relationship between our general schema and a specific schema belong to $m \in M$. Thus through this methodology we can then identify the target database schema, relation, and attribute, for each source instance of the key–value pair a.

5. Experimental setup

5.1. Data acquisition

In order to collect and prepare the initial conversational dataset, we first obtained official consent of the two participating medical centers (DHQ-Kotli and Care+ MC-Islamabad) in Pakistan to collect data for this study. Two practicing physicians, then

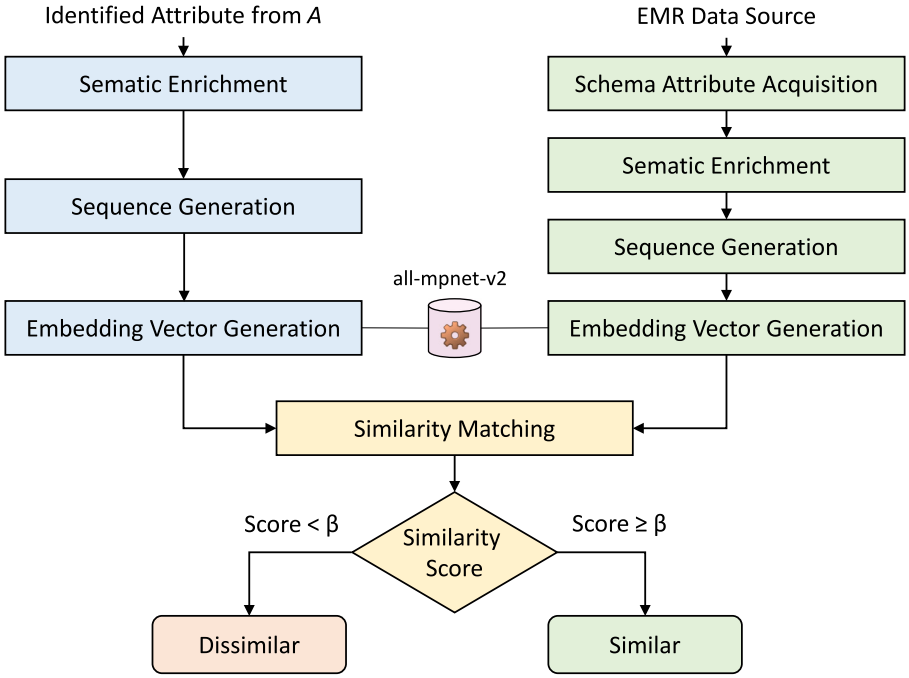

Fig. 5. Methodology for semantic reconciliation between the attribute set (A) and The EMR schemas (M).

recorded their conversations with patients and guardians at these hospitals. Overall, 148 unique clinical interactions were collected from DHQ-Kotli and 19 from Care+ MC-Islamabad. Each participant signed a consent form before the start of the conversation and was explained the necessity of this research work, verbally. Since the conversations between the physicians and patients/guardians were conducted in a local language (Urdu), three human transcribers were hired to transcribe the contents of each conversation, and translate it into English. The transcribers were supported by two practicing clinicians to resolve any ambiguities in the data. Two transcribers processed 74 audio files each, while one processed 19 conversations. All three transcribers were female with at least 14 years of education.[9]

The data and source code used in this study are available at https://github.com/desertzebra/ClinicalConversations.

5.2. Model training for sequence encoding

In order to convert the textual sequences obtained from conversations into embedding vectors, optimized for sentence similarity in the medical domain, we fine-tuned the DistilBERT-base-uncased model. To create this dataset, we first created a combination of the sequence set with itself ($S \times S$), to produce a set of unique sequence pairs. With 508 sequences in S, the combination set produced 129,272 pairs. For each pair, we then manually marked the two sequences as similar if they were intuitively equal and dissimilar otherwise. A pair of sequences, such as "how old is he? 5 years" and "whats her age? She's 15 years old" are semantically similar, however, the pair "what is child's name? h*****" and "the child has cough" are dissimilar. This produced a set of 6464 similar sequences. We then randomly selected 6464 dissimilar sequences from this set to produce a balanced dataset of 12,928 pairs. These pairs were further split into 70% training instances and 30% validation ones.

We tested various hyperparameters,[10] to optimize the sentence similarity evaluation, eventually selecting the batch size of 32, the "Sparse Categorical Cross Entropy" loss function, "Sparse Categorical Accuracy" as the evaluation metric, and AdamW optimizer (Loshchilov & Hutter, 2019), with an initial learning rate of 1e−4, 10% warmup steps, and 12 epochs. As a result of this fine-tuning activity, our model shows an accuracy of 95% on the test instances.

5.3. Threshold selection for sequence classification

In order to determine the optimal cosine similarity above which a test sequence can be classified as semantically similar to MASS, we evaluate 100 thresholds between 0.0 to 1.0 with a step size of 0.01. At each step, we calculate the AuROC score. Threshold

[9] Annotator 1 has a Bachelors in Business Administration from Bahria University, Islamabad, Pakistan.
Annotator 2 has a Bachelors in Computer Software Engineering from Foundation University, Rawalpindi, Pakistan.
Annotator 3 has a Bachelor of Dental Surgery from Lahore Medical and Dental College, Lahore, Pakistan.

[10] The details of the fine-tuning process is left out to keep this manuscript concise.

Table 2
Dataset division in terms of its usage.

Sentences	Activity
508	Model fine-tuning & MASS creation
464	Threshold selection
1,176	Testing sequences

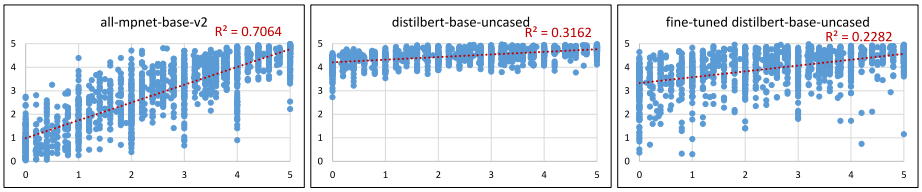

Fig. 6. Squared Pearson Correlation (r^2) of semantic sequence similarity between the annotated similarity and similarity computed by (a) by pretrained all-mpnet-base-v2 model, (b) pretrained DistilBERT-base-uncased model, and (c) fine-tuned DistilBERT-base-uncased model.

($\alpha_{proposed}$) at 0.87 is the point, where auROC is maximized for our custom DistilBERT-base-uncased model, while $\alpha_{baseline}$ for the baseline model is 0.49. Hence, for all evaluations, when a test instance has cosine similarity equal to or greater than α, with any instance from MASS, it is considered as medically aligned. A more detailed explanation of the threshold selection process is present in Appendix B.

6. Results

In order to evaluate the correctness of our methodology, and its conformance to the challenges stated in Section 3, we conducted several experiments. Some of the most important results are presented as follows.

6.1. Pre-processing

The conversational instances, in text form, were pre-processed to convert them into the set of sequences S. The text data was then anonymized and the introductory explanations of the study were removed. The text is then pre-processed to produce the set of sequences, required by our methodology. The division of the dataset for various operations and phases of the proposed methodology is described in Table 2. In particular, the conversations were split into three parts, with 508 sequences used for fine-tuning DistilBERT-base-uncased model and MASS creation, 464 sentences for threshold selection, and 1202 unique sentences, with labeled truth identifiers for attributes and their values.

6.2. Evaluation of the sequence similarity model

In order to evaluate the performance of our fine-tuned DistilBERT-base-uncased model, we have used the Semantic Textual Similarity benchmark (STSb) dataset (Cer, Diab, Agirre, Lopez-Gazpio, & Specia, 2017). The test dataset[11] contains, 1379 sentence pairs, which have been built from news items (500 instances), captions (625 instances), and forum (254 instances).

For each sentence pair instance, we first encoded the textual sentences to create embedding vector using[12] the pretrained all-mpnet-base-v2 model, the pretrained DistilBERT-base-uncased model, and the proposed fine-tuned DistilBERT-base-uncased model. We then calculated the cosine similarity between the embedding vectors. The similarity measure is then rescaled from 0–1 to 0–5, so as to identify the annotated labels for each sentence pair. Then we calculated the Pearson Correlation (r) between the computed similarity and the ground truth for the STSb dataset. The final results on the STSb test dataset for squared Pearson Correlation (r^2) are shown in Fig. 6(a)–(c). On this cross-domain dataset, the performance of the pretrained all-mpnet-base-v2 model at r^2 of 0.70, far exceeds the pretrained DistilBERT-base-uncased model at r^2 of 0.31, which is itself higher than the fine-tuned DistilBERT-base-uncased model at r^2 of 0.22.

[11] http://ixa2.si.ehu.eus/stswiki/index.php/STSbenchmark
[12] Without downstream training.

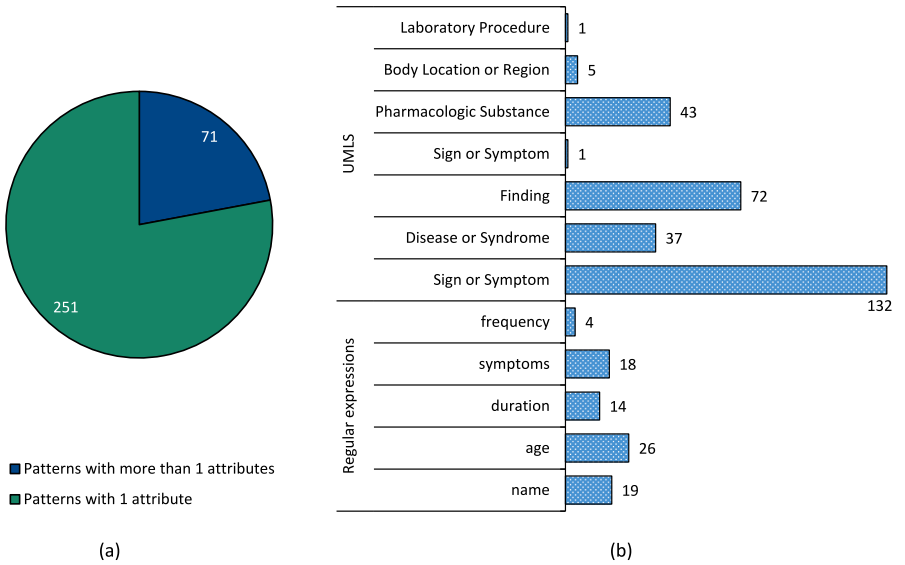

Fig. 7. Statistical information for MASS instances (a) showing the ratio of e with single vs multiple attributes, and (b) showing the unique labels and their extraction methodology.

6.3. Model development

In order to create MASS, we processed 508 sentences and created 322 true instances. Each instance is a partial representation of some sequences from the training dataset, marked by identifiers, such as [CLS], [SEP], and [MASK]. Thus the sequence, "what is your issue? sir i am having severe flu and cough along with little fever", has the following corresponding enriched sequence,

"[CLS] high fever or general fever? [SEP] just cough and sore throat;;Sign or Symptom,Sign or Symptom;;umls".

Here the three elements are separated by ";;", where the first is used for generating the embedding vector, the second contains the labels ("Sign or Symptom"), and the third part contains the extraction function ("umls").

The distribution of labels in MASS is shown in Fig. 7(a) where 251 instances have only one label and 71 have multiple labels. The count of unique labels in MASS is shown in Fig. 7(b), where 81 instances utilize regular expressions for five types of labels, while the 7 label types from UMLS are repeated 291 times.

6.4. Validation on test dataset

In order to evaluate the performance of our proposed methodology in terms of correctly identifying the attribute and its value, we used the Test dataset, which contains 1176 unique sequences. A sequence with at least one associated attribute-value pair is considered to be in the "True" class, of which there are 610 instances. While a sequence without any associated attribute-value pair is in the "False" class, which contains 566 instances.

Using our fine-tuned DistilBERT-base-uncased model, we encoded each of the unseen test sequence and matched them with the MASS. With a minimum similarity threshold $\alpha_{proposed}$ of 0.87, we collected all matched instances with the highest similarity, to produce the set of probable medical sequences. In this way, 837 instances were produced by the sequence classification process, where 648 test sequences were classified as medically probable.

We then apply the value extraction function to these instances. In case of UMLS, we first extract the unigram and bigram tokens from text associated with the sequence. We then query UMLS to identify the concepts associated with the token. The list of concepts returned by UMLS are then used to verify the correctness of the attribute associated to the probable medical sequence. If the attribute marked by MASS for a test sequence, exists in the concept list sent by UMLS, the predicted attribute is verified as partially correct. Next, we compare this attribute with the actual label of the test sequence to determine complete correctness of the attribute. Then the token used to query the UMLS is used to validate against the actual value of the test sequence. If both the attribute name and the value are correct, the predicted class for this test sequence is "True", otherwise it is considered "False". The process is repeated for all attributes, labeled by MASS. In case any actual label is not identified, or MASS predicted label is incorrect, or a value is not identified, the test sequence is again considered "False".

Finally, our proposed method is able to achieve an accuracy of 56.21% as shown by "Proposed Approach" in Table 3.

To compare our results with a baseline model, we replaced our fine-tuned DistilBERT model with the pre-trained DistilBERT-base-uncased model and the pre-trained "all-mpnet-base-v2" sentence similarity model. The DistilBERT-base-uncased model is trained on the BookCorpus dataset containing 11,038 unpublished books and English Wikipedia dataset, while the all-mpnet-base-v2 model

Table 3
Performance evaluation of the proposed methodology on test instances and its comparison with the baseline methodology.

Methodology	Sequence ClassificationThreshold	Accuracy	Precision	Recall	F1 score
Fine-Tuned DistilBERT-base-uncased (Proposed Approach)	0.87	**56.21**%	**65.47**%	32.95%	43.84%
DistilBERT-base-uncased (Baseline Approach)	0.85	47.75%	49.30%	46.39%	47.80%
all-mpnet-base-v2 (Baseline Approach)	0.49	53.23%	55.81%	**47.21**%	**51.15**%

is trained on over 1 billion tuples and provides the best results[13] for Sentence Embeddings (69.57% on 14 diverse datasets) and Semantic Search (57.02% on 6 diverse datasets) tasks. Using the same strategy of identifying the attribute name and value correctly, we evaluated the results of the baseline, referred to as "Baseline Approach" in Table 3.

These results show that the proposed method shows better accuracy and precision than the baseline methodologies, while in terms of identifying a large number of correct test sequences the all-mpnet-base-v2 based baseline methodology shows better performance.

7. Discussion

7.1. Results and their implications

Identifying structured data, with attribute names and values from the domain of healthcare is a challenging task due to the difficulty in obtaining the source data and its sensitivity. Once the data has been obtained, several operational challenges in processing the text and extracting a structured representation from it, in a timely manner is a nontrivial task.

The results presented in this manuscript indicate two important implications. Firstly, while the pre-trained textual similarity models are able to identify similarity between various, cross-domain texts, it is pertinent to apply domain adaptation before utilizing these models as a part of a domain-specific, solution. In particular, the evaluation of sentence similarity task on the complete STSb dataset, the pre-trained all-mpnet-base-v2 model and the pre-trained DistilBERT-base-uncased model show excellent agreement, in terms of the achieved Pearson Correlation. On the other hand, for the same dataset, and using the fine-tuned DistilBERT-base-uncased model, which was subsequently used in this study, indicates mediocre results. It is also important to note here that the STSb dataset has been collected from news items, image captions, and forum discussions, which produces text from various different domains. The performance of a domain-specific model is bound to be reduced on this dataset, especially when compared to the models trained on a large

Secondly, the classification threshold and semantic similarity score between the instances of MASS and the test sequences is of great importance. The threshold used by the proposed method is at "0.87", while the one used by the baseline method is "0.49". These values have been calculated using a dedicated portion of the dataset, with AuROC determining the optimal threshold value for semantic similarity. In order to test the relevance of a lower threshold value for the semantic similarity, we evaluated the performance of our proposed method with a decreased threshold value of "0.6", which produces an accuracy of 57.74% and F1-Score of 55.43% (both higher than the proposed approach and the baseline one). This increase in performance is due to the fact that with a lower threshold, a larger number of test sequences will be classified, and cause an increase in the number of actually true instances being found. It will also cause an increase in the number of incorrect results being found, thereby causing an imbalance between the precision and recall evaluations. In the real world, the test sequences are un-labeled and have to be manually verified by a clinician, before they can because a permanent part of a patient's medical record. Thus a balance has to be established between the accuracy and the number of classified instances. Essentially, this is what our proposed methodology achieves over the baseline methodology. With just 837 instances our proposed method achieves an accuracy of 56.21%. However, when the threshold is reduced from "0.87" to "0.6" the accuracy increases by less than 2% and an additional 622 instances are classified, which would nearly double the verification time for the physician. In scenarios, where it is important to capture a wide variety of correct instances and the focus is on improving the performance in terms of recall, a multi-modal approach, such as the one presented in Hussain et al. (2021) can be used.

These results can be further improved by increasing the volume of MASS and adding more medically aligned sequences and attribute labels into it, adding more and better defined extraction functions, and improving the text segmentation.

7.2. Value extraction via patterns vs UMLS

As an example, consider the sequence, "The patient has acute fever". Here, when we split this sequence into unigrams and use each non-stop word token to check the UMLS browser, we find that "Patient" has a semantic type of "Patient or Disabled Group", "acute" has "Temporal Concept ", "fever" has "Sign or Symptom" and "Finding", as determined by UMLS. With a bigram lookup, "acute fever" has a semantic type of "Disease or Syndrome" and "Finding", and "patient acute" has an approximate match with "Finding". These are only some of the concept types associated with the unigram and bigram tokens, provided by UMLS. However, "Finding" is the semantic concept in the label of the trained sequence, matching with this test sequence, we can easily make remote calls to the UMLS API and identify the best matching values, for "Finding" in it.

[13] https://www.sbert.net/_static/html/models_en_sentence_embeddings.html

On the other hand, the sentence “I am 8 years old”, is better identified through the use of regular expressions. This is because the semantic concepts associated with the unigrams “8”, “years”, “old”, and bigrams, “8 years”, “years old”, are not able to identify the value for the attribute “age”.

7.3. Clinical perspective on formalizing the encounters

The clinical adage that about two-thirds of diagnoses can be made on the basis of history alone has retained its validity despite the technological advances of the modern hospital. Once a rapport is built between a physician and a patient it helps boost the self-esteem of ill patients who are already struggling with their illnesses. The correct guidance by the physician is always relieving for the patients but for that, the art of interviewing a patient should be mastered (Walker, Hall, & Hurst, 1990). Objective questioning is a helpful tool in guiding patients and reaching the right diagnosis. It is also very helpful to cut-down unnecessary investigations which are a waste of time and money for the patients. While open-ended questions give the patients and their attendants the opportunity to explain the symptoms in detail, they often lead to cognitive overload and necessitate continuous note-taking and recording so as not to miss any important detail. Instead, for the physician, it is better to utilize short, targeted questions, and for the patient to provide detailed answers, so that contextual information can be collected (Chen, Guo, Wu, & Ju, 2020). Additionally, by recording these conversations, and extracting a correct summary from them, a lot of time and money can be saved for the medical center, physician, and patient.

7.4. Limitations of the current study

One of the limitations of this study pertains to the large amount of manual work required to transcribe and translate the clinical conversations into text, prepare the MASS instances, and expert-driven validation. The clinical environment in the real-(developing)-world is already under extreme stress, due to availability of resources and skewed patient load, towards the public infrastructure. In these circumstances, it is not possible to expect the clinical staff to provide the transcription and translation services for the clinical conversations. In fact, in designing this study, we have ensured that the transcription and translation of the clinical conversations remain closer to their real audios, rather than using any written medical reports, which are almost non-existent in our target environment. Another limitation of this study is the relatively small amount of data used for its evaluation. The manual processed involved in preparing the data, make it difficult to acquire large quantities of appropriate data. Additionally, there is a general derth of clinical conversations in literature (especially in English and Urdu), making it difficult to benchmark the proposed methodology.

In this pilot study we have been able to only partially automate the task of converting clinical conversations into a structured format. While some of the remaining portions of this task, can be automated for the general case (such as transcribing English language audios), the clinical domain requires careful, expert driven interventions and validations to ensure that the outcome of each individual task is safe and accurate. Through this pilot study and its associated data, we hope that more inclusive solutions, which can operate on a wide variety of real-world data, can be designed and implemented to eventually achieve ubiquitous healthcare.

8. Conclusion

In this study, we have proposed and evaluated a methodology to extract important data from clinical conversations. Using transformer-based machine learning models, medical dictionaries, and a novel pipeline, we apply various transformation and matching functions, to eventually extract a summarized and structured representation of the interaction between a physician and a patient. Our proposed methodology achieves an accuracy of 56.21% for identifying the attribute–value pairs correctly, which is better than the baseline methodology. In future, we aim to augment this methodology with automated methods for transcribing and translating the data. It is also important to provide enhanced NLP models for languages used in the developing world to enable them to grow at a faster pace.

CRediT authorship contribution statement

Fahad Ahmed Satti: Conceptualization, Methodology, Investigation, Software, Writing – original draft. **Musarrat Hussain:** Methodology, Writing – review & editing, Visualization. **Syed Imran Ali:** Validation, Writing – review & editing. **Misha Saleem:** Validation, Investigation, Data curation, Writing – review & editing. **Husnain Ali:** Supervision, Resources, Writing – review & editing. **Tae Choong Chung:** Supervision. **Sungyoung Lee:** Supervision, Resources, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data used in this study is available at https://github.com/desertzebra/ClinicalConversations.

Acknowledgments

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) supervised by the IITP (Institute for Information & communications Technology Promotion), and No. 2017-0-00655 (Lean UX core technology and platform for any digital artifacts UX evaluation, South Korea) and IITP-2020-0-01489 (2020 Grand ICT Research Center) and IITP-2021-0-00979 (AI based ESS scheduling and anomaly detection, South Korea) and 2022-0-00078, (Explainable Logical Reasoning for Medical Knowledge Generation, South Korea) by Institute for Information & communications Technology Promotion (IITP). We are thankful to Dr. Nisar Malik (MBBS, MD. pediatrics, Civil Medical Officer, District Headquarters Hospital, Kotli, Azad Jammu and Kashmir, Pakistan) for his valuable cooperation towards data collection from patients/guardians.

Appendix A. Data transformation

Note: While the data shown in section is from a real world case, several portions of it have been retracted to anonymize it. As such, any resemblance to a person, or a medical condition, would be a coincidental only. This should also not be used to provide any clinical diagnosis.

The following shows content of a short conversation in Urdu, written using English Alphabets by a Human.

Doctor: Ye study hai jisme baatien hongi mere aur apke beech mai bachay ki bemari k mutaliq wo record krni hai , theek hai na? Patient: Theek hai. Doctor: Kia naam hai bachi ka? Patient: ******. Doctor: Umar kitni hai? Patient: * maah. Doctor: * maah! kia masla hai? Patient: Bukhar hai. aur jhatke lage hain. Doctor: Bukhar hai aur jhatke lage hain, kis tarah k jhatke thay. Patient: Jis tarah tez saans leta hai banda aur khansi bhi hai. Doctor: Tez saans leta hai aur khansi bhi hai? Theek hai. aur bukhar tez hota hai? Patient: Je kal se tez hai. Doctor: Doodh pee rahi hai ya nahi pee rahi? Patient: Subha 10 baje piya hai uske bad nahi. Doctor: Eik min dikhaye, Saans bhi tez hai iska. Patient: ****** se check karaya hai isko waha bhaap bhi lagwai hai. Doctor: Han isko hai thora sa masla to isko dakhil kr raha hu mai, Theek hai?

The translated form of this conversation, into English is shown below. As can be seen, this text contains some spelling mistakes, grammatical mistakes, and typographical errors.

Doctor: There's a study for which i will have to record the conversation between us regardignt the child's health, is it okay with you? Patient: Yes! Doctor: What is her name? Patient: ******. Doctor: How old is she? Patient: * months. Doctor: * months.! And what is the problem? Patient: She has temperature along with seizures. Doctor: Okay! what kind of seizures? Patient: Rapid breathing along with coughing fit. Doctor: a coughing fit along with rapid breaths and was the temerature high? Patient: Yes, its high simce yesterday. Doctor: Is she taking any feed or not? Patient: Yes she did take at 10 in the morning. Doctor: Okay, let me have a look at her, yes her respiratory rate is high. Patient: We took her to a doctor in ******, they nebulized her. Doctor: Okay, she's not fine so I'm addmitting her here, will that be okay?

The text used for pre-processing is as follows.

Doctor: What is her name? Patient: ******. Doctor: How old is she? Patient: * months. Doctor: * months.! And what is the problem? Patient: She has temperature along with seizures. Doctor: Okay! what kind of seizures? Patient: Rapid breathing along with coughing fit. Doctor: a coughing fit along with rapid breaths and was the temerature high? Patient: Yes, its high simce yesterday. Doctor: Is she taking any feed or not? Patient: Yes she did take at 10 in the morning. Doctor: Okay, let me have a look at her, yes her respiratory rate is high. Patient: We took her to a doctor in ******, they nebulized her. Doctor: Okay, she's not fine so I'm addmitting her here, will that be okay?

The list of sequences generated after pre-processing is as follows.

What is her name? ******
How old is she? * months
And what is the problem? She has temperature along with seizures
what kind of seizures? Rapid breathing along with coughing fit
a coughing fit along with rapid breaths
was the temerature high?
its high simce yesterday
a coughing fit along with rapid breaths and was the temerature high? Yes, its high simce yesterday
Is she taking any feed or not? Yes she did take at 10 in the morning
let me have a look at her
yes her respiratory rate is high
Okay, let me have a look at her, yes her respiratory rate is high
We took her to a doctor in ******
they nebulized her
We took her to a doctor in ******, they nebulized her
she's not fine so I'm addmitting her here
will that be okay?
Okay, she's not fine so I'm addmitting her here, will that be okay?

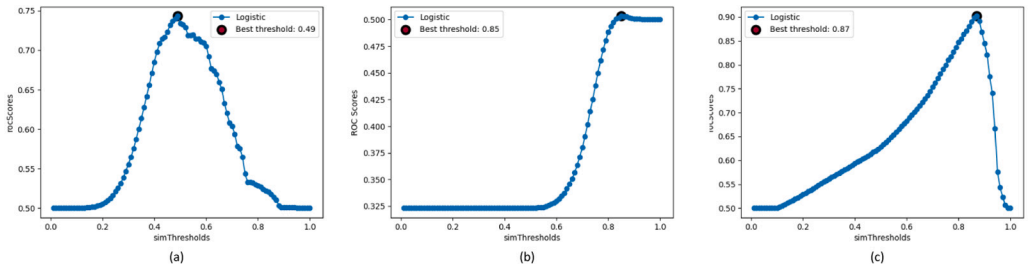

Fig. 8. Plot between AuROC and threshold values between 0.0 and 1.0. Where (a) shows the optimal threshold for all-mpnet-base-v2, (b) shows the optimal threshold for pre-trained DistilBERT-base-uncased model, and (c) shows the optimal threshold for fine-tuned DistilBERT-base-uncased model.

Appendix B. Threshold selection for sequence classification

For threshold selection, each instance contains the sequence text d, its embedding vector $\vec{V}$, and the associated label l. While the keys for this label are the same as in MASS, they also additionally contain a correct value, for each key. This set H is represented as shown in Eq. (4).

$$h = \left\langle d, \vec{V}, l \right\rangle | h \in H \wedge encode(d) \rightarrow \vec{V} \tag{4}$$

In the threshold selection process, we processed each instance in H by calculating the cosine similarity (shown in Eq. (5)) between its "embedding vectors" ($\vec{V}^H$) and the "embedding vector" ($\vec{V}^E$) of all instances in E. We then dropped all matches with similarity scores under 0.1, to reduce the number of comparisons in the next stage. In this way, we obtain a pair (ρ) of enriched sequences and their similarity score, represented in Eq. (6).

$$sim = \frac{\vec{V}^H \cdot \vec{V}^E}{\sqrt{\vec{V}^H \cdot \vec{V}^H}\sqrt{\vec{V}^E \cdot \vec{V}^E}} \tag{5}$$

$$\rho = \left\langle e_i, h_j, sim \right\rangle | e_i \in E \wedge h_j \in H \wedge 0.1 \leq sim \leq 1.0 \tag{6}$$

We then compare the labels of each pair, to validate the similarity ρ in terms of the keys and values obtained from e_i and h_j. The total function, representing the computed match between the keys and values of the labels from the threshold set and its corresponding match with the MASS instance is shown in Eq. (7). This process assigns one of three values to ρ, including "0", " ", and "1". If the two keys from any of the labels in the pair ρ are not equal, a value of 0 is assigned to it. On the other hand, if the two labels are equal, but the value annotated with the threshold selection set ($l^H.value$) and the value extracted from the application of regular expression or through the use of UMLS, as identified by the corresponding instance from MASS (x^E) on the text sequence from threshold selection set (d^H) are not equal, "∼" is assigned to ρ. Otherwise, if the labels and the value extracted match the annotated value, "1" is assigned to ρ. The "∼" matches were then manually verified and updated to either "0" or "1".

For all ρ, if $\chi(\rho)$ is zero, this indicates that while there is some cosine similarity (> 0.0) between the sequence in MASS and in the threshold selection set, their label keys and the expected values do not match with what can be achieved by the current matched instance. The value for $\chi(\rho)$ defines the computed actual class label ("0", "∼", "1"), which is converted into verified actual class label by expert intervention ("0" or "1"). This final value is used as the actual class score while the semantic similarity score provided by the fine-tuned DistilBert model, is used to calculate the predicted class label.

$$\chi(\rho) = \left\{ \begin{array}{ll} 1 & if \ \left(l_i^E \cdot key = l_i^H \cdot key \ \wedge \ x_i^E(d^H) \in l_i^E \cdot key\right) \\ \sim & if \ \left(l_i^E \cdot key = l_i^H \cdot key\right) \\ 0 & otherwise \end{array} \right\} \tag{7}$$

In order to define the predicted class label as similar or dissimilar, we move a threshold iterator from 0.0 to 1.0, with a step of 0.01. At each iteration, if the value of sim inside ρ is below the threshold iterator, the predicted class is assigned as dissimilar, and if it is equal to or above the threshold iterator the predicted class becomes similar.

Thus for 100 iterations, we have one set of actual class labels and 100 sets of predicted class labels based on the value of H. At each step, we calculate the area under ROC (AuROC) which provides a numeric value representing the ratio between the True Positive rate and False Positive rate. The maximum value of AuROC across all iterations then provides the semantic similarity threshold (α) between the two expressions, whereby the pair is considered, actually similar. After plotting all the values, as shown in Fig. 8, the best AuROC is achieved at 0.87 for our custom model, at 0.85 for the pre-trained DistilBERT-base-uncased model and for the all-mpnet-base-v2 it is achieved at 0.49. This is the value of α, which is used for classifying a test instance as similar to one of the instances in MASS.

References

Abdullah, M. F., & Ahmad, K. (2013). The mapping process of unstructured data to structured data. In *2013 international conference on research and innovation in information systems* (pp. 151–155). IEEE.

Andrabi, S. A. B., & Wahid, A. (2022). Machine translation system using deep learning for English to Urdu. *Computational Intelligence and Neuroscience*, *2022*.

Aramaki, E., Miura, Y., Tonoike, M., Ohkuma, T., Masuichi, H., & Ohe, K. (2009). Text2table: Medical text summarization system based on named entity recognition and modality identification. In *Proceedings of the BioNLP 2009 workshop* (pp. 185–192).

Beks, H., King, O., Clapham, R., Alston, L., Glenister, K., McKinstry, C., et al. (2022). Community health programs delivered through information and communications technology in high-income countries: Scoping review. *Journal of Medical Internet Research*, *24*, Article e26515.

Bodenreider, O. (2004). The unified medical language system (UMLS): Integrating biomedical terminology. *Nucleic Acids Research*, *32*, D267–D270.

Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., & Specia, L. (2017). Semantic textual similarity-multilingual and cross-lingual focused evaluation. In *Proceedings of the 2017 SEMVAL international workshop on semantic evaluation, vol. 2017*. http://dx.doi.org/10.18653/V1/S17-2001.

Chandra, M., Kumar, K., Thakur, P., Chattopadhyaya, S., Alam, F., & Kumar, S. (2022). Digital technologies, healthcare and COVID-19: Insights from developing and emerging nations. *Health and Technology*, 1–22.

Chen, S., Guo, X., Wu, T., & Ju, X. (2020). Exploring the online doctor-patient interaction on patient satisfaction based on text mining and empirical analysis. *Information Processing & Management*, *57*, Article 102253.

Chew, R., Wenger, M., Guillory, J., Nonnemaker, J., Kim, A., et al. (2022). Identifying electronic nicotine delivery system brands and flavors on instagram: Natural language processing analysis. *Journal of Medical Internet Research*, *24*, Article e30257.

Chiticariu, L., Krishnamurthy, R., Li, Y., Reiss, F., & Vaithyanathan, S. (2010). Domain adaptation of rule-based annotators for named-entity recognition tasks. In *Proceedings of the 2010 conference on empirical methods in natural language processing* (pp. 1002–1012).

Chiu, J. P., & Nichols, E. (2016). Named entity recognition with bidirectional LSTM-CNNS. *Transactions of the Association for Computational Linguistics*, *4*, 357–370.

Chiu, C., Villena, F., Martin, K., Núñez, C., & Dunstan, J. (2022). Training and intrinsic evaluation of lightweight word embeddings for the clinical domain in Spanish. *Frontiers in Artificial Intelligence*, *5*.

Coppersmith, G. (2022). Digital life data in the clinical whitespace. *Current Directions in Psychological Science*, *31*, 34–40.

Daud, A., Khan, W., & Che, D. (2017). Urdu language processing: A survey. *Artificial Intelligence Review*, *47*, 279–311.

Du, M., Wang, W., Wang, S., & Xu, B. (2020). A unified framework for attribute extraction in electronic medical records. In *2020 3rd international conference on algorithms, computing and artificial intelligence* (pp. 1–7).

Dutta, B., & Hwang, H.-G. (2020). The adoption of electronic medical record by physicians: A prisma-compliant systematic review. *Medicine*, *99*.

Furtado, D., Gygax, A. F., Chan, C. A., & Bush, A. I. (2022). Time to forge ahead: The Internet of Things for healthcare. *Digital Communications and Networks*.

Gasparetto, A., Marcuzzo, M., Zangari, A., & Albarelli, A. (2022). A survey on text classification algorithms: From text to predictions. *Information*, *13*, http://dx.doi.org/10.3390/info13020083, URL: https://www.mdpi.com/2078-2489/13/2/83.

Hussain, M., Satti, F. A., Hussain, J., Ali, T., Ali, S. I., Bilal, H. S. M., et al. (2021). A practical approach towards causality mining in clinical text using active transfer learning. *Journal of Biomedical Informatics*, *123*, Article 103932.

Ismail, L., Materwala, H., Karduck, A. P., & Adem, A. (2020). Requirements of health data management systems for biomedical care and research: Scoping review. *Journal of Medical Internet Research*, *22*, Article e17508. http://dx.doi.org/10.2196/17508.

Kanwal, S., Malik, K., Shahzad, K., Aslam, F., & Nawaz, Z. (2019). Urdu named entity recognition: Corpus generation and deep learning applications. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, *19*, 1–13.

Kaplan, B. (2020). Revisiting health information technology ethical, legal, and social issues and evaluation: Telehealth/telemedicine and COVID-19. *International Journal of Medical Informatics*.

Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. In *International conference on learning representations*.

Masroor, H., Saeed, M., Feroz, M., Ahsan, K., & Islam, K. (2019). Transtech: Development of a novel translator for Roman Urdu to English. *Heliyon*, *5*, Article e01780.

Mehta, K., Oprea, I., & Rasiwasia, N. (2021). Latex-Numeric: Language agnostic text attribute extraction for numeric attributes. In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies: industry papers* (pp. 272–279).

Narayanan, A., Rao, A., Prasad, A., & Das, B. (2020). Character level neural architectures for boosting named entity recognition in code mixed tweets. In *2020 international conference on emerging trends in information technology and engineering* (pp. 1–6). IEEE.

Nasar, Z., Jaffry, S. W., & Malik, M. K. (2019). Textual keyword extraction and summarization: State-of-the-art. *Information Processing & Management*, *56*, Article 102088.

Putthividhya, D., & Hu, J. (2011). Bootstrapped named entity recognition for product attribute extraction. In *Proceedings of the 2011 conference on empirical methods in natural language processing* (pp. 1557–1567).

Roy, K., Goyal, P., & Pandey, M. (2021). Attribute value generation from product title using language models. In *Proceedings of the 4th workshop on e-commerce and NLP* (pp. 13–17).

Satti, F. A., Ali, T., Hussain, J., Khan, W. A., Khattak, A. M., & Lee, S. (2020). Ubiquitous health profile (UHPR): A big data curation platform for supporting health data interoperability. *Computing*, *102*, 2409–2444.

Satti, F. A., Hussain, M., Hussain, J., Ali, S. I., Ali, T., Bilal, H. S. M., et al. (2021). Unsupervised semantic mapping for healthcare data storage schema. *IEEE Access*, *9*, 107267–107278. http://dx.doi.org/10.1109/ACCESS.2021.3100686.

Si, Y., Wang, J., Xu, H., & Roberts, K. (2019). Enhancing clinical concept extraction with contextual embeddings. *Journal of the American Medical Informatics Association*, *26*, 1297–1304.

Sun, C., Yang, Z., Wang, L., Zhang, Y., Lin, H., & Wang, J. (2021). Biomedical named entity recognition using BERT in the machine reading comprehension framework. *Journal of Biomedical Informatics*, *118*, Article 103799.

Vandic, D., Van Dam, J.-W., & Frasincar, F. (2012). Faceted product search powered by the semantic web. *Decision Support Systems*, *53*, 425–437.

Walker, H. K., Hall, W. D., & Hurst, J. W. (1990). Clinical methods: The history, physical, and laboratory examinations.

Wang, Y., Liu, S., Afzal, N., Rastegar-Mojarad, M., Wang, L., Shen, F., et al. (2018). A comparison of word embeddings for the biomedical natural language processing. *Journal of Biomedical Informatics*, *87*, 12–20.

Wang, Q., Yang, L., Kanagal, B., Sanghai, S., Sivakumar, D., Shu, B., et al. (2020). Learning to extract attribute value from product via question answering: A multi-task approach. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 47–55).

World Health organization (2021). Global strategy on digital health 2020–2025. URL: https://apps.who.int/iris/bitstream/handle/10665/344249/9789240020924-eng.pdf.

Xu, H., Wang, W., Mao, X., Jiang, X., & Lan, M. (2019). Scaling up open tagging from tens to thousands: Comprehension empowered attribute value extraction from product title. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 5214–5223).

Yan, H., Gui, T., Dai, J., Guo, Q., Zhang, Z., & Qiu, X. (2021). A unified generative framework for various NER subtasks. (pp. 5808–5822). URL: https://doi.org/10.18653/v1/2021.acl-long.451.

Yan, J., Zalmout, N., Liang, Y., Grant, C., Ren, X., & Dong, X. L. (2021). Adatag: Multi-attribute value extraction from product profiles with adaptive decoding. In *ACL/IJCNLP, no. 1*.

Yang, L., Wang, Q., Yu, Z., Kulkarni, A., Sanghai, S., Shu, B., et al. (2022). MAVE: A product dataset for multi-source attribute value extraction. In *Proceedings of the fifteenth ACM international conference on web search and data mining* (pp. 1256–1265).

young Jung, S., Lee, K., & Hwang, H. (2022). Recent trends of healthcare information and communication technologies in pediatrics: A systematic review. *Clinical and Experimental Pediatrics*, *65*, 291.

Zeng, D., Sun, C., Lin, L., & Liu, B. (2017). LSTM-CRF for drug-named entity recognition. *Entropy*, *19*(283).

Zheng, G., Mukherjee, S., Dong, X. L., & Li, F. (2018). Opentag: Open attribute value extraction from product profiles. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 1049–1058).