Communications in Mathematics and Applications
Vol. 15, No. 2, pp. 929–945, 2024
ISSN 0975-8607 (online); 0976-5905 (print)
Published by RGN Publications
DOI: 10.26713/cma.v15i2.2939

http://www.rgnpublications.com

Research Article

The ScispaCy Clinical Term Extraction and Snomed CT Synonymy Elimination From Clinical Data For Clustering: A Novel Study

E.K. Jasila*, N. Saleena and K. A. Abdul Nazeer

Department of Computer Science and Engineering, National Institute of Technology Calicut, Kozhikode, Kerala, India
***Corresponding author:** jasilaabhilash@gmail.com, jasila_p170058cs@nitc.ac.in

Received: May 2, 2024 **Accepted:** August 29, 2024

Abstract. A clinical document is a written or electronic record that encompasses details regarding a patient's medical procedure, clinical trial, or test outcomes. Standard information mining approaches have challenges in clustering clinical documents due to their unstructured nature. This work introduced a new approach for grouping clinical documents to address problems related to synonymy, abbreviation extension, and extraction of key features. The clinical document collection for coronary artery disease consists of 1304 records obtained from 296 patients. These records have been chosen for preprocessing with the aim of removing any irregularities. The scispaCy model extracts relevant information after a simple letter-matching algorithm identifies and extends abbreviations. Furthermore, the features are examined using SNOMED CT ontology to eradicate medical terms that have similar meanings. The TF-IDF method is employed to convert the recovered features into vectors. The BERT model's word embeddings were employed in this study to represent features. Nevertheless, the TF-IDF model surpasses the BERT model in performance. The clustering process utilises an enhanced k-means algorithm that incorporates the Red-Black Tree data structure. The recommended strategy was evaluated with several existing clustering algorithms in this study. It has been found that the proposed method produces clusters with higher scores for Normalised Mutual Information (NMI) and accuracy. Based on the results of this investigation, the model has the ability to detect individuals with similar diseases and provide assistance to healthcare professionals.

Keywords. Clinical document, SNOMED CT ontology, Red-black tree

Mathematics Subject Classification (2020). 92C50, 62P10

1. Introduction

Electronic Medical Record (EMR) has gained much popularity with the advancement of *Hospital Information System* (HIS) and information technology. Electronic Health Record (EHR) or EMR is used to store charts, symbols, graphics, texts, and other digital information produced by HIS. As a result, medical-related records could be efficiently transmitted, stored, reproduced, and managed. With the rapid adoption of EMR, various sources of clinical data (such as diagnostic history, demographics, medications, vital signs, and laboratory test results) are becoming widely accessible, facilitating large-scale data analysis. In the EMR, data is classified into three types: structured, along with semi-structured and unstructured data (Roberts *et al*. [22]). Structured data is typically kept in fixed-mode databases, contains fundamental details like personal information, medications taken, allergies, and health related data such as weight, height, blood pressure, etc. Semi-structured data typically has a flowchart format with name, value, and time stamp, just like RDF (Resource Description Framework). Unstructured text is a kind of narrative data which include surgical records, clinical notes, pathology reports, discharge records, and radiology reports. These unstructured texts contain a vast amount of medical information, but it lack standard structural frameworks and are riddled with mistakes like improper grammar use, spelling mistakes, regional dialects, and semantic ambiguities that make data processing and analysis more difficult.

The recent research on EHR data using deep learning techniques (Shickel *et al*. [26]) demonstrates the importance of health record data in hospitals and ambulance care settings. The datasets used in the various clinical approaches are *Medical Information Mart for Intensive Care* (MIMIC) (Johnson *et al*. [13]), and clinical notes data from *Informatics for Integrating Biology and the Bedside* (i2b2)[1]. The data mining techniques for EHR and its applications are proposed by Renganathan [21], and Yadav *et al*. [29]. The various information extraction techniques for EHR (Meystre *et al*. [18]) are based on pattern matching, ontology, machine learning techniques etc. The preprocessing along with its vector representations uses different word embedding models like Word2Vec, FastText, BERT, and clinical BERT (Khattak *et al*. [14]). In the clinical domain, classification and prediction tasks using word embedding for text representation achieve better results than traditional feature extraction methods.

Clustering is a technique of grouping data objects with low inter-class similarity and high intra-class similarity (Tan *et al*. [27]). In another way, 'similarity' is a metric that measures how closely data are linked. Translating data into a similarity space and then doing an analysis of the data relationships can be used as an approach for evaluating similarity (Doing-Harris *et al*. [6], Patterson *et al*. [20]). Previous studies (Patterson and Hurdle [20], Zhang *et al*. [30], and Cohen *et al*. [2]) show that before deploying an information extraction system, it is critical to assess the text's structure and overall similarity of material within the corpus. When the order of content changes, current methodologies for determining similarity within a corpus utilizing global alignment algorithms do not seem to work out. A sliding window global alignment methodology can be used to mitigate some of the speed loss, but this method is inefficient. Zhang *et al*. [31] discovered duplication and tested their method with 178 complete patient records

[1]i2b2 NLP Challenges and Datasets, https://www.i2b2.org/NLP/DataSets/, 2011, Online: accessed 19-July-2021.

of remarkably similar patients using the sliding window global alignment methodology. It is relevant to explore an effective, efficient, and scalable approach for recognizing similar clinical texts within a large corpus.

The stages in a traditional method for clinical document clustering are as follows: (1) Preprocessing the clinical notes, (2) Feature vector representation, and (3) Clustering the documents. In the voluminous collection of datasets, the traditional feature extraction methods extract many terms for representing the documents. In this paper, (i) the abbreviations are detected and replaced by a letter matching algorithm, (ii) the preprocessing part works with a scispaCy model for processing the text, and (iii) SNOMED CT ontology is used to eliminate synonymous medical terms. These are represented in vector form by the TF-IDF model. The clustering section employs an enhanced k-means algorithm based on the Red-Black Tree. The experiments are done with the dataset from the i2b2 workshop[1]. Two data sets are used, one set with two clusters and another with three clusters. This paper uses two types of evaluation measures viz Accuracy and Normalized Mutual Information (NMI). Our method outperforms other similar methods available in the literature, in terms of NMI scores and Accuracy.

1.1 Key Highlights

This paper proposes a novel approach for the clustering of clinical documents with the following contributions:

(i) A novel method for extracting clinical terms using the scispaCy model and elimination of the synonymy problem using the SNOMED CT ontology.

(ii) An enhanced k-means algorithm based on Red Black Tree for improving the accuracy and NMI scores of clusters.

1.2 Organization of the paper

The subsequent sections of the paper are as follows: Section 2 has a summary of similar works. Our proposed strategy for clinical document clustering is discussed in Section 3. Then, Section 4 discusses the experimental results, after which Section 5 concludes the paper.

2. Related Work

Several researchers have proposed various frameworks for grouping medical documents. A method for clustering clinical documents based on multi-view nonnegative matrix factorization is presented in (Ling *et al*. [16]). Here, the features are combined with all three views, i.e., words, medicine names, and symptoms. This strategy performs better when compared to feature representation by a single view. A method for grouping clinical templates based on SNOMED CT ontology is presented in (Gøeg *et al*. [8]).

A method for finding features based on two new technologies, IBM Watson and Framester, is discussed in (Dessi *et al*. [5]). Here, the problem of the curse of dimensionality is resolved by reducing the features using singular value decomposition. A UMLS meta map-based biomedical document clustering methodology based on disease concepts is explained in (Shah and Luo [24]). According to El-Sappagh *et al*. [7], an SCT Ontology (upper-level) based on general medical

science can enhance the meaning of clinical documents. Marcińczuk *et al*. [17] compared the modern language modelling approaches (doc2vec and BERT) with traditional methods (TF-IDF and wordnet-based). The findings of the experiment demonstrated that wordnet-based similarity measures could outperform contemporary embedding-based methods.

An information retrieval system based on a corpus of COVID-19-related research publications is described by Das *et al*. [4]. They created a similarity network out of the papers, using shared citations and biological domain-specific language embeddings to determine similarity. On the article network, the articles are clustered using ego-splitting community detection, and then the queries are matched with the clusters. To offer responses to the questions, extractive summarization using BERT and PageRank approaches are applied. Si and Roberts[2] presented a network for learning from words to sentences, sentences to notes, and notes to patients that comprises three levels of Transformer-based encoders. Before the final patient representation is given as input into the classification layer for clinical predictions, the first level applies a pre-trained BERT model from word to phrase. The second and third layers both employ a stack of 2-layer encoders.

A document embedding-based clustering algorithm (Tang *et al*. [28]) is suggested for generating clinical note templates. It uses a modified k-means algorithm. This method performs better in terms of precision and executes in lesser time complexity.

3. Methodology

The different stages of the proposed framework are Document Preprocessing, Document representation, and Clustering. Figure 1 depicts the proposed framework's overall architecture.

3.1 Document Preprocessing

The document preprocessing strategy makes the documents appropriate for employing data mining functionalities. We require complete data to process documents correctly using the following techniques: preprocessing, abbreviation detection and replacement, named entity recognition, and SNOMED CT ontology inclusion. We only get the improved quality features for representing documents after employing these techniques in order. Thus, we do not use typical preprocessing techniques.

(1) *Preprocessing*: Initially, XML tags are removed from the clinical notes.

(2) *Abbreviation detection and replacement*: We used the method reported in (Schwartz *et al*. [22]) for identifying abbreviations in clinical documents. The abbreviations are detected, and they are replaced with corresponding expansions. This step is to obtain more meaningful clinical terms while avoiding the differences in abbreviations and their expansions. For example, in the clinical notes, the pattern appears as "Coronary Artery Disease (CAD)... CAD". The abbreviation CAD is replaced by its expansion Coronary Artery Disease which appears elsewhere in the text.

[2]Y. Si and K. Roberts, Three-level hierarchical transformer networks for long-sequence and multiple clinical documents classification, arXiv:2104.08444 (2021), DOI: 10.48550/arXiv.2104.08444.

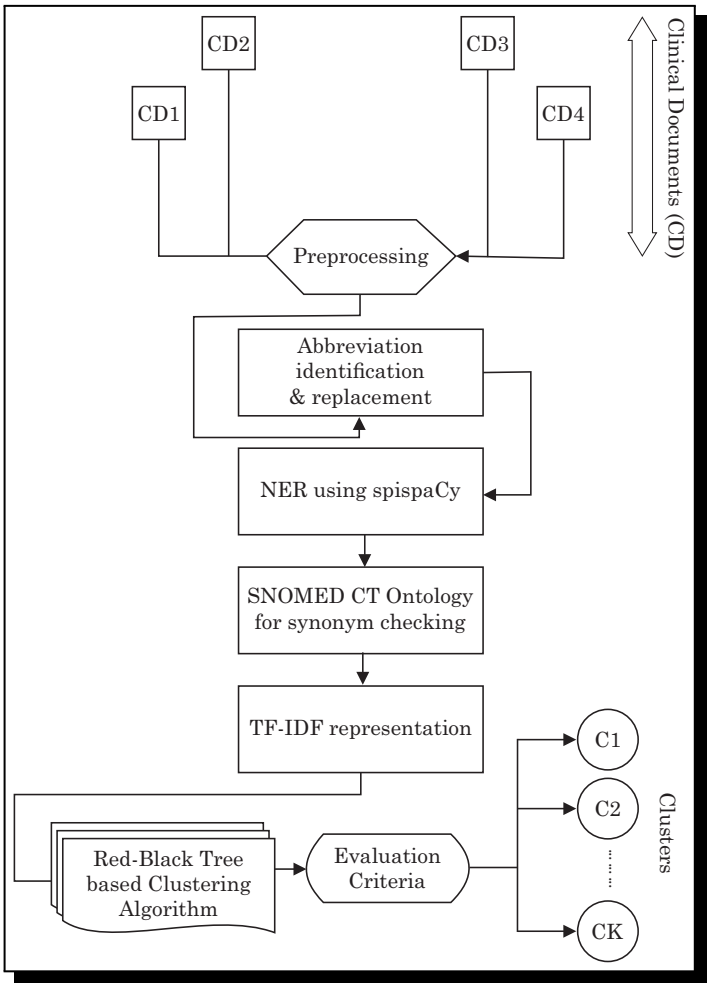

Figure 1. Flowchart of the proposed method for clustering of clinical documents

Abbreviation detection methods can be classified into three groups: heuristic, statistical, and machine learning. Heuristic rules compare the short and long forms lexically to see if an expansion form is correct. The proposed method uses a simple letter-matching algorithm that contains the following steps (Schwartz and Hearst [23]):

(a) Each character in the short form must correspond to a character in the long form.

(b) The first character of the short form must correspond to the first word of the long form.

(c) The long form's matched characters must be in the same order as the short form's characters.

(3) *Named Entity Recognition* (NER): NER stage extracts the pertinent clinical features from within the documents utilising the scispaCy core model (Shelar *et al*. [25]). This set of spaCy models includes biological, scientific, and clinical text processing. ScispaCy is a valuable tool for NER, the process of recognizing and categorizing items. The spaCy library contains many useful text-processing features in a variety of languages. SpaCy models have become the de facto standard for practical *Natural Language Processing* (NLP) owing to their speed. We chose to construct a biomedical text processing pipeline on top of the spaCy library since potential clients are familiar with the spaCy models.

The proposed method uses *en_core_sci_lg* model for extracting the relevant clinical named entities. ScispaCy's core packages are *en_core_sci_sm*, *en_core_sci_md*, and *en_core_sci_lg*. The *en_core_sci_lg* package has a larger vocabulary and includes 600k word vectors, whereas the *en_core_sci_sm* package has a smaller vocabulary and does not include word vectors. The *en_core_sci_md* package contains medium vocabulary with 50k word vectors.

(4) *SNOMED CT Ontology*: Systematized Nomenclature of Medicine-Clinical Terms (SNOMED CT) (Ivanović and Budimac [10]) is a global clinical nomenclature. It assists to provide cost-effective, high-quality healthcare by displaying clinical data in an effective and comprehensible manner. SNOMED CT is a crucial component of the EHR. It was designed and validated using clinical, technological, and terminological skills. Implementation and practical application are key steps in realizing its full potential.

Figure 2 shows the representation of a SNOMED CT. In SNOMED CT, descriptions, concepts, and relationships are the main component types. A distinct, numerical, and machine-readable SNOMED CT identifier is the point of reference for each concept, representing a unique clinical meaning. Each concept is represented by a pair of descriptions: Synonym and Fully Specified Name (FSN). The FSN stands for a distinct, clear explanation of a concept's meaning. There can only be a unique FSN per concept in each language.

A synonym is a term that can be employed to indicate or choose a concept. There may be many synonyms for a given concept. This enables SNOMED CT users to denote a specific clinical meaning using the terms they prefer. An association between two concepts is represented by a relationship. Individual patients and clinicians benefit from SNOMED CT-based clinical information, which supports evidence-based care. The use of an EHR enhances communication and expands access to pertinent data. The advantages are significantly increased if clinical data is stored in a manner that permits meaning-based retrieval.

Our proposed method uses the following procedure to check each extracted entity (terms) using SNOMED CT ontology. The steps are:

(1) If the entity is found in SNOMED CT

- Find the ConceptID = ConceptID(entity) using SNOMED CT
- Take all active descriptions in ConceptID (Different descriptions are available - Synonyms and Fully Specified Name)
- Select Fully Specified Name (FSN) from the previous step
- Replace entity by its FSN

(2) Else

- Keep the entity with no change

3.2 Document Representation

We compared two types of models for the document's feature representations: TF-IDF and BERT models. The TF-IDF model performs better when compared to BERT model.

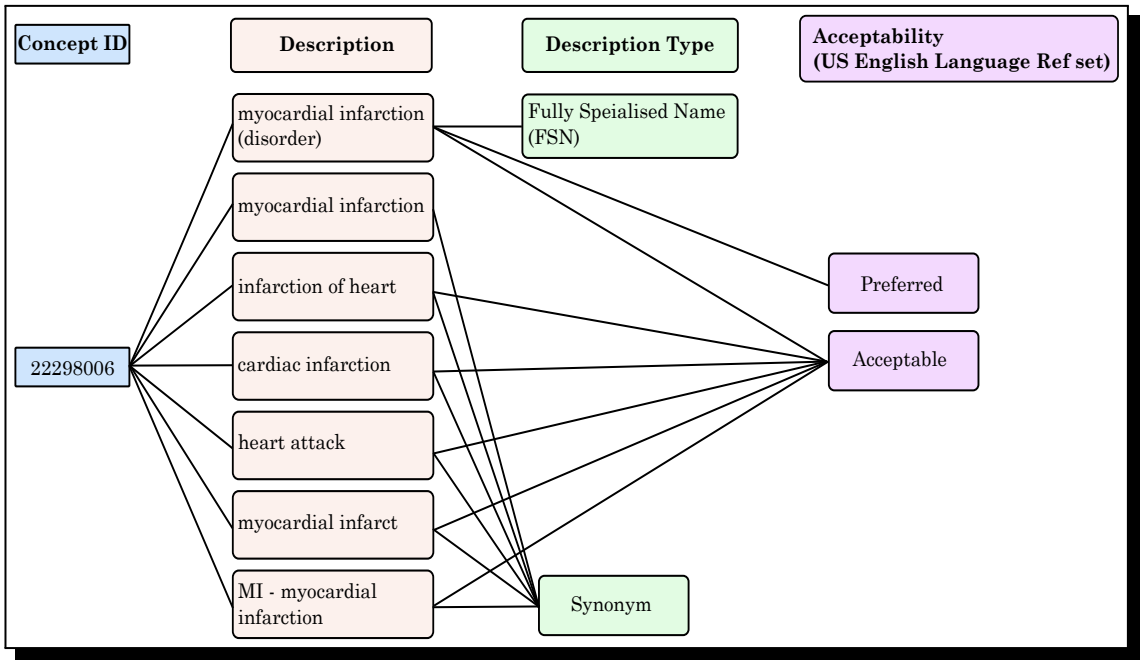

Figure 2. SNOMED CT representation SNOMED CT[2]

(1) *TF-IDF model*: The features are represented using the vector space model, TF-IDF. The TF-IDF value is calculated by eqn. (3.1). Eqn. (3.2) is used to find the IDF value, and the calculated TF-IDF values are normalized by the Euclidean Norm, eqn. (3.3). TF(t_1,d_1) indicates how many times a term t_1 occurs in document d_1. IDF(t_1) is calculated as the total document's logarithmic fraction divided by the number of documents containing the term t_1. The IDF computation is added with "1" to prevent the result from being zero.

$$\text{TF}-\text{IDF}(t_1,d_1) = tf(t_1,d_1) * \text{IDF}(t_1), \tag{3.1}$$

$$\text{IDF}(t_1) = \log\left(\frac{nd}{df(dt_1)}\right), \tag{3.2}$$

$$u_n = \frac{u}{\|u\|_2} = \frac{u}{\sqrt{u_1^2+u_2^2+\cdots+u_n^2}}. \tag{3.3}$$

(2) *BERT model*: Bidirectional Encoder Representations from Transformers (BERT) (Kattak *et al*. [14]) is based on a transformer technique, i.e., an attention mechanism. The BERT model learns context-based relationships among the word in the text. A transformer is composed of two mechanisms: an encoder that takes the text input along with a decoder that makes a task prediction. BERT only needs an encoder mechanism to achieve its goal of generating a language model. Here we used BERT base model that contains transformer blocks (12 layers), 110 million parameters, along with 12 attention heads, along with an output size of 768 dimensions. The BERT model which considers the polysemy and synonymy relations among the text documents. This model accepts only maximum of 512 tokens from each documents and the final output has each document with 768 dimensions. The output is taken from the last hidden state of the BERT model. Figure 3 shows the architecture of the BERT model.

[2]SNOMED CT Document Library, International Health Terminology Standards Development Organisation, http://snomed.org/doc, 2022, Online: accessed 19 July, 2022.

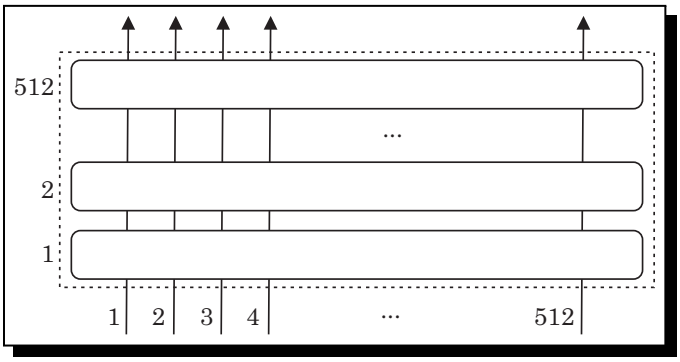

Figure 3. BERT model [15]

3.3 Clustering Algorithm using Red Black Tree

The clustering method has two phases. The first phase differs slightly from the algorithm used in (Jasila *et al*. [11]). It uses a sort-based tactic to find the initial centroids. During the second phase, data items are allocated to the suitable clusters. using an improved method that is the same as in (Jasila *et al*. [11]).

A self-balancing binary search tree containing an extra bit in each node, which is interpreted as red or black, is known as a Red-Black tree (Cormen *et al*. [3]). We enhance the standard k-means clustering algorithm by using a Red-Black tree-based approach (Jasila *et al*. [12]). The method employs a systematic procedure for determining initial centroids, in contrast to the random selection of initial centroids used in the k-means algorithm. Algorithm 2 illustrates the method of finding initial centroids. This algorithm slightly differs from the algorithm described in our previous work. Normally two kinds of datasets are available: one has evenly distributed data, and the other has unevenly distributed data. The datasets are dispersed unevenly, which means some clusters contain a large amount of similar data, while others contain less. In this instance, x has a value greater than 1. The rearranged data items are divided into k sets, with $x * n/k$ data items in each of the $k-1$ sets. The kth set, or the final set, has $n-(x*n/k)*(k-1)$ data items. This paper uses the clinical document dataset, which are unevenly distributed. For unevenly distributed datasets, x has a value greater than 1. Algorithm 1 provided below illustrates the clustering method.

Algorithm 1: Clustering Approach

Input: A set of data items $Q = q_1, q_2, \ldots, q_n$, and k is the number of clusters

Output: k Clusters

Steps:

1. Get Initial set of centroids from Algorithm 2.
2. Assign each data item q_i, $i \in n$ to appropriate clusters by Algorithm 4.

Algorithm 3 and Algorithm 4 are same as that of the algorithms described in (Jasila *et al*. [11]. Algorithm 3 explains the procedure for sorting the data (obtained from Algorithm 2 - *Step* 4) based on Red Black Tree. The data points are assigned to the appropriate clusters by Algorithm 4.

When comparing two documents, the cosine similarity (Jasila *et al*. [11]) measure is the most suitable similarity measure. In Algorithm 4, our method employs cosine similarity to assess the similarity between the documents. The similarity between two vectors E and F, $E = (e_1, e_2, \ldots, e_n)$ and $F = (f_1, f_2, \ldots, f_n)$ is determined by the eqn. (3.4). The cosine similarity values range from $(0, 1)$, where "1" denotes the vectors are the same and "0" denotes they are different.

$$\left(\sum_{i=1}^{n} E_i * F_i\right) \Big/ \sqrt{\left(\sum_{i=1}^{n} E_i^2\right) * \left(\sum_{i=1}^{n} F_i^2\right)}. \tag{3.4}$$

Algorithm 2: Finding Initial centroids

Input: $Q = q_1, q_2, \ldots, q_n$ // k is the Number of clusters

Output: Initial centroids

Steps:

1. Begin
2. For each column of the dataset Q
3. Calculate Range, max_value−min_value
4. Select the column with greatest Range
5. Sort the data Specified in the step 4, using Algorithm 3
6. For each value s_i, $i \in n$, sorted data S from Algorithm 3
7. Retrieve the corresponding data items from the dataset Q
8. Add the data points to set R
9. // Use a parameter x in the initial centroids calculation
10. Set k_1 to $x * n/k$
11. // x's value depends on the dataset Q, For unevenly distributed data $x > 1$
12. // Other case $0.9 \leq x \leq 1$
13. Split the rearranged dataset R to k parts
14. If $x > 1$
15. Add $k-1$ parts with k_1 data points and kth part contain $n - k_1 * (k-1)$ data items
16. Else
17. Add each k parts with k_1 data points
18. Set initial centroids to the mean of each of the k parts
19. End

Algorithm 3: Red Black Tree based Sorting

Input: $U = u_1, u_2, \ldots, u_n$, n values for the column having the greatest range

Output: S, Sorted data

Steps:

1. Begin
2. Set s_1 to root and set color to Red
3. For each data u_j, $j = 2$ to n
4. If $u_j <$ ROOT
5. Insert the value recursively into the root's left subtree.
   // Balance the tree using Recoloring and Rotation.
6. Else
7. Insert the value recursively into the root's right subtree.
   // Balance the tree using Recoloring and Rotation.
8. End For
9. Return the value according to the tree's inorder traversal order.
10. End

Algorithm 4: Assignment of data items to appropriate clusters

Input: $Q = q_1, q_2, \ldots, q_n$ (n data points)

k (Number of clusters)

$IC = ic_1, ic_2, \ldots, ic_k$, k initial centroids (Algorithm 2)

Output: k Clusters

Steps:

1. Begin
2. Calculate the cosine_similarity, $\cos(q_i, ic_j)$ between q_i, $i = 1$ to n and ic_j, $j = 1$ to k
3. For each data point p_i, $i = 1$ to n
4. Locate the adjacent centroid
5. Assign q_i to its adjacent centroid ic_j
6. Set cluster number to j
7. Set similarity to $\cos(q_i, ic_j)$
8. For all clusters j, $1 \le j \le k$, recompute the centroids
9. Repeat
10. For each data item q_i, $i = 1$ to n
11. Calculate the cosine_similarity between each data item and its nearest centroids
12. If the calculated similarity > the similarity stored
13. Keep the data item in the cluster
14. Refresh the similarity value to the newly calculated value
15. Else
16. Calculate the cosine_similarity of the data item with every other centroid

Contd. Algorithm

17. Locate the adjacent centroid
18. Assign q_i to its adjacent centroid ic_j
19. Set cluster number to j
20. Set similarity to $\cos(q_i, ic_j)$
21. End for
22. For all clusters j, $1 \le j \le k$, recompute the centroids.
23. Until Stagnation of data items from one cluster to another
24. End

4. Experimental Analysis

4.1 Experimental Setup

Our proposed work was executed with the clinical notes dataset. The clinical notes datasets for *Coronary Artery Diseases* (CAD) were collected from i2b2 NLP data sets from Harvard Medical School[1]. This clinical notes 2014 dataset contain 1304 records. These records are from 296 patients, and each patient has 3 to 5 records. The dataset is an XML-tagged dataset. The records are divided into two groups. In one group, three types of cases are included, i.e., patients with CAD, without CAD, and with CAD in their first records. In the second group, two types of cases are included, i.e., patients with CAD and patients without CAD in their records. For our implementation, Python, Python IDE Spyder, scispaCy and PyMedTermino are used. The data related to SNOMED CT is available through a licence issued by the National Release Centre (NRC)[3].

4.2 Evaluation Metrics

We use two common metrics to assess clustering performance, viz. *Normalized Mutual Information* (NMI) (Huang *et al*. [9]) and Accuracy(Cai *et al*. [1]). NMI and Accuracy have values ranging from 0 to 1, with 1 being the best clustering result and 0 being the worst. These two measuring metrics are widely utilised in the clustering literature, and also each has benefits and drawbacks. But, combining them shows how effective the clustering methods are. In terms of known class labels, accuracy is the percentage of correctly classified data. The NMI is calculated using eqn. (4.1),

$$NMI(X,CS) = 2I(X;CS)/H(X) + H(CS). \tag{4.1}$$

Here, the terms X and CS stand for class labels and cluster labels, H stands for the Entropy of the class labels (eqn. (4.2)), $I(X;CS)$ stands for the Mutual Information between the class labels and cluster labels (eqn. (4.3)). Each cluster's class label entropy is $H(X \mid CS)$, as shown in eqn. (4.4).

$$H(X) = -\sum p(x) * \log p(x), \tag{4.2}$$

[3]SNOMED CT Ontology, http://www.nrces.in/standards/snomed-ct, 2022, Online: accessed 21 July, 2022.

$$I(X;CS) = H(X) - H(X \mid CS), \tag{4.3}$$

$$H(X \mid CS) = -\sum\sum p(x,cs) * \log p(x \mid cs). \tag{4.4}$$

4.3 Results and Discussions

We compared our results with various clustering algorithms like k-means, k-medoid, heuristic k-means (Nazeer *et al*. [19]), and *Nonnegative Matrix Factorization* (NMF). In comparison to other models, our suggested model performs better. We contrasted the two feature representation models, TF-IDF and BERT. In the case of the NMF algorithm, three views are compared: words, symptoms/medication names, and all three views together with the two feature methods count and TF-IDF. These NMF-based results are taken from a previously published literature (Ling *et al*. [16]). Table 1 depicts the overall analysis of various clustering algorithms concerning $k = 2$ over accuracy. The results in Table 1 show that the vector representation of documents for clustering works better in TF-IDF compared to the BERT word embedding model. So we also calculated the NMI scores with TF-IDF and BERT model representations. Table 2 shows a summary analysis of different clustering algorithms when $k = 2$ over NMI. These clinical notes datasets are unevenly distributed, some of the clusters have a lot of similar data, while others have less. In this case, the x's value is greater than 1. The x's value of the clinical notes dataset is 1.45 when $k = 2$. The accuracy and NMI scores comparison for $k = 3$ is shown in Tables 3 and 4. The value of x is 1.3. Tables 1, 2, 3, and 4 results show that the proposed method with TF-IDF model performs better when compared to the existing methods.

Figure 4 shows the graphical representation of the accuracy of the clinical notes dataset when $k = 2$ and $k = 3$.

Table 1. Accuracy of various methods at $k = 2$

Algorithms	Feature Type	Views	Accuracy (%)
NMF (Ling *et al*. [16])	Count	Words	57.77
		Symptom/Medication	55.07
		All 3 views	59.80
NMF (Ling *et al*. [16])	TF-IDF	Words	53.38
		Symptom/Medication	73.31
		All 3 views	75.00
k-means	TF-IDF	Clinical terms + SNOMED CT	69.17
	BERT	Words	56.36
k-medoid	TF-IDF	Clinical terms + SNOMED CT	69.17
	BERT	Words	54.06
Heuiristic k-means	TF-IDF	Clinical terms + SNOMED CT	69.17
	BERT	Words	43.63
Proposed Method	TF-IDF	Clinical terms + SNOMED CT	78.00
	BERT	Words	43.63

Table 2. NMI scores of various methods at $k = 2$

Algorithms	Feature Type	Views	Accuracy (%)
NMF (Ling *et al*. [16])	Count	Words Symptom/Medication All 3 views	0.0198 0.0924 0.1751
NMF (Ling *et al*. [16])	TF-IDF	Words Symptom/Medication All 3 views	0.0034 0.1844 0.2283
k-means	TF-IDF BERT	Clinical terms + SNOMED CT Words	0.0229 0.0197
k-medoid	TF-IDF BERT	Clinical terms + SNOMED CT Words	0.0088 0.0931
Heuiristic k-means	TF-IDF BERT	Clinical terms + SNOMED CT Words	0.0248 0.0720
Proposed Method	TF-IDF BERT	Clinical terms + SNOMED CT Words	0.2314 0.0720

Table 3. Accuracy of various methods at $k = 3$

Algorithms	Feature Type	Views	Accuracy (%)
NMF (Ling *et al*. [16])	Count	Words Symptom/Medication All 3 views	40.54 52.03 53.38
NMF (Ling *et al*. [16])	TF-IDF	Words Symptom/Medication All 3 views	35.47 52.36 52.36
k-means	TF-IDF BERT	Clinical terms + SNOMED CT Words	50.84 34.59
k-medoid	TF-IDF BERT	Clinical terms + SNOMED CT Words	26.92 30.36
Heuiristic k-means	TF-IDF BERT	Clinical terms + SNOMED CT Words	26.69 32.13
Proposed Method	TF-IDF BERT	Clinical terms + SNOMED CT Words	60.04 28.53

Table 4. NMI scores of various methods at $k = 3$

Algorithms	Feature Type	Views	NMI score
NMF (Ling *et al*. [16])	Count	Words Symptom/Medication All 3 views	0.0228 0.1273 0.1459
NMF (Ling *et al*. [16])	TF-IDF	Words Symptom/Medication All 3 views	0.0020 0.1606 0.1711
k-means	TF-IDF BERT	Clinical terms + SNOMED CT Words	0.1219 0.0021
k-medoid	TF-IDF BERT	Clinical terms + SNOMED CT Words	0.0122 0.0015
Heuiristic k-means	TF-IDF BERT	Clinical terms + SNOMED CT Words	0.0115 0.0022
Proposed Method	TF-IDF BERT	Clinical terms + SNOMED CT Words	0.1790 0.1321

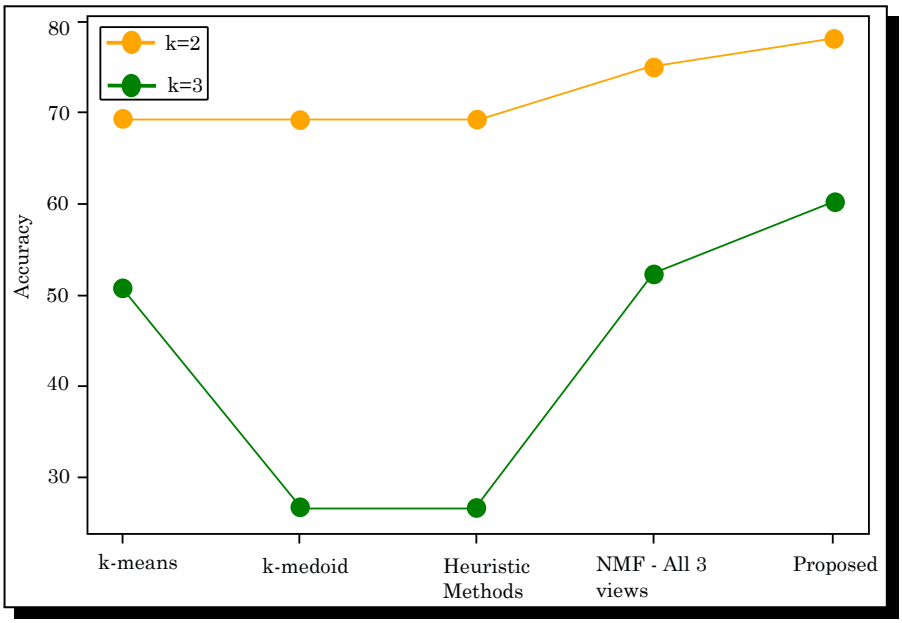

Figure 4. Overall accuracy at $k = 2$ and $k = 3$

5. Conclusion and Future Work

This article demonstrates the significance of clustering in clinical documents in a straightforward and concise manner. Documents are initially collected and cleaned up. A letter matching algorithm is used to replace abbreviations with expansions. For the purpose of extracting the entities, the scispaCy core model is applicable. Following that, the quality of

the data from clinical notes improved with the help of the SNOMED CT ontology. Afterward, the enhanced data was processed using a clustering algorithm based on the Red-Black Tree. By reading this work, other researchers will be able to delve deeper and gain a better understanding of the numerous methodologies and measurements that may be used to analyze and, as a result, develop an efficient model. In the future, these findings could be utilized in a comparison analysis of coronary artery disease patients with risk factors such as gender, genetic variables, family history, age, smoking, and dietary habits. This would help healthcare providers to deliver more effective interventions for both fatal and non-fatal heart attacks.

Competing Interests

The authors declare that they have no competing interests.

Authors' Contributions

All the authors contributed significantly in writing this article. The authors read and approved the final manuscript.

References

[1] D. Cai, X. He and J. Han, Locally consistent concept factorization for document clustering, *IEEE Transactions on Knowledge and Data Engineering* **23**(6) (2010), 902 – 913, DOI: 10.1109/TKDE.2010.165.

[2] R. Cohen, M. Elhadad and N. Elhadad, Redundancy in electronic health record corpora: analysis, impact on text mining performance and mitigation strategies, *BMC Bioinformatics* **14** (2013), Article number: 10, DOI: 10.1186/1471-2105-14-10.

[3] T. H. Cormen, C. E. Leiserson, R. L. Rivest and C. Stein, *Introduction to Algorithms*, 4th edition, The MIT Press, Cambridge, 1312 pages (2022).

[4] D. Das, Y. Katyal, J. Verma, S. Dubey, A. D. Singh, K. Agarwal, S. Bhaduri and R. K. Ranjan, Information retrieval and extraction on COVID-19 clinical articles using graph community detection and Bio-BERT embeddings, in: *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020 (NLP-COVID19)*, K. Verspoor, K. B. Cohen, M. Dredze, E. Ferrara, J. May, R. Munro, C. Paris and B. Wallace (editors), Association for Computational Linguistics, (2020), URL: https://aclanthology.org/2020.nlpcovid19-acl.7.

[5] D. Dessi, D. R. Recupero, G. Fenu and S. Consoli, Exploiting cognitive computing and frame semantic features for biomedical document clustering, in: *Proceedings of the Workshop on Semantic Web Solutions for Large-scale Biomedical Data Analytics* (*SeWeBMeDA 2017*) co-located with *14th Extended Semantic Web Conference* (*ESWC 2017*) (Portoroz, Slovenia, May 28, 2017), A. Hasnain, A. Sheth, M. Dumontier and D. Rebholz-Schuhmann, Vol. 1948 (2017), pp. 20 – 34, URL: https://ceur-ws.org/Vol-1948/paper3.pdf.

[6] K. Doing-Harris, O. Patterson, S. Igo and J. Hurdle, Document sublanguage clustering to detect medical specialty in cross-institutional clinical texts, in: *DTMBIO'13: Proceedings of the 7th International Workshop on Data and Text Mining in Biomedical Informatics*, pp. 9 – 12 (2013), DOI: 10.1145/2512089.2512101.

[7] S. El-Sappagh, F. Franda, F. Ali and K.-S. Kwak, SNOMED CT standard ontology based on the ontology for general medical science, *BMC Medical Informatics and Decision Making* **18** (2018), Article number: 76, DOI: 10.1186/s12911-018-0651-5.

[8] K. R. Gøeg, R. Cornet and S. K. Andersen, Clustering clinical models from local electronic health records based on semantic similarity, *Journal of Biomedical Informatics* **54** (2015), 294 – 304, DOI: 10.1016/j.jbi.2014.12.015.

[9] X. Huang, X. Zheng, W. Yuan, F. Wang and S. Zhu, Enhanced clustering of biomedical documents using ensemble non-negative matrix factorization, *Information Sciences* **181**(11) (2011), 2293 – 2302, DOI: 10.1016/j.ins.2011.01.029.

[10] M. Ivanović and Z. Budimac, An overview of ontologies and data resources in medical domains, *Expert Systems with Applications* **41**(11) (2014), 5158 – 5166, DOI: 10.1016/j.eswa.2014.02.045.

[11] E. K. Jasila, N. Saleena and K. A. A. Nazeer, An efficient document clustering approach for devising semantic clusters, *Cybernetics and Systems* (2023), 1 – 18, DOI: 10.1080/01969722.2023.2175135.

[12] E. K. Jasila, N. Saleena and K. A. A. Nazeer, Ontology based document clustering - An efficient hybrid approach, in: *IEEE 9th International Conference on Advanced Computing* (IACC) (Tiruchirappalli, India, 2019), pp. 153 – 157 (2019), DOI: 10.1109/IACC48062.2019.8971594.

[13] A. E. W. Johnson, T. J. Pollard, L. Shen, L.-W. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi and R. G. Mark, MIMIC-III, a freely accessible critical care database, *Scientific Data* **3** (2016), Article number: 160035, DOI: 10.1038/sdata.2016.35.

[14] F. K. Khattak, S. Jeblee, C. Pou-Prom, M. Abdalla, C. Meaney and F. Rudzicz, A survey of word embeddings for clinical text, *Journal of Biomedical Informatics* **100** (2019), 100057, DOI: 10.1016/j.yjbinx.2019.100057.

[15] Y. Li, J. Cai and J. Wang, A text document clustering method based on weighted Bert model, in: *2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference* (ITNEC) (Chongqing, China, 2020), Vol. **1**, pp. 1426 – 1430, (2020), DOI: 10.1109/ITNEC48623.2020.9085059.

[16] Y. Ling, X. Pan, G. Li and X. Hu, Clinical documents clustering based on medication/symptom names using multi-view nonnegative matrix factorization, *IEEE Transactions on NanoBioscience* **14**(5) (2015), 500 – 504, DOI: 10.1109/TNB.2015.2422612.

[17] M. Marcińczuk, M. Gniewkowski, T. Walkowiak and M. Będkowski, Text document clustering: Wordnet vs. TF-IDF vs. Word embeddings, in: *Proceedings of the 11th Global Wordnet Conference*, University of South Africa, 2021, P. Vossen and C. Fellbaum (editors), Global Wordnet Association pp. 207 – 214 (2021), URL: https://aclanthology.org/2021.gwc-1.24.

[18] S. M. Meystre, G. K. Savova, K. C. Kipper-Schuler and J. F. Hurdle, Extracting information from textual documents in the electronic health record: A review of recent research, *Yearbook of Medical Informatics* **17**(1) (2008), 128 – 144, DOI: 10.1055/s-0038-1638592.

[19] K. A. A. Nazeer, S. D. M. Kumar and M. P. Sebastian, Enhancing the k-means clustering algorithm by using a $O(n \log n)$ heuristic method for finding better initial centroids, in: *2011 Second International Conference on Emerging Applications of Information Technology* (Kolkata, India, 2011), pp. 261 – 264 (2011), DOI: 10.1109/EAIT.2011.57.

[20] O. Patterson and J. F. Hurdle, Document clustering of clinical narratives: a systematic study of clinical sublanguages, *AMIA Annual Symposium Proceedings* **2011** (2011), 1099 – 1107.

[21] V. Renganathan, Text mining in biomedical domain with emphasis on document clustering, *Healthcare Informatics Research* **23**(3) (2017), 141 – 146, DOI: 10.4258/hir.2017.23.3.141.

[22] K. Roberts and S. M. Harabagiu, A flexible framework for deriving assertions from electronic medical records, *Journal of the American Medical Informatics Association* **18**(5) (2011), 568 – 573, DOI: 10.1136/amiajnl-2011-000152.

[23] A. S. Schwartz and M. A. Hearst, A simple algorithm for identifying abbreviation definitions in biomedical text, in: *Biocomputing*, pp. 451 – 462 (2003), DOI: 10.1142/9789812776303_0042.

[24] S. Shah and X. Luo, Exploring diseases based biomedical document clustering and visualization using self-organizing maps, in: *2017 IEEE 19th International Conference on e-Health Networking, Applications and Services (Healthcom)* (Dalian, China, 2017), pp. 1 – 6, IEEE (2017), DOI: 10.1109/HealthCom.2017.8210791.

[25] H. Shelar, G. Kaur, N. Heda and P. Agrawal, Named entity recognition approaches and their comparison for custom NER model, *Science & Technology Libraries* **39**(2) (2020), 324 – 337, DOI: 10.1080/0194262X.2020.1759479.

[26] B. Shickel, P. J. Tighe, A. Bihorac and P. Rashidi, Deep EHR: A survey of recent advances in deep learning techniques for Electronic Health Record (EHR) analysis, *IEEE Journal of Biomedical and Health Informatics* **22**(5) (2018), 1589 – 1604, DOI: 10.1109/JBHI.2017.2767063.

[27] P.-N. Tan, M. Steinbach, A. Karpatne and V. Kumar, *Introduction to Data Mining*, 2nd edition, Pearson, London (2019).

[28] C. Tang, J. M. Plasek, Y. Xiong, Z. Zhang, D. W. Bates and L. Zhou, A clustering algorithm based on document embedding to identify clinical note templates, *Annals of Data Science* **8** (2021), 497 – 515, DOI: 10.1007/s40745-020-00296-8.

[29] P. Yadav, M. Steinbach, V. Kumar and G. Simon, Mining Electronic Health Records (EHRs): A survey, *ACM Computing Surveys* **50**(6) (2018), Article number: 85, 1 – 40, DOI: 10.1145/3127881.

[30] R. Zhang, S. Pakhomov and G. B. Melton, Longitudinal analysis of new information types in clinical notes, *AMIA Summits on Translational Science – Proceedings* **2014** (2014), 232 – 237.

[31] R. Zhang, S. Pakhomov, B. T. McInnes and G. B. Melton, Evaluating measures of redundancy in clinical texts, in: *AMIA Annual Symposium Proceedings*, Vol. 2011, p. 1612, American Medical Informatics Association (2011).