HHS Public Access

Author manuscript

Comput Biol Med. Author manuscript; available in PMC 2025 November 01.

Published in final edited form as:

Comput Biol Med. 2024 November ; 182: 109144. doi:10.1016/j.compbiomed.2024.109144.

Dementia Risk Prediction Using Decision-Focused Content Selection from Medical Notes

Shengyang Li[a], **Paul Dexter**[b,c], **Zina Ben-Miled**[a,c,*], **Malaz Boustani**[b,c]

[a]Department of Electrical and Computer Engineering, Indiana University-Purdue University Indianapolis (IUPUI), Indianapolis, Indiana, 46202, USA

[b]Indiana University School of Medicine, 340 W. 10th St, Indianapolis, Indiana, 46202, USA

[c]Regenstrief Institute, Inc., 1101 W. 10th Street, Indianapolis, Indiana, 46202, USA

Abstract

Several general-purpose language model (LM) architectures have been proposed with demonstrated improvement in text summarization and classification. Adapting these architectures to the medical domain requires additional considerations. For instance, the medical history of the patient is documented in the Electronic Health Record (EHR) which includes many medical notes drafted by healthcare providers. Direct processing of these notes may not be possible because the computational complexity of LMs imposes a limit on the length of input text. Therefore, previous applications resorted to content selection using truncation or summarization of the text. Unfortunately, these text processing techniques may lead to information loss, redundancy or irrelevance. In the present paper, a decision-focused content selection technique is proposed. The objective of this technique is to select a subset of sentences from the medical notes of a patient that are relevant to the target outcome over a predefined observation period. This decision-focused content selection methodology is then used to develop a dementia risk prediction model based

*Corresponding author: zmiled@iu.edu (Z. Ben-Miled).

Declaration of Competing Interest

Dr. Ben Miled has a financial interest in DigiCare Realized and could benefit from the results of this research. Dr. Boustani serves as the Chief Scientific Officer and co-Founder of BlueAgilis; and the Chief Health Officer of DigiCare Realized, Inc. He has equity interest in Blue Agilis, Inc; DigiCare Realized, Inc; he sold his equity in Preferred Population Health Management LLC; and MyShift, Inc (previously known as RestUp, LLC). He serves as an advisory board member for Eli Lilly and Co; Eisai, Inc; Merck & Co Inc; Biogen Inc; and Genentech Inc. These conflicts have been reviewed by Indiana University and have been appropriately managed to maintain objectivity. The remaining authors declare no competing interests.

CRediT authorship contribution statement

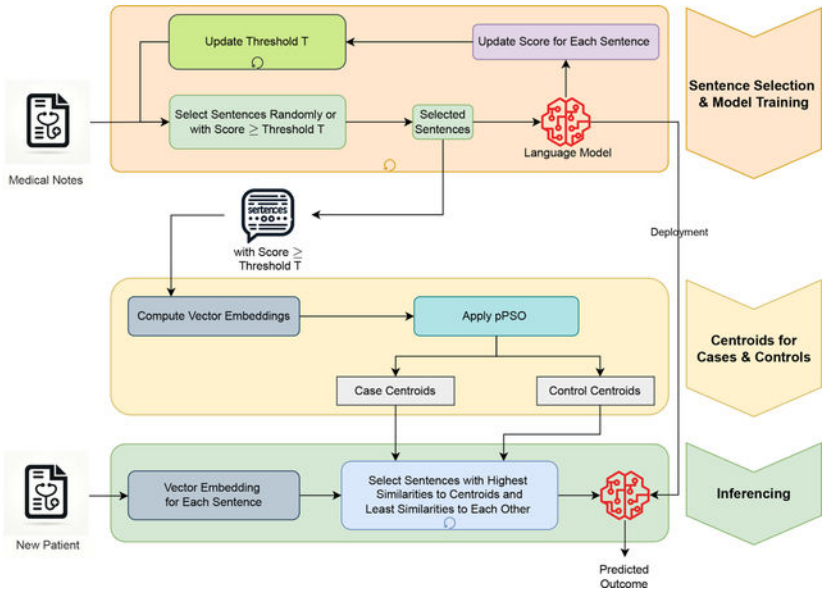
Shengyang Li: Conceptualization, Methodology, Software, Validation, Investigation, Writing - Original Draft. **Paul Dexter:** Conceptualization, Validation, Investigation, Resources, Writing - review & editing. **Zina Ben-Miled:** Conceptualization, Methodology, Validation, Investigation, Writing - review & editing. **Malaz Boustani:** Conceptualization, Methodology, Validation, Investigation, Writing - review & editing.

Institutional Review

This study was approved by the Review Board of Indiana University (IRB number: 2008372812). The institutional review entitled a) the collection of data through noninvasive procedures routinely employed in clinical practice, b) research involving materials (data, documents, records, or specimens) that have been collected and c) research on individual or group characteristics or behavior, or research employing survey, interview, oral history, focus group, program evaluation, human factors evaluation, or quality assurance methodologies. We confirm that all methods were performed in accordance with the relevant guidelines and regulations.

on the Longformer LM architecture. The results show that the proposed framework delivers an AUC of 78.43 when the summary is restricted to 1024 tokens, outperforming previously proposed content selection techniques. This performance is notable given that the model estimates dementia risk with a one year prediction horizon, relies on an observation period of only one year and solely uses medical notes without other EHR data modalities. Moreover, the proposed techniques overcome the limitation of machine learning models that use a tabular representation of the text by preserving contextual content, enable feature engineering from raw text and circumvent the computational complexity of language models.

Graphical Abstract:



Keywords

Dementia; Transformers; Supervised learning; Unsupervised learning; Medical notes; Risk prediction

1. Introduction

One of the promises of digital health is the ability to assist health providers with scalable decision support systems that can identify patients who are at risk for chronic diseases such as dementia [1, 2, 3]. Preferably, these systems should use routinely collected Electronic Health Record (EHR) data so they can be applied to a large patient population without incurring the cost or burden associated with additional laboratory or assessment tests. The patient's history is documented in the EHR in the form of structured (e.g., medications, diagnosis, lab results) or unstructured (e.g., discharge notes) data. Structured data are easier to model because they consist of well-defined exposure variables and value ranges which can be encoded in a tabular format [4, 5]. Medical notes are information rich. However, translating notes into features for classification purposes remains an open research problem [6, 7].

Feature engineering from medical notes follows two main strategies: topic modeling [8, 9] and text encoding using a pre-trained language model (LM) [7, 10]. Each of these strategies has advantages and disadvantages. Topic models consist of a vocabulary of keywords that discriminate among the classes of the target outcome (e.g., at risk for dementia versus healthy patients). These models are hard to develop and often involve substantial human-in-the-loop efforts. However, once developed, they can be applied to medical notes of any length. The second strategy applies a LM to encode the text with minimal pre-processing. LMs are transformer models that automatically extract features from the input text. They have been showing increasingly good performance for several tasks including question-answering [11, 12] and inferencing [13, 14]. However, LMs have difficulties processing long text such as the collection of several providers' notes from the patient's medical history. LMs tokenize the words in the text and encode the semantic relationship between these tokens using an attention mechanism. The complexity of this attention mechanism imposes a limit on the number of tokens that can be processed. This limit is 512 for BERT [15] and 4,096 for Longformer [16]. Techniques to improve the span of the input text either 1) reduce the complexity of the attention mechanism [17, 18] or 2) pre-process the input text to reduce its length.

Reduced complexity attention mechanisms were accomplished with various types of attention patterns [16, 19], attention approximation [20, 21], and sparse activation of partial parameters [22, 23]. For instance, Longformer [16] was able to reduce the original quadratic attention complexity in BERT to a linear complexity by using a limited left and right context window which is augmented with a sparse global attention mechanism. A similar approach is used in compute-intensive large language model (LLM) which are now able to accommodate up to 10M tokens [24]. Unfortunately, the result of recent studies [25, 26] indicate that LLMs still struggle to retrieve and use relevant information when it is not located at the beginning or end of long texts. As such, many LLMs may not deliver the expected performance if they fail to retrieve information relevant to the target outcome from lengthy texts.

The second technique which can increase the span of the input text focuses on pre-processing the text prior to modeling by using methodologies such as:

- Splitting the text into fixed-length segments which are processed independently. Outcomes induced by the segments are then combined using linear or non-linear regression [27, 28]. This methodology ignores the relationships between different segments.
- Truncating the text by sampling the first, last, or randomly selected subset of the tokens [29, 30]. Truncation may disregard crucial information [31].
- Selecting only sentences of interest from the text using data mining [32, 33].

The present paper introduces a meta-algorithm which integrates content selection [32] and summarization [34] with modeling. Content selection, a data mining approach, identifies content relevant to the target outcome. Summarization imposes a limit on the desired number of tokens from the selected content to meet the input restrictions of the model. The classification model closes the meta-learning feedback loop by scoring the sentences in

the selected content according to their contributions towards the prediction of the expected outcome.

The ability to process long text is necessary to transform medical notes stored in EHR systems into useful information (e.g., risk for dementia) to health providers. The decision focused content selection takes advantage of the ability of LM to extract contextual features from the text in these notes while overcoming the computational complexity and the limitation of these models to extract relevant information from long text.

2. Related Works

The ultimate objective of the present study is to show that a dementia risk prediction classifier can be developed using a LM classifier coupled with a decision-focused content selection methodology from medical notes. This section provides a background related to this research objective including a summary of most recently proposed dementia risk models, while highlighting the differences between detection and prediction models, and the practical implications of the characteristics of the feature space of these models. This summary is followed by a brief review of content selection techniques and advanced transformer architectures for long text.

2.1. Dementia Risk Models using EHR data

Early risk prediction and accurate detection of dementia are important for appropriate care management and potential intervention. Several previous studies, including those by the authors, proposed dementia risk models using retrospective EHR data. These models are distinguished based on two important characteristics: observation period and prediction horizon. For instance, a prediction horizon of one year and an observation period of two years indicates that the model relies on the patient medical history over the past two years to estimate dementia risk one year prior to the onset of the disease. Longer observation periods are not clinically practical since the patient's older records may not be available. However, they are more likely to yield better performance. Extended prediction horizons differentiate between prediction and detection models. Prediction models are less accurate since they do not have access to the most recent health status of the patient.

The above trade-offs are rarely clarified in the literature. For instance, a dementia detection model (i.e., prediction horizon = 0 years) using medical notes was able to achieve 81% accuracy with a fine-tuned LM and truncation for content selection [35]. Another LM detection model achieved 90% accuracy using doctor-patient speech transcripts [36].

Compared to dementia detection, the goal of dementia risk prediction is to identify patients at risk as early as possible when they are pre-symptomatic. A dementia risk prediction model using random forest (RF) was described in [37]. This prediction model achieved an AUC of 0.72 for a 7-year prediction horizon compared to the baseline dementia detection model which had an AUC of 0.81. Both models relied on structured features encompassing the demographics and history of diagnosis of the patient.

Another dementia risk prediction model was developed using gradient boosting trees (XGB) across prediction horizons of 1, 3 and 5 years [4]. The features of the model included demographics, medications, history of diagnosis, and procedures. AUCs of 0.9 and 0.85 were reported for prediction horizons of 1 and 5 years, respectively. However, the models also had low positive predictive values (PPV ≈ 0.4–0.5) indicating a high proportion of false positives.

The authors of the present paper have also been actively engaged in developing, validating, improving and deploying dementia risk prediction models. In an initial study, we compared models with 1 and 3 year prediction horizons which were derived form multiple EHR modalities: demographics, history of diagnosis, medications, and medical notes [8]. The modalities were all represented in a tabular format and used to train a RF model. The data from all the EHR modalities under consideration were aggregated over a ten-year observation period when available in the EHR. For medical notes, accommodating this extended observation period was possible since the modality was encoded in a tabular format following the topic modeling feature strategy mentioned above.

The above study provided some important results: 1) as expected, the 1 year prediction horizon model achieved an accuracy of 0.72 outperforming the 3 year prediction horizon model; 2) an extended observation period is restrictive in practice since the required EHR data may not be available for a large number of the patients; and 3) medical notes make an important contribution since the difference in accuracy of the model based solely on the medical notes versus all the EHR modalities was less than 3%. Subsequent improvements focused on mapping the topic keywords from medical notes to UMLS concepts [38] and reducing the observation period from 10 years to 2 years. The UMLS mapping reduced the dimensionality of the feature space and improved the explainability of the model while maintaining its performance despite the shorter observation period. The resulting 1 year prediction horizon model was deployed and integrated in the clinical workflow of two health systems and is being validated against gold standard cognitive assessments as part of an ongoing pragmatic trial [39].

Ongoing research efforts by the authors continue to focus on improving dementia risk prediction from medical notes. An XGB model with an observation period of 2 years and a prediction horizon of 1 year using the UMLS concepts described above delivered better performance than two models that encoded the topic keywords using two LMs (BERT and clinical BERT [40]) [6]. We conjuncture that the lower performance of the latter LM models is due to the use of embeddings of the topic keywords rather than the raw medical notes; thereby nullifying context learning, one of the main strengths of LMs.

The present study takes into consideration practical and technical lessons learned from previous studies by the authors and other researchers. Namely, the observation period of the model should ideally be limited to two years and the encoding of the medical notes has to observe context, input length and relevance.

2.2. Content Selection

Several studies evaluated classifiers based on LMs with simple content selection strategies such as truncation. Selecting the first subset of tokens, a random subset of tokens, and tokens with the highest term frequency inverse document frequency (TF-IDF) scores were compared in [41]. These various truncation techniques fared better with different datasets. Similar findings were reported in [29] indicating that simple content selection strategies may be dataset-depend. Therefore, an advanced, dataset-agnostic content selection strategy is needed for long texts.

The idea of decision-focused content selection is based on the assumption that only some of the sentences in the text include information that is relevant to the target task [32]. As argued in [34], information related to a fracture may not help with understanding the risk of insomnia. CogLTX [32] introduced a novel approach to sentence selection, employing a two-step process involving a judge BERT model and a reasoner BERT model. Initially, the long text is partitioned into blocks, and the judge model assesses the relevance of each token within each block. A relevance score is assigned to each block which is calculated as the average of the token scores within the block. The blocks with the highest scores are used to train the reasoner model.

DecSum offers a different alternative to decision-focused sentence selection [34]. A predictive model is first trained on the entire text. This model is then used to select representative sentences from the original text. The selection process is guided by an objective function that considers three factors: 1) the ability of the selected sentences to infer similar decisions as the full-text model, 2) similarity in the distributions of these decisions, and 3) textual non-redundancy in the selected sentences. One of the major limitations of DecSum is the assumption that a sufficiently accurate model that can handle long text is already available for the target task where in fact creating such a model is the ultimate goal of content selection.

2.3. Transformer Architectures for Long text

Transformer-XL [42] uses segment-level recurrence and relative positional encoding to allow for the processing of an unlimited number of tokens. The segment recurrence resembles the structure of recurrent neural networks. The vector representations computed for the previous segments are reused to extend the context of the current segment. This allows contextual inter-segment information to be learned beyond the individual segment boundaries. The absolute positional encoding in the standard transformer was initially designed to allow the attention mechanism to consider the relative position of tokens within a segment. Using the representations of the previous segments along with the current segment in Transformer-XL can lead to ambiguity at the level of the positional encoding between two segments. Therefore, the absolute positional encoding is replaced with a sinusoidal position encoding matrix that represents the distance between two words instead of the position of the word within a segment, and a trainable parameter is added to compute the inter-segment attention score. This extended inter-segment contextual reach in Transformer-XL comes at the expense of limiting the context to only the left-to-right direction as opposed to the bidirectional context used in BERT.

Longformer-Encoder-Decoder (LED) is a variant of the Longformer model [16] with an extended token length limit of 16,384. This LM is a sequence-to-sequence model with an encoder-decoder architecture which is specifically designed for generative tasks such as summarization and translation. In contrast, its predecessor, Longformer, has an encoder-only architecture. As in Longformer, the encoder of LED utilizes a modified attention mechanism which consists of a combination of local and global attention. Local attention limits the local context to a window of tokens surrounding the target token. Global attention refers to an attention technique where only a selected set of task-specific tokens can attend to and be attended by all tokens in the sequence. LED is based on the BART [30] LM. In the present study, the versions of LED that were fine-tuned on the BookSum dataset [43] and the PubMed datasets [44] are used as baselines for benchmarking purposes.

Bigbird [19] also uses the combination of a window-based local attention and an extended global attention, a reduced complexity attention mechanism which was first introduced in Longformer. However, the complexity of the attention mechanism is further reduced by using sparsification following the Watts-Strogatz graph. The resulting sparse attention was shown to preserve the properties of the full attention mechanism.

3. Methods

A meta-algorithm for dementia risk prediction is proposed. This algorithm uses a transformer architecture to identify sentences that are relevant to the target classification task. The selected sentences, subject to a non-redundancy criterion, are re-used to fine-tune the same transformer model. Once this model is trained, it is used along with the centroids of the selected sentences for inferencing as shown in Figure 1.

The present study is non-interventional and uses retrospective EHR data. All methods were subject to IRB approval and performed in accordance with the relevant guidelines and regulations.

3.1. Data Extraction

The dataset used in this study is collected from the Indiana Network for Patient Care (INPC) through the Regenstrief Data Core. INPC is a centralized community-wide electronic medical record that collects data from 19 hospitals in seven health systems, the Marion County Health Department, regional laboratories, radiology centers, and various physician practices. The data collected include patients' demographics, laboratory results, emergency department, inpatient, and outpatient encounter data, free-text chief complaints, coded diagnoses and procedures, vital signs, and other clinical data.

The EHR data under consideration in the present study consists of medical notes which are represented in textual format. These health providers' notes are extracted for two cohorts: dementia (cases) and healthy control (controls). Only patients with incident dementia were included in the first cohort. Patients with mild cognitive impairment (MCI) or previous dementia are excluded from the study. Matching controls, with no dementia or MCI diagnosis, are identified for each case based on sex, race, and age (within 5 years of the

disease index date of the matching case). Moreover, two observation periods are considered: 1 and 2 years in conjunction with two prediction horizons: 0.5 (6 months) and 1 year.

All the medical notes within the observation period are collected in a single document organized chronologically starting with the oldest notes. As discussed earlier, the order of notes is important for content selection techniques that are based on truncation since older notes most likely include less relevant information than more recent notes. Standard text pre-processing is then applied to the document containing all the medical notes including replacing numerals with spaces and removing text formatting tags and report titles. This step is necessary for health data as suggested in [45]. Finally, the medical notes are split into sentences using the NLTK sentence tokenizer [46] and duplicate sentences are removed.

3.2. Definitions

Few parameters and variables are introduced in Table 1 to facilitate the description of the proposed meta-algorithm. When applicable, the label (i) is used to refer to the values of these parameters with respect to iteration $1 \leq i \leq f$ of the meta-algorithm. Within each iteration of the meta-algorithm, there are nested iterative processes. The term epoch k is used to refer to these inner iterations.

3.3. Model Development

The patients in the cohort are split 67/33 into training (Dtr) and testing (Dts) datasets while maintaining an equal balance between cases and controls. As mentioned above, the pre-processed medical notes of each patient over the observation period are collected in a document d^p which consists of multiple sentences. The objective is to iteratively construct $\widetilde{d}^{p(f)}$ and develop the model $M^{(f)}$ such that the predicted disease outcome $\widetilde{y}^{p(f)} \approx y^p$ where y^p is the actual disease outcome indicating cases (i.e. patients at risk of dementia) or healthy controls.

In each iteration i of the meta-algorithm and for each patient p in Dtr, the algorithm selects the most relevant sentences from d^p to construct $\widetilde{d}^{p(i)}$. This subset and the corresponding outcomes y^p are then used to train $M^{(i)}$. The predicted outcome $\widetilde{y}^{p(i)}$ produced by the trained model $M^{(i)}$ is in turn used to update the sentence relevance score $r^{(i)}(s_x^p)$ and the threshold $T^{(i)}$ which are then used to generate the new set of sentences $\widetilde{d}^{p(i+1)}$ for each patient in the next iteration. This iterative process stops when $i = f$ iterations of the meta-model are completed.

The following components are executed in each iteration of the meta-algorithm (Fig. 2).

a) Constructing $\widetilde{d}^{p(i)}$: Non-repetitive sentences are selected one at a time from d^p until the limit L is reached. This selection is modeled after the multi-armed bandit problem since there is a trade-off between choosing random sentences from d^p (i.e., exploration) versus choosing sentences with relevance scores $r^{(i)}(s_x^p) \geq T^{(i)}$ (i.e., exploitation). This trade-off is controlled by the parameter $\epsilon^{(i)}$. That is, sentences in d^p are randomly selected with a

probability $\epsilon^{(i)}$. Conversely, sentences with relevance score greater than $T^{(i)}$ are selected with a probability $1 - \epsilon^{(i)}$. Moreover, when selecting sentences with a relevance score greater than $T^{(i)}$, the sentence s_x^p with minimum sum of maximum cosine similarities to previously selected sentences is chosen as defined in Equation 1:

$$s_x^p = \underset{r^{(i)}(s_x^p) \geq T^{(i)}}{\arg\min} \sum_{s_w^p} \max_{s_y^p} S^p(s_w^p, s_y^p), \tag{1}$$

where $s_w^p \in \tilde{d}^{p(i)} + s_x^p$ and $s_y^p \in \tilde{d}^{p(i)} + s_x^p - s_w^p$. The purpose of the selection criterion in Equation 1 is to promote textual non-redundancy so that the sentences selected are dissimilar from one another.

In the first iteration of the meta-algorithm, $\epsilon^{(1)} = 1$ and all the sentences are selected at random from d^p. Effectively, $T^{(1)} = 1$ and $r^{(1)}(s_x^p) = 0$. Thereafter, $\epsilon^{(i)} = \left(f - i + 1\right)/f$ follows a decaying trajectory that encourages less exploration and more exploitation when approaching the last iteration, f, of the meta-algorithm.

b) Training the classifier model $M^{(i)}$: The classifier model is trained with $\tilde{d}^{p(i)}$ for each patient $p \in Dtr$. A classification head is added to the Longformer model. This head consists of multiple layers ending with a softmax linear layer which is optimized using the cross-entropy loss. The classification layer accepts the pooled output from the Longformer model.

c) Updating the relevance scores: The classifier model $M^{(i)}$ produces the estimated disease outcome $\tilde{y}^{p(i)}$ for each patient $p \in Dtr$. However, only sentences that were included in $\tilde{d}^{p(i)}$ participate in this classification. Therefore, the relevance score for these sentences is updated as follows:

$$r^{(i)}(s_x^p) = \frac{\sum_{s_x^p \in \tilde{d}^{p(j)}}^{j \leq i} q^{(j)}(s_x^p)}{\sum_{s_x^p \in \tilde{d}^{p(j)}}^{j \leq i} 1}. \tag{2}$$

where

- $q^{(i)}(s_x^p) = 1$ if $\tilde{y}^{p(i)} = y^p$. These are the true positive and true negative predictions.
- $q^{(i)}(s_x^p) = -1$ if $\tilde{y}^{p(i)} \neq y^p$. These are the false positive and false negative predictions.

The normalization denominator of Equation 2 represents the total number of iterations in which s_x^p participated in the training of model $M^{(i)}$ for each iteration i of the meta-algorithm.

d) Calculating $T^{(i+1)}$: The optimal threshold $T^{(i)}$ is used to select sentences from d^p at each iteration of the meta-algorithm. As mentioned above, for the first iteration, this threshold is set to 1. Thereafter, $T^{(i+1)}$ is calculated using an iterative process over k_{max} epochs. For each epoch $1 \leq k \leq k_{max}$, a candidate threshold $\widetilde{T}_k^{(i+1)} = 1 - 2(k-1)/k_{max}$ is proposed. The next step decides if the proposed new threshold value is accepted or the old value is retained.

A subset of patient medical notes is randomly selected from Dtr. From those notes, sentences with $r^{(i)}(s_x^p) \geq \widetilde{T}_k^{(i+1)}$ are considered one at a time following the textual-non redundaney requirement of Equation 1. Sentence sampling stops when the limit L is reached or when there are no more sentences that satisfy $r^{(i)}(s_x^p) \geq \widetilde{T}_k^{(i+1)}$. The average loss $\mathscr{L}_{M^{(i)}}(\widetilde{T}_k^{(i+1)})$ generated by $M^{(i)}$ is computed for the selected sentences. Let $\Delta\mathscr{L}^{(i)}$ represent the difference in average loss of $M^{(i)}$ when the old versus proposed thresholds are used. That is,

$$\Delta\mathscr{L}^{(i)} = \mathscr{L}_{M^{(i)}}(T^{(i+1)}) - \mathscr{L}_{M^{(i)}}(\widetilde{T}_k^{(i+1)}). \tag{3}$$

In the first epoch, $T^{(i+1)}$ is initialized to $T^{(i)}$. Thereafter, following a simulated annealing update, $T^{(i+1)} = \widetilde{T}_k^{(i+1)}$ with a probability p where

$$p = \begin{cases} 1 & \text{if } \Delta\mathscr{L}^{(i)} \geq 0, \\ e^{\frac{\Delta\mathscr{L}^{(i)}}{t}} & \text{if } \Delta\mathscr{L}^{(i)} < 0, \end{cases} \tag{4}$$

and $T^{(i+1)}$ remains the same with a probability of $1 - p$.

The decaying temperature in Equation 4 is defined as $t = \lambda(1 - i/f + 0.01)$, where λ is a configurable parameter that is used to control the sensitivity of the simulated annealing with respect to the loss value $\Delta\mathscr{L}^{(i)}$. In the present study, λ is set to 0.1. When the iteration i of the meta-algorithm approaches the final one (i.e., f) exploitation (i.e., $T^{(i+1)}$ remains unchanged) is preferred over exploration (i.e., $T^{(i+1)} = \widetilde{T}_k^{(i+1)}$).

It should be noted that for some patients with false predictions at every iteration, the relevance score of the sentences is always negative or none of the sentences will have a relevance score greater than $T^{(i)}$. However, these patients still participate in the training of M in subsequent iterations and the sentences for these patients will be selected from d^p as if this was the first iteration of the meta-algorithm. Also when constructing $\widetilde{d}^{p(i)}$, if there are no non-overlapping sentences that can be selected from d^p or no more non-overlapping sentences with relevance scores $r^{(i)}(s_x^p) \geq T^{(i)}$, the sentence selection terminates.

3.4. Inferencing

The meta-algorithm used to construct $M^{(f)}$ outlined above follows a supervised learning approach that depends on the availability of the disease outcome y^p for each patient. The objective of this learning is two-fold: 1) selecting the best sentences from d^p for each patient p in Dtr and 2) accurately predicting y^p. Specifically, the disease outcome y^p was used in the calculation of the relevance score of each sentence in d^p (Equation 2).

The goal of inferencing is to estimate the disease outcome for new patients. Prospectively, the true disease outcome is not available for these patients. Therefore, an unsupervised sentence selection process is needed. The held-out set of patients (Dts) in this study is a representative set of these new patients. The documents d^p for patients in Dts are constructed following the same standard pre-processing mentioned above for the patients in Dtr. An unsupervised sentence selection is then proposed to identify the relevant sentences that make up the summaries $\widetilde{d^p}$. These summaries are submitted to $M^{(f)}$ in order to estimate the disease outcome $\tilde{y}^p$. The steps involved in this inferencing process are depicted in Figure 3.

As mentioned above, the major difference between model development and inferencing is sentence selection. The sentence selection technique used to develop $M^{(f)}$ is supervised whereas the sentence selection used for inferencing has to be unsupervised. That said, the sentence selection proposed for inferencing re-uses the results of the sentence selection technique used for model development as described next.

Let $\widetilde{Rc}^{cases}$ and $\widetilde{Rc}^{controls}$ represent the set of sentences that satisfy $r^{(f)}(s_x^p) \geq T^{(f)}$ for all cases and controls in Dtr, respectively. The embeddings of all the sentences in $\widetilde{Rc}^{cases}$ and $\widetilde{Rc}^{controls}$ are generated using sentence-BERT. The optimization technique pPSO [48], a variant of Particle Swarm Optimization (PSO), is then used to identify α representative sentences, called centroids, from each set over the embedded space of the sentences. These centroids are denoted Rc^{cases} and $Rc^{controls}$, respectively. In the present study, the configurable hyper-parameter α is set to 200. This process reduces the number of sentences used in the inferencing and eliminates redundancy resulting from similar sentences from multiple patients.

To construct $\tilde{d}^p$ for a new patient p in Dts, non-repetitive sentences are selected one at a time until the limit L is reached while taking into consideration the following two criteria:

- The sum of the highest cosine similarities between each selected sentence and the centroids in Rc^{cases} or $Rc^{controls}$ is maximized.
- A selected sentence s_x^p must satisfy textual non-redundancy with previously selected sentences following Equation 1.

The above criteria are realized by minimizing a joint objective loss function:

$$\mathscr{L}(s_x^p) = \sum_{s_w^p} \max_{s_y^p} S^p(s_w^p, s_y^p) - \sum_{s_w^p} \max_{s_z^p} S^p(s_w^p, s_z^p), \tag{5}$$

where $s_w^p \in \widetilde{d^p} + s_x^p$, $s_y^p \in \widetilde{d^p} + s_x^p - s_w^p$, and $s_z^p \in Rc^{cases} \cup Rc^{controls}$. Once, $\widetilde{d^p}$ is constructed, it is submitted to $M^{(f)}$ which produces the estimated disease outcome $\tilde{y}^p$.

4. Results

Initially, the cohort consisted of a total of 26,236 patients: 4,796 cases and 21,440 controls. To avoid class imbalance, the healthy control patients were randomly down-sampled to 4,796 for a total number of 9,592 patients equally split between cases and controls. The mean age, sex, race and ethnicity distribution of these two cohorts are shown in Table 2.

The number of cases and controls in Table 2 are split for training (Dtr) and testing (Dts) as shown in Table 3. This table also shows that on average, the documents d^p collected from the medical notes of cases and controls over a 1 year observation exceeds 5,000 tokens and 200 sentences with a high standard deviation. In fact, nearly 40% of all d^p have a length over the limit $L = 4096$.

The proposed decision-focused content selection approach is benchmarked against baseline and state-of-the-art content selection techniques including:

- Late Truncation: Sentences from d^p are selected up to the allowable limit L starting from the earlier medical notes of the patient. Sentences that are chronologically closest to the index date within the observation period are truncated.
- Early Truncation: Sentences from d^p are selected up to the allowable limit L starting from the latest medical notes of the patient. Sentences that are chronologically the farthest from the index date within the observation period are truncated.
- Random Truncation: Sentences from d^p are randomly selected up to the allowable limit L regardless of their chronological order within the observation period.
- pPSO: Sentences sampling from d^p is based on the centroids created by pPSO from the training dataset without additional processing.
- LED_BookSum [43]: Sentences are selected from d^p using the LED LM which was fine-tuned on the BookSum dataset.
- LED_PubMed [44]: Sentences are selected from d^p using the LED LM which was fine-tuned on the PubMed dataset.

All of the above methods, including the techniques proposed in the present study, produce summaries $\widetilde{d^p}$ of d^p which consist of a set of sentences within a limit L. Both late and early truncations eliminate sentences in d^p according to a chronological criterion. For early truncation, the most recent sentences are retained. As mentioned earlier, sentences closer to the index date are more likely to have relevant information. Random truncation does not consider the chronological order of the sentences over the observation period.

pPSO creates centroids directly from the training dataset which are then used during inferencing. First, pPSO is applied to the sentence-BERT embeddings of all the sentences from the medical records of the cases in the training dataset. This step creates a set of representative centroid sentences (Rc^{cases}). The same process is repeated for all the sentences in the medical records of the control patients in the training dataset to generate the set of centroids $Rc^{controls}$. For inferencing over new patients, the summaries are generated using the same steps outlined in Section 3.4 except that in this case the centroids Rc^{cases} and $Rc^{controls}$ are directly developed using sentences from Dtr. In contrast, the approach proposed in the present study to construct Rc^{cases} and $Rc^{controls}$ first performs decision-focused sentence selection on the sentences in Dtr and then uses only the selected sentences (i.e., $\widetilde{Rc}^{cases}$ and $\widetilde{Rc}^{controls}$) to construct the centroids Rc^{cases} and $Rc^{controls}$.

LED_BookSum and LED_PubMed are fine-tuned summarization models of the original LED model that consume input text up to $L = 16{,}384$ tokens and produce an output summary with a maximum length of 1024 tokens [49]. Various parameters are available to control the generation technique of these LMs as shown in Table 4. For instance, the parameter num_beams refers to the number of beams used in the search algorithm; the repetition_penalty [50] controls the penalty for repeated content in the generated text; the parameter no_repeat_ngram_size ensures that all ngrams of that size occur only once in the output. The length of the summary is controlled by three parameters: length_penalty, max_length and min_length. A length_penalty value bigger than 0 promotes longer output summaries. The max_length is set to the minimum of the number of input tokens and the maximum number of output tokens allowable by LED (i.e., 1024). Similarly, the min_length is set to the minimum of 0.75 * the number of input tokens and the maximum number of output tokens allowable by LED (i.e., 1024). These ranges were selected to force LED to generate summaries as close as possible to the maximum allowable number of tokens (i.e., 1024) while accommodating short medical notes.

For each content selection technique, a classifier that can predict the disease outcome is developed using the same Longformer architecture described in Section 3.3 following 3-fold cross-validation. Longformer has a maximum input token length of $L = 4096$. As mentioned above, the summarization methods LED_BookSum and LED_PubMed can only generate summaries that are up to $L = 1024$ tokens. For a fair comparison, L is first set to 1024 tokens for all content selection techniques and then extended to 4096 tokens for all the techniques that are able to generate longer summaries (i.e., excluding LED_BookSum and LED_PubMed). The Longformer classifiers are trained on Nvidia A100 GPU with half-precision over 5 epochs with a batch size of 4 using the AdamW [51] optimizer and a linear warmup over 500 steps.

For the proposed method, the meta-algorithm is executed for 10 iterations ($f = 10$) with a batch size of 4, and warmup steps set to 1000. Since all the classification models, including the proposed model, use the same classification architecture (i.e., Longformer), any observed differences in the performance of the classification can be attributed to the content selection approach.

Typically, content selection techniques are evaluated using the ROUGE [52] performance metrics which are based on the structural similarities (e.g., longest common sub-sequences and n-grams) between the original documents and the generated summaries. In the present study, more conservative metrics are used. These are based on the accuracy of the dementia risk prediction produced by the respective classifiers that use the summaries as input for specific observation periods and prediction horizons. The metrics include AUC, precision, sensitivity, specificity, F1 score, and accuracy.

These measures provide a comprehensive, prevalence-independent, assessment of the performance of the model. The AUC curve is a graphical representation of the sensitivity of the model against its false positive rate (i.e., 1 - specificity) when varying binary classification thresholds are applied to the probability risk scores generated by the model. AUC is a good measure of the discrimination performance of the model (i.e., at risk dementia patients versus healthy controls), as stated in [3]. Precision, sensitivity, specificity, F1 score and accuracy are reported for a specific binary classification threshold. Changing this threshold can increase sensitivity and lower specificity. In practice, health providers require a binary decision. Therefore, the binary classification threshold can be customized depending on the prevalence of the disease condition and the tolerance for false positives in a given health care institution.

The above metrics are shown in tables 5 and 6 for $L = 1024$ and $L = 4096$, respectively when the observation period and prediction horizon are both set to 1 year. The AUC curves for these classifiers are also depicted in Figure 4 and Figure 5.

For diagnostic tests AUCs between 0.7–0.8 are considered acceptable; those between 0.8 and 0.9 are considered excellent and values above 0.9 are considered outstanding [53]. Using this scale, the proposed method is within the high acceptable range. Moreover, the results indicate that the proposed approach achieves the highest performance across all metrics. It is followed by pPSO and the truncation methods, all outperforming both LED_BookSum and LED_PubMed. This indicates that the proposed approach is more effective at selecting sentences that are relevant to dementia risk prediction. It is interesting to note that when L reduces from 4096 (Table 6) to 1024 (Table 5), the performance of the proposed method improves slightly while a performance degradation is observed for all other approaches. This may be an indication of the sensitivity of the corresponding classifier to the content selected from the original medical notes where longer summaries tend to include more irrelevant sentences.

Random sentence selection does not always perform better than both early and late truncations. This indicates that randomly selected sentences may not always be relevant to the disease outcome, especially for longer summaries. Early truncation always outperforms

late truncation for $L = 1024$ and $L = 4096$ confirming the importance of the medical notes closer to the index date.

Both pPSO and the proposed content selection technique are guided by the target outcome, whereas the other techniques are generic regardless of the outcome. According to the results (Tables 5 and 6), guided content selection techniques are needed for improved accuracy of disease specific risk predictions.

The above results are for dementia risk predictors with an observation period of 1 year and a prediction horizon of 1 year. In order to analyze the impact of varying the observation period and the prediction horizon in conjunction with the length of the summary, the observation period is extended from 1 year to 2 years and the prediction horizon is reduced from 1 year to 0.5 year (i.e., 6 months). These results are shown in tables 7 and 8. The performance metrics of the proposed risk prediction models with 1 year observation period and 1 year prediction horizon from tables 5 and 6 are repeated in tables 7 and 8, respectively for comparative convenience.

An increase in performance is observed when the prediction horizon is reduced from 1 year to 0.5 year (i.e., risk prediction closer to the onset of the disease) for both observation periods and document lengths. These classifiers show AUCs > 80% reaching the excellent range for diagnostic tests. Moreover, a performance improvement is also observed when the observation period increases from 1 to 2 years and L increases from 1024 to 4096. This result suggests that longer observation periods may make more relevant information accessible allowing the corresponding model to benefit from longer summaries.

5. Conclusion

Risk prediction for chronic diseases such as dementia is important for appropriate care management and planning. This is especially true if the prediction can be established one year prior to the onset of the disease in order to allow for medical interventions when possible. The present study introduces a model that reviews the medical notes of the patient over an observation period and produces a score indicating the likelihood that a patient will develop dementia within a predefined prediction horizon. The ultimate beneficiaries of the proposed model are health care providers in primary care clinics. These providers are over-burden and often lack the training needed to administer specialized cognitive tests. The model can be applied at scale with minimal burden since it only relies on routinely collected EHR data and it can offer primary care physicians with additional information which can support their decisions to refer at risk patients to suitable specialty care practices.

To develop this early dementia risk prediction model, a LM with a transformer architecture was used. While this architecture is suitable for text processing, it has several limitations in the context of disease risk prediction. First, the length of the notes when combined over an observation period may exceed the input limit of the LM. Second, an arbitrary content selection technique may not identify content which is relevant to the disease of interest. Therefore, a decision-focused content selection is proposed. This technique optimizes both content selection and risk prediction simultaneously. Experimental results demonstrate that

the proposed approach outperforms other content selection techniques which are based on truncation and summarization.

The results also indicate that decision-focused content selection may need to consider different lengths of summaries for different observation periods and different patients. In the present study, a static limit was used (i.e., $L = 1024$ or $L = 4096$). A variable limit, which is defined for each observation period and patient, may be more suitable. Along with exploring dynamic limits for summaries, future work also includes the fusion of medical notes with other EHR modalities as well as environmental data. Our ongoing clinical trial which enabled the deployment of the initial dementia risk prediction model in two large health care systems is in its final year. Lessons learned from this ongoing trial and the development of a new model which is based on the proposed techniques will allow us to continue improving the utility of the dementia risk prediction framework in primary care practices.

Acknowledgments

This research was supported by the National Institute on Aging, #076071-00010C. The authors would like to thank Evgenia Teal of the Regenstrief Institute for her support. Katrina Jo Coppedge reviewed and edited the manuscript. The high-performance computing services used in this study are supported in part by Lilly Endowment, Inc., through its support for the Indiana University Pervasive Technology Institute.

Biography

Dr. Ben Miled has a financial interest in DigiCare Realized and could benefit from the results of this research. Dr. Boustani serves as the Chief Scientific Officer and co-Founder of BlueAgilis; and the Chief Health Officer of DigiCare Realized, Inc. He has equity interest in Blue Agilis, Inc; DigiCare Realized, Inc; he sold his equity in Preferred Population Health Management LLC; and MyShift, Inc (previously known as RestUp, LLC). He serves as an advisory board member for Eli Lilly and Co; Eisai, Inc; Merck & Co Inc; Biogen Inc; and Genentech Inc. These conflicts have been reviewed by Indiana University and have been appropriately managed to maintain objectivity. The remaining authors declare no competing interests.

Data Availability

The data that support the findings of this study are available from the Regenstrief Institute but restrictions apply to the availability of these data, which were used under a research agreement for the current study, and so are not publicly available. Data are however available from the corresponding author upon reasonable request and with permission of the Regenstrief Institute. The code for model training and inferencing is available under the PDM laboratory.

References

[1]. Taylor B, Barboi C, Boustani M, Passive digital markers for alzheimer’s disease and other related dementias: A systematic evidence review, Journal of the American Geriatrics Society (2023).

[2]. Ford E, Milne R, Curlewis K, Ethical issues when using digital biomarkers and artificial intelligence for the early detection of dementia, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery (2023) e1492. [PubMed: 38439952]

[3]. Grant SW, Collins GS, Nashef SA, Statistical primer: developing and validating a risk prediction model, European Journal of Cardio-Thoracic Surgery 54 (2018) 203–208. [PubMed: 29741602]

[4]. Li Q, Yang X, Xu J, Guo Y, He X, Hu H, Lyu T, Marra D, Miller A, Smith G, et al. , Early prediction of alzheimer's disease and related dementias using real-world electronic health records, Alzheimer's & Dementia (2023).

[5]. Coley RY, Smith JJ, Karliner L, Idu AE, Lee SJ, Fuller S, Lam R, Barnes DE, Dublin S, External validation of the eradar risk score for detecting undiagnosed dementia in two real-world healthcare systems, Journal of General Internal Medicine 38 (2023) 351–360. [PubMed: 35906516]

[6]. Ben Miled Z, Dexter PR, Grout RW, Boustani M, Feature engineering from medical notes: A case study of dementia detection, Heliyon 9 (2023).

[7]. Novikova J, Robustness and sensitivity of bert models predicting alzheimer's disease from text, in: Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021), 2021, pp. 334–339.

[8]. Ben Miled Z, Haas K, Black CM, Khandker RK, Chandrasekaran V, Lipton R, Boustani MA, Predicting dementia with routine care emr data, Artificial Intelligence in Medicine 102 (2020) 101771. [PubMed: 31980108]

[9]. Dieng AB, Ruiz FJ, Blei DM, Topic modeling in embedding spaces, Transactions of the Association for Computational Linguistics 8 (2020) 439–453.

[10]. Ji S, Hölttä M, Marttinen P, Does the magic of bert apply to medical code assignment? a quantitative study, Computers in biology and medicine 139 (2021) 104998. [PubMed: 34739971]

[11]. Yang X, Chen A, PourNejatian N, Shin HC, Smith KE, Parisien C, Compas C, Martin C, Costa AB, Flores MG, et al. , A large language model for electronic health records, NPJ Digital Medicine 5 (2022) 194. [PubMed: 36572766]

[12]. Jin Q, Yuan Z, Xiong G, Yu Q, Ying H, Tan C, Chen M, Huang S, Liu X, Yu S, Biomedical question answering: a survey of approaches and challenges, ACM Computing Surveys (CSUR) 55 (2022) 1–36.

[13]. He Y, Zhu Z, Zhang Y, Chen Q, Caverlee J, Infusing disease knowledge into bert for health question answering, medical inference and disease name recognition, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 4604–4614.

[14]. Zhang T, Cai Z, Wang C, Qiu M, Yang B, He X, Smedbert: A knowledge-enhanced pre-trained language model with structured semantics for medical text mining, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021, pp. 5882–5893.

[15]. Devlin J, Chang M-W, Lee K, Toutanova K, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pp. 4171–4186.

[16]. Beltagy I, Peters ME, Cohan A, Longformer: The long-document transformer, arXiv preprint arXiv:2004.05150 (2020).

[17]. Lin T, Wang Y, Liu X, Qiu X, A survey of transformers, AI Open (2022).

[18]. Tay Y, Dehghani M, Bahri D, Metzler D, Efficient transformers: A survey, ACM Comput. Surv. 55 (2022).

[19]. Zaheer M, Guruganesh G, Dubey KA, Ainslie J, Alberti C, Ontanon S, Pham P, Ravula A, Wang Q, Yang L, et al. , Big bird: Transformers for longer sequences, Advances in Neural Information Processing Systems 33 (2020) 17283–17297.

[20]. Peng H, Pappas N, Yogatama D, Schwartz R, Smith N, Kong L, Random feature attention, in: International Conference on Learning Representations (ICLR 2021), 2021.

[21]. Katharopoulos A, Vyas A, Pappas N, Fleuret F, Transformers are rnns: Fast autoregressive transformers with linear attention, in: International conference on machine learning, PMLR, 2020, pp. 5156–5165.

[22]. Fedus W, Zoph B, Shazeer N, Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity, The Journal of Machine Learning Research 23 (2022) 5232–5270.

[23]. Zoph B, Designing effective sparse expert models, in: 2022 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW), IEEE Computer Society, 2022, pp. 1044–1044.

[24]. Reid M, Savinov N, Teplyashin D, Lepikhin D, Lillicrap T, Alayrac J.-b., Soricut R, Lazaridou A, Firat O, Schrittwieser J, et al. , Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, arXiv preprint arXiv:2403.05530 (2024).

[25]. Li D, Shao R, Xie A, Sheng Y, Zheng L, Gonzalez J, Stoica I, Ma X, Zhang H, How long can context length of open-source llms truly promise?, in: NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following, 2023.

[26]. Liu NF, Lin K, Hewitt J, Paranjape A, Bevilacqua M, Petroni F, Liang P, Lost in the middle: How language models use long contexts, Transactions of the Association for Computational Linguistics 12 (2024) 157–173.

[27]. Wu C, Wu F, Qi T, Huang Y, Hi-transformer: Hierarchical interactive transformer for efficient and effective long document modeling, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), 2021, pp. 848–853.

[28]. Pappagari R, Zelasko P, Villalba J, Carmiel Y, Dehak N, Hierarchical transformers for long document classification, in: 2019 IEEE automatic speech recognition and understanding workshop (ASRU), IEEE, 2019, pp. 838–844.

[29]. Park H, Vyas Y, Shah K, Efficient classification of long documents using transformers, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2022, pp. 702–709.

[30]. Lewis M, Liu Y, Goyal N, Ghazvininejad M, Mohamed A, Levy O, Stoyanov V, Zettlemoyer L, Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 7871–7880.

[31]. Koh HY, Ju J, Liu M, Pan S, An empirical survey on long document summarization: Datasets, models, and metrics, ACM computing surveys 55 (2022) 1–35.

[32]. Ding M, Zhou C, Yang H, Tang J, Cogltx: Applying bert to long texts, Advances in Neural Information Processing Systems 33 (2020) 12792–12804.

[33]. Fiok K, Karwowski W, Gutierrez-Franco E, Davahli MR, Wilamowski M, Ahram T, Al-Juaid A, Zurada J, Text guide: improving the quality of long text classification by a text selection method based on feature importance, IEEE Access 9 (2021) 105439–105450.

[34]. Hsu C-C, Tan C, Decision-focused summarization, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, pp. 117–132.

[35]. Liu M, Beare R, Collyer T, Andrew N, Srikanth V, Leveraging natural language processing and clinical notes for dementia detection, in: Proceedings of the 5th Clinical Natural Language Processing Workshop, 2023, pp. 150–155.

[36]. Matošević L, Jović A, Accurate detection of dementia from speech transcripts using roberta model, in: 2022 45th Jubilee International Convention on Information, Communication and Electronic Technology (MIPRO), IEEE, 2022, pp. 1478–1484.

[37]. Tang AS, Rankin KP, Cerono G, Miramontes S, Mills H, Roger J, Zeng B, Nelson C, Soman K, Woldemariam S, et al. , Leveraging electronic health records and knowledge networks for alzheimer's disease prediction and sex-specific biological insights, Nature Aging (2024) 1–17. [PubMed: 38177328]

[38]. Bodenreider O, The unified medical language system (umls): integrating biomedical terminology, Nucleic acids research 32 (2004) D267–D270. [PubMed: 14681409]

[39]. Kleiman MJ, Plewes AD, Owora A, Grout RW, Dexter PR, Fowler NR, Galvin JE, Ben-Miled Z, Boustani M, Digital detection of dementia (d3): a study protocol for a pragmatic cluster-

randomized trial examining the application of patient-reported outcomes and passive clinical decision support systems, Trials 23 (2022) 868. [PubMed: 36221141]

[40]. Alsentzer E, Murphy J, Boag W, Weng W-H, Jin D, Naumann T, McDermott M, Publicly available clinical BERT embeddings, in: Proceedings of the 2nd Clinical Natural Language Processing Workshop, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 72–78. URL: https://www.aclweb.org/anthology/W19-1909. doi:10.18653/v1/W19-1909.

[41]. Dai X, Chalkidis I, Darkner S, Elliott D, Revisiting transformer-based models for long document classification, in: Findings of the Association for Computational Linguistics: EMNLP 2022, 2022, pp. 7212–7230.

[42]. Dai Z, Yang Z, Yang Y, Carbonell JG, Le Q, Salakhutdinov R, Transformer-xl: Attentive language models beyond a fixed-length context, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 2978–2988.

[43]. led-large-book-summary, https://huggingface.co/pszemraj/led-large-book-summary, 2023. Accessed: 2023-07-15.

[44]. led-large-pubmed, https://huggingface.co/patrickvonplaten/led-large-16384-pubmed, 2023. Accessed: 2023-07-15.

[45]. Mavrogiorgos K, Kiourtis A, Mavrogiorgou A, Kleftakis S, Kyriazis D, A multi-layer approach for data cleaning in the healthcare domain, in: Proceedings of the 2022 8th International Conference on Computing and Data Engineering, 2022, pp. 22–28.

[46]. Bird S, Klein E, Loper E, Natural language processing with Python: analyzing text with the natural language toolkit, “ O’Reilly Media, Inc.”, 2009.

[47]. Reimers N, Gurevych I, Sentence-bert: Sentence embeddings using siamese bert-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 3982–3992.

[48]. Miles S, Yao L, Meng W, Black CM, Ben Miled Z, Comparing pso-based clustering over contextual vector embeddings to modern topic modeling, Information Processing & Management 59 (2022) 102921.

[49]. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, Cistac P, Rault T, Louf R, Funtowicz M, et al., Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations, 2020, pp. 38–45.

[50]. Keskar NS, McCann B, Varshney LR, Xiong C, Socher R, Ctrl: A conditional transformer language model for controllable generation, arXiv preprint arXiv:1909.05858 (2019).

[51]. Loshchilov I, Hutter F, Decoupled weight decay regularization, arXiv preprint arXiv:1711.05101 (2017).

[52]. Lin C-Y, Rouge: A package for automatic evaluation of summaries, in: Text summarization branches out, 2004, pp. 74–81.

[53]. Mandrekar JN, Receiver operating characteristic curve in diagnostic test assessment, Journal of Thoracic Oncology 5 (2010) 1315–1316. [PubMed: 20736804]

Highlights

- Medical notes document the medical history of the patient.
- The collection of medical notes exceeds the computational limits of language models.
- A decision-focused summary extracts content relevant to dementia.
- Dementia risk prediction is accomplished using decision-focused summaries.

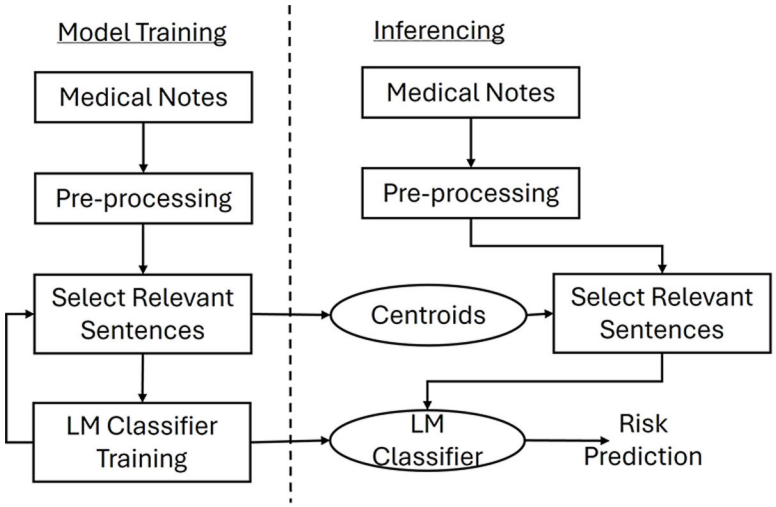


Figure 1:
Training and inferencing processes involved in the meta-algorithm for dementia risk prediction.

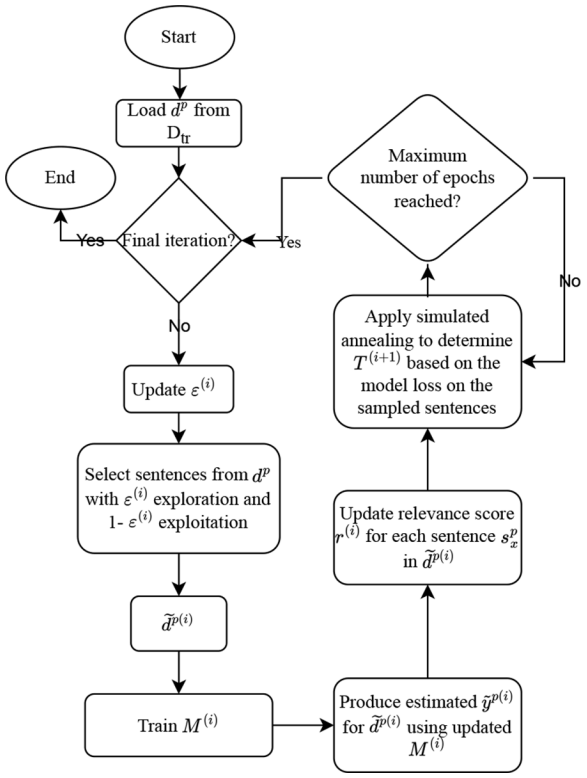


Figure 2:
Iterative sentence selection and classifier algorithm.

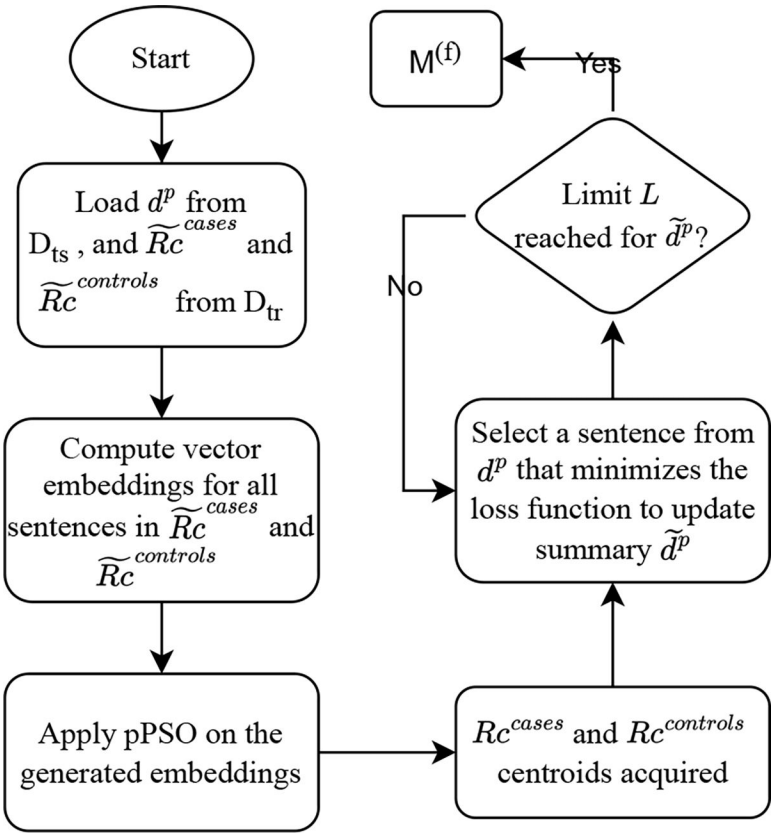


Figure 3:
The workflow for generating decision-focused summaries from the medical notes of new patients.

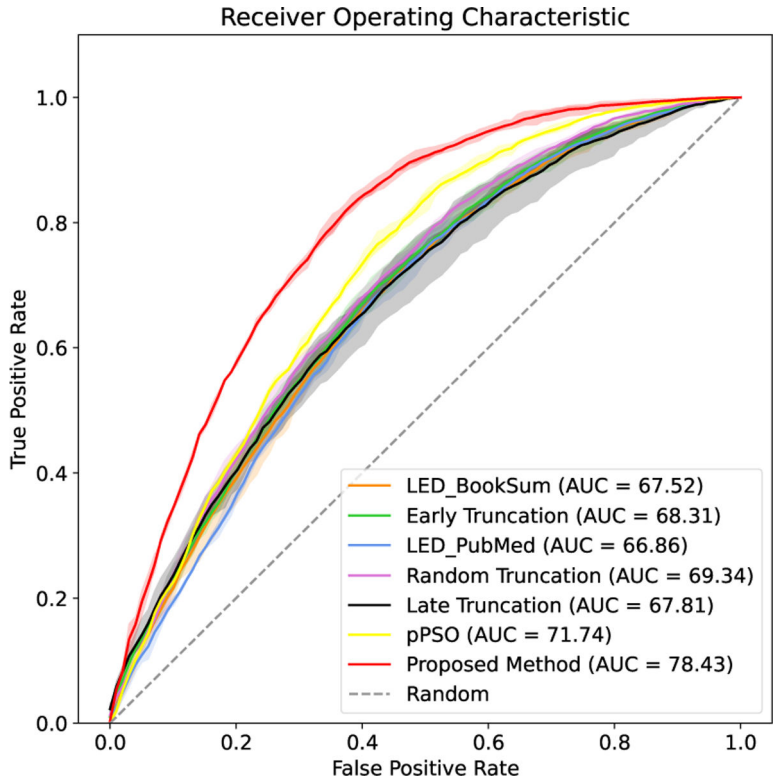


Figure 4:

AUC curves for the classifiers that use the benchmark and proposed decision-focused content selection methodologies with a limit $L = 1024$ on the number of input tokens, an observation period = 1 year and a prediction horizon = 1 year. The standard deviation in the curves is obtained from the 3-fold cross validation.

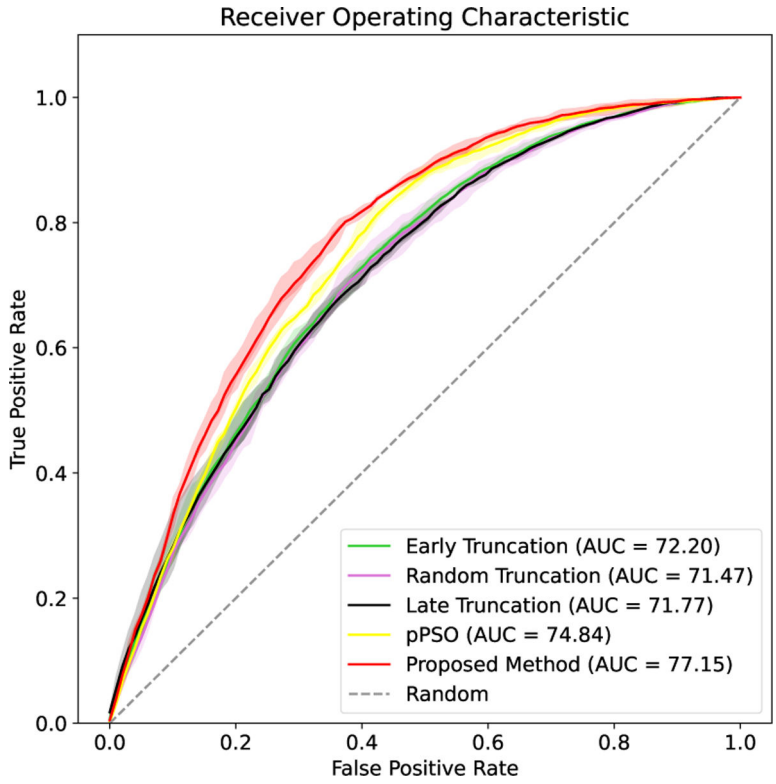


Figure 5:

AUC curves for the classifiers that use the benchmark and proposed decision-focused content selection methodologies with a limit $L = 4096$ on the number of input tokens, an observation period = 1 year and a prediction horizon = 1 year. The standard deviation in the curves is obtained from the 3-fold cross validation.

Table 1

Definitions of variables and parameters used in the proposed decision-focused sentence selection algorithm.

Variable	Definition
d^p	$d^p = \{s_1^p, s_2^p \cdots, s_{n^p}^p\}$ where s_x^p refers to sentence x and n^p is the total number of sentences in d^p collected over the observation period for a given patient p.
y^p	True disease outcome (i.e., 1 for cases or 0 for controls) for a patient p.
Dtr, Dts	The set of documents d^p or patients used for training and testing the classifier model, respectively.
$S^p(s_x^p, s_y^p)$	A matrix S where each entry represents the cosine similarities between the embeddings can the sentences s_x and s_y calculated using sentence-BERT [47]. This matrix can be pre-computed for all pairs on sentences in d^p or computed on demand.
$\tilde{d}^{p(i)}$	A set of sentences selected from d^p following decision-focused criteria and used to train a classification model in iteration i.
$M^{(i)}$	A classifier model which is trained with the content-focused sentences in $\tilde{d}^{p(i)}$.
L	The maximum number of tokens allowed by $M^{(i)}$.
$\tilde{y}^{p(i)}$	Disease risk prediction for patient p which is generated by model $M^{(i)}$ in iteration i.
$r^{(i)}(s_x^p)$	The normalized relevance score of s_x^p to the outcome y^p.
$T^{(i)}$	The threshold for $r^{(i)}(s_x^p)$ used to select sentences.
$\epsilon^{(i)}$	The probability of selecting a sentence from d^p at random. Conversely, $1 - \epsilon^{(i)}$ represents the probability of selecting a sentence from d^p with $r^{(i)}(s_x^p) \geq T^{(i)}$.

Table 2

Number and summary statistics of cases and control patients.

	Cases	Controls
Number of patients	**4796**	**4796**
Mean Age (yrs)	74.43	72.80
Female/Male (%)	63/37	64/36
White/Non-White (%)	67/33	70/30

Table 3

Average (standard deviation) number of tokens and sentences in d^p for the training and testing datasets when the observation period = 1 year. The last two rows show the percentages of d^p with more than $L = 1024$ and $L = 4096$ tokens in each dataset.

	Dtr	*Dts*
Number of patients	**6139**	**1918**
Tokens	5543.2	5395.0
(Std)	(9049.1)	(8849.8)
Sentences	230.0	218.7
(Std)	(299.0)	(322.0)
d^p > 1024 tokens (%)	73.8	73.4
d^p > 4096 tokens (%)	39.9	37.4

Table 4

Values of important parameters for LED_BookSum and LED_PubMed that are used during the generation of the summaries.

	LED_BookSum/LED_PubMed
num_beams	4
repetition_penalty	1.5
no_repeat_ngram_size	3
length_penalty	1.0
max_length	min(input token length, 1024)
min_length	min(input token length*0.75, 1024)

Table 5

Mean AUC, specificity, sensitivity, F1 score and accuracy for the classifiers that use the different content selection techniques with L = 1024 tokens, an observation period = 1 year and a prediction horizon = 1 year. Mean and standard deviation are obtained from the 3-fold cross validation.

	AUC	Precision	Sensitivity	Specificity	F1	Accuracy
Late Truncation	67.81 (2.74)	62.97 (2.08)	62.90 (2.06)	62.90 (2.06)	62.86 (2.06)	62.91 (2.07)
Early Truncation	68.31 (1.50)	63.34 (1.15)	63.30 (1.14)	63.30 (1.14)	63.27 (1.15)	63.30 (1.14)
Random Truncation	69.34 (0.55)	63.83 (0.20)	63.79 (0.21)	63.79 (0.21)	63.77 (0.21)	63.79 (0.21)
LED_BookSum	67.52 (1.15)	62.47 (0.28)	62.44 (0.29)	62.44 (0.29)	62.42 (0.30)	62.44 (0.29)
LED_PubMed	66.86 (0.58)	62.55 (0.84)	62.53 (0.83)	62.53 (0.83)	62.52 (0.83)	62.53 (0.83)
pPSO	71.74 (0.64)	70.03 (0.89)	64.13 (0.68)	64.13 (0.68)	62.50 (0.65)	64.13 (0.68)
Proposed Method	78.43 (0.60)	73.78 (0.52)	70.64 (0.90)	70.64 (0.90)	70.21 (0.97)	70.64 (0.90)

Table 6

Mean AUC, specificity, sensitivity, F1 score and accuracy for the classifiers that use the different content selection techniques with $L = 4096$ tokens, an observation period = 1 year and a prediction horizon = 1 year. Mean and standard deviation are obtained from the 3-fold cross validation.

	AUC	Precision	Sensitivity	Specificity	F1	Accuracy
Late Truncation	71.77 (1.15)	65.95 (0.81)	65.78 (0.79)	65.78 (0.79)	65.69 (0.80)	65.78 (0.79)
Early Truncation	72.20 (0.82)	66.56 (1.17)	66.36 (1.03)	66.36 (1.03)	66.26 (0.97)	66.36 (1.03)
Random Truncation	71.47 (1.11)	66.17 (1.48)	66.07 (1.44)	66.07 (1.44)	66.01 (1.43)	66.07 (1.44)
pPSO	74.84 (0.64)	71.43 (0.24)	67.69 (0.87)	67.69 (0.87)	67.01 (1.04)	67.69 (0.87)
Proposed Method	77.15 (0.66)	72.38 (0.23)	69.89 (0.94)	69.89 (0.94)	69.57 (1.13)	69.89 (0.94)

Table 7

Mean AUC, specificity, sensitivity, F1 score and accuracy for the classifiers that use the proposed content selection technique with varying observation periods (i.e., 1 and 2 years) and prediction horizons (0.5 and 1 year) when L = 1024 tokens. Mean and standard deviation are obtained from the 3-fold cross validation.

Observation Period (yrs)	Prediction Horizon (yrs)	AUC	Precision	Sensitivity	Specificity	F1	Accuracy
1	0.5	80.63 (1.34)	74.60 (0.68)	72.31 (0.82)	72.31 (0.82)	72.08 (0.83)	72.31 (0.82)
2	0.5	81.26 (1.45)	75.04 (0.64)	73.05 (0.71)	73.05 (0.71)	72.89 (0.72)	73.05 (0.71)
1	1	78.43 (0.60)	73.78 (0.52)	70.64 (0.90)	70.64 (0.90)	70.21 (0.97)	70.64 (0.90)
2	1	79.90 (1.22)	74.34 (0.52)	71.53 (0.60)	71.53 (0.60)	71.20 (0.62)	71.53 (0.60)

Table 8

Mean AUC, specificity, sensitivity, F1 score and accuracy for the classifiers that use the proposed content selection technique with varying observation periods (i.e., 1 and 2 years) and prediction horizons (0.5 and 1 year) when $L = 4096$ tokens. Mean and standard deviation are obtained from the 3-fold cross validation.

Observation Period (yrs)	Prediction Horizon (yrs)	AUC	Precision	Sensitivity	Specificity	F1	Accuracy
1	0.5	80.34 (1.17)	74.49 (0.64)	72.02 (0.67)	72.02 (0.67)	71.76 (0.68)	72.02 (0.67)
2	0.5	81.64 (1.33)	74.53 (0.74)	73.60 (0.77)	73.60 (0.77)	73.59 (0.77)	73.60 (0.77)
1	1	77.15 (0.66)	72.38 (0.23)	69.89 (0.94)	69.89 (0.94)	69.57 (1.13)	69.89 (0.94)
2	1	80.07 (1.25)	74.52 (0.57)	71.85 (0.63)	71.85 (0.63)	71.55 (0.65)	71.85 (0.63)