Natural Language Processing Journal 6 (2024) 100046

Contents lists available at ScienceDirect

Natural Language Processing Journal

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

Context is not key: Detecting Alzheimer's disease with both classical and transformer-based neural language models

Behrad TaghiBeyglou [a,b,*], Frank Rudzicz [c,d,e]

[a] Institute of Biomedical Engineering, University of Toronto, Toronto, Canada
[b] KITE Research Institute, Toronto Rehabilitation Institute – University Health Network, Toronto, Canada
[c] Faculty of Computer Science, Dalhousie University, Halifax, Canada
[d] Department of Computer Science, University of Toronto, Toronto, Canada
[e] Vector Institute for Artificial Intelligence, Toronto, Canada

ARTICLE INFO

Keywords:
Alzheimer's disease
ADReSS
Dementia detection
Word2vec
Feature engineering
Classical models
BERT
GPT

ABSTRACT

Natural language processing (NLP) has exhibited potential in detecting Alzheimer's disease (AD) and related dementias, particularly due to the impact of AD on spontaneous speech. Recent research has emphasized the significance of context-based models, such as Bidirectional Encoder Representations from Transformers (BERT). However, these models often come at the expense of increased complexity and computational requirements, which are not always accessible. In light of these considerations, we propose a straightforward and efficient word2vec-based model for AD detection, and evaluate it on the Alzheimer's Dementia Recognition through Spontaneous Speech (ADReSS) challenge dataset. Additionally, we explore the efficacy of fusing our model with classic linguistic features and compare this to other contextual models by fine-tuning BERT-based and Generative Pre-training Transformer (GPT) sequence classification models. We find that simpler models achieve a remarkable accuracy of 92% in classifying AD cases, along with a root mean square error of 4.21 in estimating Mini-Mental Status Examination (MMSE) scores. Notably, our models outperform all state-of-the-art models in the literature for classifying AD cases and estimating MMSE scores, including contextual language models.

1. Introduction

Alzheimer's disease (AD) is a progressive neurodegenerative disease that impairs cognitive functioning and is increasingly common in our aging society (Luz et al., 2021; Ilias and Askounis, 2022). According to the World Health Organization, approximately 55 million people currently suffer from dementia, with this number expected to surge to 78 million by 2030 and 139 million by 2050 (Ilias and Askounis, 2022). Symptoms of AD include memory decline, disorientation, confusion, and behavioural changes (Geda et al., 2013). Importantly, AD progression can lead to loss of independence, significantly impacting patients, their families, and society as a whole (Pappagari et al., 2021). Given that late-stage AD progression is currently inevitable, early detection of AD through cost-effective and scalable technologies is critical. Currently, AD is most conclusively diagnosed using positron emission tomography (PET) imaging and cerebrospinal fluid exams to measure the concentration of amyloid plaques in the brain by postmortem histology, which is a costly and invasive process (Land et al., 2020). Thus, there is a need for more accessible, non-invasive, and efficient methods of AD diagnosis.

Accessible assessment methods for AD include neuropsychological and cognitive tests such as the Mini-Mental Status Examination (MMSE) (Kurlowicz and Wallace, 1999) and the Montréal Cognitive Assessment (MoCA) (Nasreddine et al., 2003). However, these methods still require active administration by an expert, and their specificity in early-stage diagnosis is questionable. During the course of AD, patients experience a gradual deterioration of cognitive function and accordingly may face a loss of lexical-semantic skills, including anomia, reduced word comprehension, object naming problems, semantic paraphasia, and a reduction in vocabulary and verbal fluency (Mirheidari et al., 2018; Pan et al., 2021; Chen et al., 2021).

Clinical information pertaining to cognition can be extracted from spontaneous speech elicited using picture descriptions (Goodglass et al., 2001). As a result, studies have used speech analysis and machine learning (ML) techniques to differentiate between the speech patterns of healthy individuals and those with cognitive impairments, particularly within datasets comprising semi-structured speech tasks such as picture description (Thomas et al., 2005; König et al., 2015; López-de-Ipiña et al., 2015).

Another line of inquiry explores natural language processing (NLP) and linguistic analyses (Rentoumi et al., 2014; Orimaye et al., 2014; Mirheidari et al., 2016; Fraser et al., 2016; Wankerl et al., 2017; Sadeghian et al., 2017; Weiner et al., 2017). These approaches provide

[*] Corresponding author at: Institute of Biomedical Engineering, University of Toronto, Toronto, Canada.
E-mail address: behrad.taghibeyglou@mail.utoronto.ca (B. TaghiBeyglou).

https://doi.org/10.1016/j.nlp.2023.100046
Received 22 July 2023; Received in revised form 19 October 2023; Accepted 4 December 2023

novel precision tools in AD diagnosis, enabling objective quantitative analyses and reliable evidence for expedited and accurate assessments. However, many of these studies have been conducted on datasets imbalanced in age, sex, or AD status (Orimaye et al., 2014; Mirheidari et al., 2016; Fraser et al., 2016; Wankerl et al., 2017; Sadeghian et al., 2017; Weiner et al., 2017). As a result, it is crucial to establish balanced and standardized datasets in order to facilitate the comparison of different methodologies, as highlighted by de la Fuente Garcia et al. (2020) and Voleti et al. (2019). To address this, the Alzheimer's Dementia Recognition through Spontaneous Speech (ADReSS) challenge dataset was introduced in 2020, providing a balanced dataset, thereby mitigating common biases associated with other AD datasets and enabling direct comparisons among the techniques (Luz et al., 2020).

NLP-based methods have spanned from-scratch training to fine-tuning context-based models, such as Bidirectional Long Short-Term Memory (bi-LSTM) (Cummins et al., 2020), bi-directional Hierarchical Attention Network (bi-HANN) (Cummins et al., 2020), Convolutional Recurrent Neural Network (CRNN) (Koo et al., 2020), Transformer-XL (Koo et al., 2020), and Bidirectional Encoder Representations from Transformer (BERT) (Balagopalan et al., 2020, 2021; Haulcy and Glass, 2021; Syed et al., 2020; Guo et al., 2021; Farzana and Parde, 2020; Taghibeyglou and Rudzicz, 2023), to either identify AD (Yuan et al., 2020), estimate MMSE score (Farzana and Parde, 2020), or both (Balagopalan et al., 2020, 2021; Koo et al., 2020; Martinc et al., 2021; Pappagari et al., 2021; Rohanian et al., 2021; Sarawgi et al., 2020; Syed et al., 2020; Cummins et al., 2020). Despite excellent performance compared to baseline methods (Luz et al., 2020), the complexity of these methodologies and the need to implement them on high-memory GPUs highlights the need to explore simpler methodologies that can ensure ease and performance in AD detection.

In this paper, we introduce a novel model based on a mixture of word2vec embeddings and linguistic features, that is simple, easy to implement, highly accurate, and designed for identifying AD from transcripts. We also estimate MMSE scores using leave-one-subject-out (LOSO) cross validation as well as on the unseen data of the ADReSS dataset. To compare the performance of our model with contextual language models, we fine-tune BERT models and the Generative Pre-trained Transformer (GPT-2). Additionally, we compare our proposed models, including both classification and regression models, with state-of-the-art studies conducted on the same dataset, thus showcasing the effectiveness of our simpler models.

Section 2 provides detailed information on the dataset and all the models implemented and proposed in this study. In Section 3, we present the results of our proposed models, along with those of fine-tuned state-of-the-art models from previous studies. Section 4 discusses the findings, limitations, and suggestions for future research. Finally, we conclude in Section 5.

2. Methods

2.1. Dataset

We use the ADReSS challenge dataset (Luz et al., 2020), which was specifically curated to facilitate research on automatic AD detection using spontaneous speech and associated transcripts. This dataset comprises 156 speech recordings and their corresponding transcripts from English-speaking participants, divided into training and testing subsets.

Participants in this challenge were given the task of verbally describing the Cookie Theft picture from the Boston Diagnostic Aphasia Exam, conducted in English (Goodglass et al., 2001; Guo et al., 2021). The transcripts were annotated using the CHAT coding system (MacWhinney, 2014). To ensure consistent audio quality across recordings and minimize variation caused by recording conditions like microphone placement, the challenge organizers applied acoustic enhancements to the recordings (Luz et al., 2020). These enhancements involved stationary noise removal and audio volume normalization applied uniformly to all speech segments.

Table 1
Participants number in each subset of ADReSS challenge dataset.

Subset	Sex	Class	
		AD	Non-AD
Train	Male	24	24
	Female	30	30
Test	Male	11	11
	Female	13	13

Furthermore, to mitigate the risk of bias in the prediction tasks, age and gender were carefully matched while constructing the dataset, as shown in Table 1. Furthermore, the MMSE scores (Cockrell and Folstein, 2002) were available for all participants (but one) in the training set. More details on the dataset can be found in Table S1 in the "Supplementary Material".

2.2. Transfer learning and fine-tuning

We used different language models originally trained on extensive corpora, including book texts, Wikipedia, clinical records, and other sources. We fine-tuned these models specifically for AD detection, which enables us to compare our proposed non-contextualized model with state-of-the-art but more computationally expensive (and possibly less clinically interpretable) contextualized models. In this regard, we investigate two distinct language model architectures: BERTs and GPT-2.

2.2.1. BERTs

The Transformer proposed by Vaswani et al. (2017) is a model architecture that relies on an attention mechanism to draw global dependencies between input and output. This is achieved through the encoder and decoder components. Encoders are used to focus on the position of the words relative to other neighbouring words, and the more encoder layers are used, the more contexts the model can focus on. As a result, BERTs, which consist of several (12 or 24) transformer encoder units, can deeply learn the context of the input sentences. In this study, four BERT models have been used:

- **uncased base BERT**: This model has 12 transformer layers, each with a hidden size of 768 and 12 self-attention heads, resulting in 110M parameters (Devlin et al., 2018). This model was pre-trained on *BooksCorpus* (800M words) (Zhu et al., 2015) and *English Wikipedia* (2.5B words). In this paper, the uncased base BERT model was used in three distinct network architectures. The model with no network adjustment is called *baseBERT1*, the model with two fully connected layers added in the last layer (768 → 64 and 64 → 1) is called *baseBERT2*, and the model with three fully connected layers (768 → $128, 128$ → 64, and 64 → 1) is called *baseBERT3*.
- **Bio-Clinical BERT** (Alsentzer et al., 2019): This model is a fine-tuned version of BioBERT (Lee et al., 2020), which was originally using the same uncased base BERT architecture and was trained with two additional sources of corpora: PubMed abstracts[1] and PMC full-text articles.[2] Bio-Clinical BERT model, introduced in Alsentzer et al. (2019), was further pre-trained using the MIMIC-III v1.4 database notes (Johnson et al., 2016). In this paper, we fine-tuned this model for the AD detection task without any further refinements in the network architecture, and refer to it as *Bio-Clinical BERT* throughout the paper.

[1] https://www.ncbi.nlm.nih.gov/pubmed/.

[2] https://www.ncbi.nlm.nih.gov/pmc/.

- **DistilBERT** (Sanh et al., 2019): This model used a knowledge distillation technique to imitate the function of uncased base BERT with a student architecture using only two transformer blocks/layers. This approach significantly reduces the number of trainable parameters to 66M while retaining 97% of the performance of the uncased base BERT. In this paper, we fine-tune the model for the AD classification task and refer to it as *DistilBERT*.
- **BioMed-RoBERTa-based** (Gururangan et al., 2020): The RoBERTa base model is essentially the uncased base BERT model with some fine refinements regarding the optimization procedure and incorporating three more large corpora (Liu et al., 2019). The fine-tuned model used in this study, called *BioMed-RoBERTa-based*, pre-trains the RoBERTa base model on 2.68 million scientific papers from the Semantic Scholar[3] corpus. In this paper, we fine-tune the model for AD detection without making any further modifications to the network architecture.

For both the classification and regression schemes, sentences from each participant's transcript are tokenized and used as input. The output is either 1 for AD and 0 otherwise in the classification task, or the MMSE score for the regression task. The final hidden state corresponding to the first special token ([CLS]) in the transcript, which summarizes the information across all tokens using the self-attention mechanism in BERT, is used as the aggregate representation and passed to the last layer. For the classification task, the last layer is accompanied by a sigmoid activation function, while for the regression task, the output is coupled with a rectified linear unit (ReLU) activation function. The aggregation of all sentences is computed as the average of the output values. A threshold of 0.5 is used as the boundary for the AD or Non-AD class in the classification task, while the arithmetic average is used for the regression task.

2.2.2. GPT-2

GPT-2, in contrast to BERTs, consists of transformer decoder blocks that allow the architecture to use the previous token of the input sequence to predict the next token that should follow Radford et al. (2019). In this study, we used 'GPT-2 small', which has 12 decoder layers, each with a hidden size of 768 and 12 self-attention heads, and accordingly has 117M parameters. This model is denoted as *GPT-2* throughout the paper. Similar to the refinements described for BERTs, we made the same modifications for the *GPT-2* model. However, instead of extracting embeddings for the first token, we use the embedding of the last token to perform AD detection or MMSE regression.

2.2.3. Training specifications

All pre-trained weights are based on sequence classification models. For classification, we use binary cross-entropy loss, while for regression models, we use Huber loss, which is a mixture of L1-loss and L2-loss and is useful for mitigating both vanishing and exploding gradient issues. Let N_b be the number of training samples in the batch size, and y_i and $\hat{y}_i$ be the original and predicted MMSE values corresponding to the ith sentence in the batch, respectively. Assuming $l(\mathbf{y}, \hat{\mathbf{y}}) = l_1, \dots, l_i, \dots, l_N$ is the loss vector for all samples in the batch, l_i is defined as:

$$l_i = \begin{cases} 0.5 \cdot (y_i - \hat{y}_i)^2, & \text{if } |y_i - \hat{y}| < \delta \\ \delta \cdot (|y_i - \hat{y}i| - 0.5 \cdot \delta), & \text{otherwise} \end{cases} \tag{1}$$

where δ is the regularization factor, set as 1 in this study. The overall loss value for the batch is calculated as $1/N \cdot \sum_{i=1}^{N} l_i$. The optimizer we used is AdamW (Adam with weight decay) (Loshchilov and Hutter, 2017) with a learning rate of $2 \cdot 10^{-5}$, β_1 of 0.9, β_2 of 0.999, and weight decay λ of 0.01. Furthermore, we applied a linear warm-up scheduler to address potential local optima issues.

For tuning training hyperparameters, such as the number of epochs and batch size, we applied a grid search using batch sizes of 1, 2, 4, 8, and 16, and epoch numbers of 1, 2, 3, and 5. A low number of epochs was selected to minimize the impact on the pre-trained models while transferring them to the task of AD detection or MMSE regression. The combination of 3 epochs and a batch size of 4 provided the best performance among all models. Additional information regarding the number of trainable parameters for each model is available in Table S3 of the "Supplementary Material".

All fine-tuning procedures were implemented using PyTorch on a computer equipped with an AMD Ryzen 9 5900X 12-Core Processor and a GeForce RTX 3080 graphics card. Additionally, all pre-trained weights were obtained from the HuggingFace transformers repository (Wolf et al., 2020).

2.3. Proposed model

Our proposed model includes two submodels: only word2vec-based model (called $model_{W2V}$), and the concatenation of word2vec embeddings and linguistic-based features, which makes $model_{W2V+LBF}$.

2.3.1. Preprocessing

For data preprocessing, we excluded the first four sentences of each transcript, as they typically involve a member of the data collection team. Furthermore, stop words were removed from each sentence using the Gensim library (Řehůřek et al., 2011).

2.3.2. word2vec-based component

To capture the underlying semantic representation of the words present in each transcript, we employed Wikipedia2Vec (Yamada et al., 2018). This model is a skip-gram-based word2vec model that converts words and entities (tokens) from Wikipedia corpora into vector embeddings of dimension 500. In this paper, we refer to the Wikipedia2Vec model as *W2V*. All preprocessed tokens in each participant's transcript are used as inputs to the *W2V* model, resulting in a set of embeddings $\mathbf{X}_k = \{\mathbf{x}_{1,1}^{(k)}, \dots, \mathbf{x}_{1,J_{k_1}}^{(k)}, \dots, \mathbf{x}_{I_k,1}^{(k)}, \dots, \mathbf{x}_{I_k,J_{k_I}}^{(k)}\}$, where k denotes the participant number, I_k is the total number of sentences for the kth participant, J_{k_i} denotes the total number of words in the ith sentence of the kth participant, and $\mathbf{x}_{i,j}^{(k)} \in \mathbb{R}^{500}$ represents the embedding corresponding to the ith word of the jth sentence of the kth participant.

Next, the embeddings in the set are used to obtain a concentrated vector representation using the following equation:

$$\mathbf{y}_k = \frac{\mathtt{med}(\mathbf{X}_k)}{\mathtt{std}(\mathbf{X}_k)}, \tag{2}$$

where `med` is the arithmetic median operator applied independently to each embedding dimension, `std` represents the standard deviation of the embeddings, and $\mathbf{y}_k$ denotes the standardized concentrated vector representation of the transcript for the kth participant.

The entire procedure is illustrated in Algorithm 1, where W_k denotes the set of all tokens in the preprocessed transcript of participant k, specifically $W_k = \{\langle w_{1,1}^{(k)}\rangle, \dots, \langle w_{1,J_{k_1}}^{(k)}\rangle, \dots, \langle w_{I_k,1}^{(k)}\rangle, \dots, \langle w_{I_k,J_{k_I}}^{(k)}\rangle\}$, $\langle w_{i,j}^{(k)}\rangle$ represents the token corresponding to the ith word of the jth sentence of the kth participant, and "[,]" denotes concatenation.

2.3.3. Linguistic-based features

In this study, we employed the `eval` command within the CLAN package (MacWhinney, 2014, 2017b,a) to extract a total of 34 linguistic-based features (LBFs) from the transcripts. These features encompass various aspects such as duration, total utterances, mean length of utterance (MLU), type-token ratio, and open-closed class word ratio. Table S1 (in Supplementary Material) presents a comprehensive list of these features, including their common names and a brief description. Additionally, we integrated demographic information such as age and sex into our analysis.

[3] https://www.semanticscholar.org/.

Table 2
Details of the classifiers used in the proposed method.

Classifier	Hyperparamter(s)
LR	penalty: ‘L2’, C: 1.0, max_iter: 100
DT	criterion: ‘gini’, splitter: ‘best’, min_samples_split: 2, min_samples_leaf: 1
Linear SVC	penalty: ‘L2’, loss: ‘squared_hinge’, C: 1.0, max_iter: 1000
Nu-SVC	nu: 0.5, kernel: ‘RBF’, degree: 3
LDA	solver: ‘SVD’
QDA	reg_param: 0.0
GNB	N/A
XGBC	learning_rate: 0.1, n_estimators: 100, max_depth: 3, min_child_weight: 1, subsample: 1.0, colsample_bytree: 1.0
AdaC	n_estimators: 50, learning_rate: 1.0
XTs	n_estimators: 100, criterion: ‘gini’, min_samples_split: 2, min_samples_leaf: 1

Algorithm 1 Extraction of Concentrated Vector Representation for the k^{th} Transcript

```
Require: Set of tokens for the $k^{th}$ participant ($W_k$).
Ensure: $\mathbf{y}_k$
  $\mathbf{X}_k = [\ \ ]$
  for $i \leftarrow 1$ to $I_k$ do
    for $j \leftarrow 1$ to $J_{k_i}$ do
      $\mathbf{x}_{tmp} \leftarrow W2V(\langle w_{i,j}^{(k)} \rangle)$
      $\mathbf{X}_k \leftarrow [\mathbf{X}_k, \mathbf{x}_{tmp}]$
    end for
  end for
  $\mathbf{y}_k \leftarrow \mathtt{med}(\mathbf{X}_k)/\mathtt{std}(\mathbf{X}_k)$    ▷ The med
  and std functions are applied to each dimension of the embeddings
  separately, and the “/” operator represents element-wise division.
  return $\mathbf{y}_k$
```

2.3.4. Feature selection

To identify the most informative features for both classification and regression tasks, we conducted correlation and variance analyses using the FeatureWiz package (AutoViML, 2020) based on the minimum redundancy maximum relevance (MRMR) method (Peng et al., 2005). We set a correlation threshold of 0.4 and repeated the analyses 5 times with different random seeds across all samples. From these analyses, we selected the features that appeared in at least 3 iterations as the most relevant for further model development. This selection process was performed for each model, either $model_{W2V}$ or $model_{W2V+LBF}$, independently. Additional information regarding the top-5 selected features for each task is available in Table S4 of the “Supplementary Material”.

2.3.5. Feature standardization

After selecting the optimal feature set, each feature is normalized using zero-mean unit-variance normalization to mitigate the impact of varying feature characteristics across different dimensions of the feature set. The normalization parameters, such as the mean and standard deviation of each dimension, are calculated solely using the training set. This ensures that none of the training parameters are utilized in evaluating the models.

2.3.6. Models

Classification: For the classification task, we employed classical machine learning models using the selected features from the previous models. Specifically, we evaluated the following classifiers: logistic regression (LR), decision tree (DT), linear Support Vector Classifier (Linear SVC), Nu-Support Vector Classifier (Nu-SVC), Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Gaussian Naive Bayes (GNB), eXtreme Gradient Boosting Classifier (XGBC), Adaptive Boosting Classifier (AdaC), and Extra Trees Classifier (XTs). More details about each model's hyperparameter(s) are mentioned in Table 2.

Regression: For regression models, we evaluate the Support Vector Regressor (SVR), Linear Regression Model (LRM), Ridge, least absolute shrinkage and selection operator (Lasso), Adaboost Regressor (AdaR), Stochastic Gradient Descent Regressor (SGDR), Random Forest Regressor (RFR), Extra Trees Regressor (XTsR) and XGB Regressor (XGBR). The details of each model are mentioned in Table 3.

All classification and regression models are implemented in Python using Scikit-learn library (Pedregosa et al., 2011; Buitinck et al., 2013).

2.4. Evaluation

2.4.1. Cross-validation and testing set

In the field of machine learning, various cross-validation methods are employed for model development, including K-fold, hold-out, and leave-one-subject-out (LOSO) techniques. While the former two approaches are commonly used in a wide range of machine learning papers, they often exhibit poor generalizability in healthcare applications. This limitation arises from the fact that samples from different classes of the same subject can appear across the train, test, and/or validation sets. We therefore chose to use the LOSO approach. By implementing LOSO, we allocate the features corresponding to the sentences of one participant as the validation set, while employing the remaining data for training the model.

In this paper, we present the performance of the proposed models using two distinct evaluations. First, we utilize the LOSO cross-validation technique on the training set to assess model performance during the development phase. This involves iteratively training the models on subsets of participants, leaving one participant out as the validation set each time. Second, we employ a separate evaluation using the testing set. The models developed during the training phase are evaluated on this independent set of participants to gauge their performance in real-world scenarios. By conducting evaluations on both the training set (via LOSO cross-validation) and the testing set, we provide a comprehensive assessment of the proposed models' effectiveness and generalizability.

2.4.2. Metrics

Classification: To evaluate the classification models, we used accuracy (AC), sensitivity (SE), specificity (SP), and harmonic score (HS).

Regression: In regression models, we will report several metrics commonly used in the literature, including the mean absolute error (MAE), mean absolute percentage error (MAPE), and root mean square error (RMSE). Since the MMSE scores, which are the target for regression models, reflect the degree of progression of AD in individuals with AD, we will also report correlation-based metrics such as Pearson's correlation coefficient (PCC), Spearman's correlation coefficient (SCC) (Hauke and Kossowski, 2011), and Kendall's correlation coefficient (KCC) (Abdi, 2007). While PCC measures the linearity between the predicted values and the original values, SCC and KCC are non-parametric tests that assess the monotonic relationship using ranked data. Given the significance of capturing the trend of estimation in MMSE regression, SCC and KCC are particularly important measures.

Table 3
Details of the regression models used in the proposed method.

Regressor	Hyperparamter(s)
SVR	kernel: ‘RBF’, C: 1.0, ϵ: 0.1
LRM	N/A
Ridge	α: 1.0, solver: ‘Auto’
Lasso	α: 1.0, max_iter: 1000
AdaR	n_estimators: 50, learning_rate: 1.0
SGDR	loss = ‘squared_error’, penalty = ‘L2’, α = 0.0001, max_iter: 1000, tol: 1e−3
RFR	n_estimators: 100, criterion: ‘MSE’, min_samples_split: 2, min_samples_leaf: 1
XTsR	n_estimators: 100, criterion: ‘MSE’, min_samples_split: 2, min_samples_leaf: 1
XGBR	learning_rate: 0.1, n_estimators: 100, max_depth: 3, min_child_weight: 1, subsample: 1.0, colsample_bytree: 1.0

Table 4
LOSO performance of the pre-trained models. AC, SP, SE, and HS represent the accuracy, specificity, sensitivity, and harmonic score, respectively.

Model	AC↑	SP↑	SE↑	HS↑
baseBERT1	0.80	0.89	0.7	0.78
baseBERT2	0.79	0.81	0.76	0.78
baseBERT3	0.78	**0.90**	0.67	0.77
Bio-Clinical BERT	**0.84**	0.85	0.83	**0.84**
DistilBERT	0.84	0.87	0.81	0.84
BioMed-RoBERTa-based	0.81	0.87	0.76	0.81
GPT-2	0.79	0.72	**0.85**	0.78

Table 5
The performance of the pre-trained models on the test set. AC, SP, SE, and HS represent the accuracy, specificity, sensitivity, and harmonic score, respectively.

Model	AC↑	SP↑	SE↑	HS↑
baseBERT1	0.77	**0.96**	0.58	0.72
baseBERT2	0.81	0.92	0.71	0.80
baseBERT3	0.80	0.89	0.70	0.78
Bio-Clinical BERT	**0.87**	0.92	0.83	**0.87**
DistilBERT	0.83	0.83	0.83	0.83
BioMed-RoBERTa-based	0.77	0.79	0.75	0.77
GPT-2	0.79	0.71	**0.87**	0.78

Table 6
LOSO performance of the proposed models ($model_{W2V}$: only word2vec-based concentrated embeddings, $model_{W2V+LBF}$: combination of word2vec-based concentrated with linguistic-based features). LR: logistic regression, DT: decision tree, SVC: support vector classifier, LDA: linear discriminant analysis, QDA: quadratic discriminant analysis, GNB: Gaussian naive Bayes, AdaC: Adaboost classifier, XTs: extra trees, XGBC: extreme gradient boosting classifier, AC: accuracy, SP: specificity, SE: sensitivity, HS: harmonic score.

Classifier	Model	AC↑	SP↑	SE↑	HS↑
LR	$model_{W2V}$	0.74	0.87	0.81	0.84
	$model_{W2V+LBF}$	0.74	0.89	0.80	0.84
DT	$model_{W2V}$	0.76	0.80	0.72	0.76
	$model_{W2V+LBF}$	0.56	0.54	0.57	0.55
Linear SVC	$model_{W2V}$	0.81	0.85	0.78	0.81
	$model_{W2V+LBF}$	0.80	0.85	0.74	0.79
Nu-SVC	$model_{W2V}$	0.85	0.85	0.85	0.85
	$\boldsymbol{model_{W2V+LBF}}$	**0.90**	**0.91**	0.89	**0.9**
LDA	$model_{W2V}$	0.73	0.74	0.72	0.73
	$model_{W2V+LBF}$	0.66	0.69	0.63	0.66
QDA	$model_{W2V}$	0.60	0.63	0.57	0.6
	$model_{W2V+LBF}$	0.44	0.33	0.56	0.42
GNB	$model_{W2V}$	0.87	0.87	0.87	0.87
	$\boldsymbol{model_{W2V+LBF}}$	**0.90**	0.89	**0.91**	**0.9**
XGBC	$model_{W2V}$	0.77	0.76	0.78	0.77
	$model_{W2V+LBF}$	0.78	0.78	0.78	0.78
AdaC	$model_{W2V}$	0.81	0.78	0.85	0.81
	$model_{W2V+LBF}$	0.82	0.81	0.83	0.82
XTs	$model_{W2V}$	0.88	0.89	0.87	0.88
	$model_{W2V+LBF}$	0.89	0.91	0.87	0.89

More information about the metrics used in this study can be found in “Supplementary Material”. All the metrics are implemented using the Scikit-learn (Pedregosa et al., 2011) and SciPy Statistics libraries (Virtanen et al., 2020) in Python.

3. Results

3.1. Classification

3.1.1. Transfer learning and fine-tuning

The performance of the fine-tuned models is presented in Tables 4 and 5 for the LOSO evaluation (on the training set) and the test set, respectively. Further investigation of Table 4 reveals that both *Bio-Clinical BERT* and *DistilBERT* demonstrate similar AC and HS. However, considering the crucial importance of sensitivity in AD detection, we select the *Bio-Clinical BERT* model as the best performer overall. This selection is further validated in Table 5, where *Bio-Clinical BERT* exhibits superior classification performance on the unseen test data.

3.1.2. Proposed models

The results of our proposed models for the classification task are presented in Tables 6 and 7, corresponding to the leave-one-subject-out (LOSO) evaluation on the training set and the test set, respectively. From Table 6, it can be observed that the best performance is achieved by $model_{W2V+LBF}$ using Nu-SVC and GNB classifiers. As mentioned in the previous section, considering the significance of sensitivity, the GNB model is selected as the best model for the classification task. Notably, this model achieves a sensitivity of 1, indicating its capability to correctly identify AD cases without any false negatives. Tables 6 and 7 also indicate that the combination of LBF with our proposed word2vec-based technique enhances the performance of $model_{W2V}$, which solely relies on the proposed word2vec embeddings, in almost all classifiers.

3.2. Regression

3.2.1. Transfer learning and fine-tuning

The results of regression using the pre-trained models are presented in Table 8. Among all the models, *DistilBERT* demonstrates the best performance, exhibiting relatively strong correlation indices between the predicted and original MMSE scores. However, it should be noted that the overall performance of the fine-tuned models is relatively unsatisfactory. As a result, the performance on the test set is not further evaluated.

3.2.2. Proposed models

The results of our proposed models alongside each regression are presented in Tables 9 and 10 for the LOSO evaluation and the test set, respectively. Table 9 shows that the models trained with AdaR exhibit the best performance, characterized by low MAE and high MAPE. Assuming the AdaR models are selected based on their performance on the train set, the results indicate that $model_{W2V+LBF}$ demonstrates better performance compared to $model_{W2V}$. However, it is noteworthy that the Lasso models outperform the AdaR models on the test data. Another important observation from Tables 9 and 10 is that the performance of almost all models improves on the test set.

Table 7
The performance of the proposed models ($model_{W2V}$: only word2vec-based concentrated embeddings, $model_{W2V+LBF}$: combination of word2vec-based concentrated with linguistic-based features) on the test set. LR: logistic regression, DT: decision tree, SVC: support vector classifier, LDA: linear discriminant analysis, QDA: quadratic discriminant analysis, GNB: Gaussian naive Bayes, AdaC: Adaboost classifier, XTs: Extra Trees, XGBC: extreme gradient boosting classifier, AC: accuracy, SP: specificity, SE: sensitivity, HS: harmonic score.

Classifier	Model	AC↑	SP↑	SE↑	HS↑
LR	$model_{W2V}$	0.88	0.83	0.92	0.87
	$model_{W2V+LBF}$	0.88	0.83	0.92	0.87
DT	$model_{W2V}$	0.54	0.67	0.42	0.51
	$model_{W2V+LBF}$	0.54	0.38	0.71	0.49
Linear SVC	$model_{W2V}$	0.83	0.79	0.88	0.83
	$model_{W2V+LBF}$	0.79	0.75	0.83	0.79
Nu-SVC	$model_{W2V}$	0.83	0.83	0.83	0.83
	$\boldsymbol{model_{W2V+LBF}}$	**0.92**	0.88	0.96	**0.91**
LDA	$model_{W2V}$	0.56	0.71	0.42	0.52
	$model_{W2V+LBF}$	0.79	0.71	0.88	0.78
QDA	$model_{W2V}$	0.58	0.58	0.58	0.58
	$model_{W2V+LBF}$	0.52	0.50	0.54	0.52
GNB	$model_{W2V}$	0.81	0.83	0.79	0.81
	$\boldsymbol{model_{W2V+LBF}}$	**0.92**	0.83	**1.00**	**0.91**
XGBC	$model_{W2V}$	0.77	0.83	0.71	0.77
	$model_{W2V+LBF}$	0.77	0.71	0.83	0.77
AdaC	$model_{W2V}$	0.77	0.79	0.75	0.77
	$model_{W2V+LBF}$	0.79	0.67	0.92	0.77
XTs	$model_{W2V}$	0.83	**0.92**	0.75	0.83
	$model_{W2V+LBF}$	0.85	0.75	0.96	0.84

Table 8
LOSO performance of the pre-trained models in predicting MMSE score. MAE: mean absolute error, MAPE: mean absolute percentage error, RMSE: root mean square error, PCC: Pearson's correlation coefficient, KCC: Kendall's correlation coefficient, SCC: Spearman's correlation coefficient.

Model	MAE↓	MAPE↓	RMSE↓	PCC↑	KCC↑	SCC↑
baseBERT1	6.68	0.57	7.48	−0.53*	−0.4*	−0.48*
baseBERT2	6.90	0.54	7.61	−0.61*	−0.48*	−0.55*
baseBERT3	7.14	0.54	7.73	−0.75*	−0.61*	−0.71*
Bio-Clinical BERT	7.05	0.50	7.79	0.23	0.12	0.14
DistilBERT	**4.94**	**0.41**	**5.57**	**0.73***	**0.57***	**0.71***
BioMed-RoBERTa-based	7.25	0.51	8.29	−0.1	−0.15	−0.18
GPT-2	7.45	0.61	9.06	−0.06	−0.12	−0.16

* P-value < 0.05 showing statistically significant correlation.

In order to investigate the prediction error of the AdaR model, we use the Bland-Altman plot (Bland and Altman, 1986) in Fig. 1. Notably, $model_{W2V}$ achieves higher performance on LOSO validation, whereas $model_{W2V+LBF}$ exhibits superior performance on the test/unseen data. Additionally, the Bland-Altman plot highlights that the error standard deviation for $model_{W2V+LBF}$ is significantly low on the test data, suggesting that the predictions not only closely align with the original values but also display minimal bias.

3.3. Comparison with previous studies

To further assess the ranking of our proposed model compared to state-of-the-art studies on the same dataset, we compared our methodology with some of the best-performing models from previous literature. The results of this comparison are presented in Table 11, highlighting the potential of our model, given its simplicity and interpretability.

4. Discussion

In this study, we evaluate a non-contextual word2vec-based model for detecting AD and estimating the corresponding MMSE scores. We also explore fine-tuning and transfer learning of pre-trained BERT models and GPT-2 to effectively compare our non-contextual model with contextual models.

Table 9
LOSO performance of our proposed models in predicting MMSE score. SVR: support vector regression, LRM: linear regression model, Lasso: least absolute shrinkage and selection operator, AdaR: Adaboost regressor, SGDR: stochastic gradient descent regressor, RFR: random forest regressor, XTsR: extra trees regressor, XGBR: extreme gradient boosting regressor, MAE: mean absolute error, MAPE: mean absolute percentage error, RMSE: root mean square error, PCC: Pearson's correlation coefficient, KCC: Kendall's correlation coefficient, SCC: Spearman's correlation coefficient.

Regressor	Model	MAE↓	MAPE↓	RMSE↓	PCC↑	KCC↑	SCC↑
SVR	$model_{W2V}$	5.30	0.54	7.03	0.86*	**0.83***	**0.93***
	$model_{W2V+LBF}$	5.24	0.54	6.97	0.87*	**0.83***	**0.93***
LRM	$model_{W2V}$	18.54	1.08	23.08	0.16	0.08	0.12
	$model_{W2V+LBF}$	6.05	0.36	7.61	0.85*	0.64*	0.80*
Ridge	$model_{W2V}$	6.74	0.44	8.69	0.77*	0.58*	0.76*
	$model_{W2V+LBF}$	5.67	0.35	7.11	0.86*	0.64*	0.80*
Lasso	$model_{W2V}$	4.57	0.44	5.63	0.89*	0.74*	0.89*
	$model_{W2V+LBF}$	4.33	0.37	5.15	0.91*	0.75*	0.89*
AdaR	$\boldsymbol{model_{W2V}}$	**3.91**	0.32	**4.79**	**0.92***	0.79*	0.91*
	$\boldsymbol{model_{W2V+LBF}}$	**4.09**	0.37	5.12	0.90*	0.73*	0.87*
SGDR	$model_{W2V}$	5.84	0.39	7.58	0.82*	0.62*	0.78*
	$model_{W2V+LBF}$	5.44	0.34	6.83	0.87*	0.66*	0.81*
RFR	$model_{W2V}$	4.29	0.36	5.17	0.91*	0.77*	0.91*
	$model_{W2V+LBF}$	4.31	0.37	5.15	0.92*	0.75*	0.89*
XTsR	$model_{W2V}$	4.19	0.36	**4.79**	0.92*	0.80*	0.92*
	$model_{W2V+LBF}$	4.26	0.35	5.04	**0.93***	0.76*	0.90*
XGBR	$model_{W2V}$	4.14	**0.30**	5.37	0.89*	0.74*	0.88*
	$model_{W2V+LBF}$	4.15	**0.32**	5.31	0.89*	0.71*	0.86*

* P-value < 0.05 showing statistically significant correlation.

Table 10
The performance of our proposed models in predicting MMSE score on the test data. SVR: support vector regression, Lasso: least absolute shrinkage and selection operator, LRM: linear regression model, AdaR: Adaboost regressor, SGDR: stochastic gradient descent regressor, RFR: random forest regressor, XTsR: extra trees regressor, XGBR: extreme gradient boosting regressor, MAE: mean absolute error, MAPE: mean absolute percentage error, RMSE: root mean square error, PCC: Pearson's correlation coefficient, KCC: Kendall's correlation coefficient, SCC: Spearman's correlation coefficient.

Regressor	Model	MAE↓	MAPE↓	RMSE↓	PCC↑	KCC↑	SCC↑
SVR	$model_{W2V}$	4.02	0.25	5.69	0.84*	**0.81***	**0.92***
	$model_{W2V+LBF}$	3.94	0.24	5.64	0.84*	**0.84***	**0.93***
LRM	$model_{W2V}$	10.96	0.52	13.78	0.48*	0.40*	0.52*
	$model_{W2V+LBF}$	4.06	0.20	5.09	0.88*	0.72*	0.87*
Ridge	$model_{W2V}$	5.19	0.25	7.15	0.74*	0.57*	0.73*
	$model_{W2V+LBF}$	3.73	0.18	4.78	0.88*	0.73*	0.88*
Lasso	$\boldsymbol{model_{W2V}}$	**3.48**	**0.18**	**4.27**	**0.91***	0.77*	0.90*
	$\boldsymbol{model_{W2V+LBF}}$	**3.31**	**0.17**	**4.21**	**0.90***	0.80*	0.92*
AdaR	$model_{W2V}$	3.92	0.21	5.07	0.84*	0.72*	0.87*
	$model_{W2V+LBF}$	3.60	0.19	4.56	0.88*	0.76*	0.89*
SGDR	$model_{W2V}$	4.58	0.22	6.25	0.79*	0.60*	0.75*
	$model_{W2V+LBF}$	3.62	0.18	4.60	0.89*	0.74*	0.88*
RFR	$model_{W2V}$	3.79	0.21	5.07	0.83*	0.73*	0.88*
	$model_{W2V+LBF}$	3.83	0.21	4.87	0.87*	0.78*	0.91*
XTsR	$model_{W2V}$	3.83	0.21	5.08	0.83*	0.72*	0.88*
	$model_{W2V+LBF}$	3.77	0.20	4.75	0.88*	0.77*	0.91*
XGBR	$model_{W2V}$	3.69	0.21	5.37	0.76*	0.72*	0.87*
	$model_{W2V+LBF}$	3.83	0.20	5.09	0.83*	0.70*	0.84*

* P-value < 0.05 showing statistically significant correlation.

As shown in Tables 4 and 5, BERT models demonstrate the ability to capture a significant amount of information for classifying AD vs. Non-AD cases using transcripts. Previous studies such as Jawahar et al. (2019) have demonstrated that BERTs can understand the lexicon, syntax, and semantics of transcribed speech. This finding has been further supported in AD detection tasks using the ADReSS dataset (Balagopalan

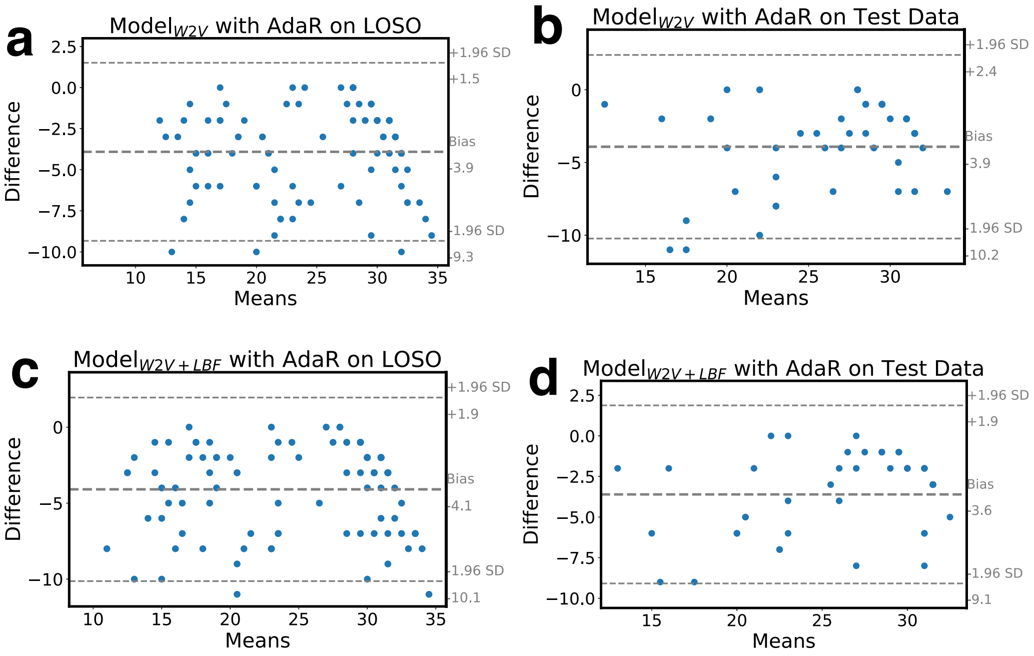

Fig. 1. Bland-Altman plots depicting the performance of the AdaR regressor alongside *model*$_{W2V}$ for LOSO validation (a) and test/unseen data (b), and *model*$_{W2V+LBF}$ for LOSO validation (c) and test/unseen data (d). The dashed lines show upper (+1.96 $*$ standard deviation) and lower (−1.96 $*$ standard deviation) confidence intervals.

Table 11
Comparing the performance of the proposed models with the previous literature on the test set. RMSE: root mean square error, AC: accuracy.

Study	Model	AC↑	RMSE↓
Balagopalan et al. (2020, 2021)	BERT, SVM, and Ridge	0.83	4.56
Rohanian et al. (2021)	LSTM with Gating using acoustics and lexical.	0.79	4.54
Farzana and Parde (2020)	Token-level psycholinguistic and sentiment features with SVR	–	4.34
Sarawgi et al. (2020)	Disfluency, acoustics, and cognitive features with ensemble MLP	0.83	4.60
Koo et al. (2020)	Ensemble of Transformer-XL, VGGish, and Global Vectors (GLoVE)	0.81	3.75
Syed et al. (2020)	Incorporating acoustics and various BERTs with SVM and SVR	0.85	4.30
Martinc et al. (2021)	Fusion of acoustics and four-character sequences models	0.94	–
Haulcy and Glass (2021)	BERT with LBF features with SVM and Gradboost	0.85	4.56
Meghanani et al. (2021)	fasttext and bi-gram with CNN	0.83	4.28
Luz et al. (2020)[a]	Linguistic features with LDA and DTR	0.75	5.20
Proposed model	**word2vec-based concentrated embedding with GNB and Lasso**	**0.92**	**4.21**

[a] The baseline of the ADReSS challenge.

et al., 2020, 2021; Haulcy and Glass, 2021). Moreover, additional investigations have revealed that "uh" and "um", which emphasize potential pauses and transitions between subsequent sentences, are important predictors within the self-attention layers of BERTs (Balagopalan et al., 2021; Yuan et al., 2020).

Table 4 demonstrates that both *Bio-Clinical BERT* and *DistilBERT* perform equally well in terms of accuracy and harmonic score. However, as indicated in Table 5, *Bio-Clinical BERT* exhibits the best results on the test set. The *Bio-Clinical BERT* model has been trained on a large corpus, incorporating both general texts and clinical notes, which enhances its generalizability compared to other models. Additionally, since the ADReSS dataset focuses on describing scenes, often using expressions from daily conversations, models trained on corpora such as books or Wikipedia are expected to perform well in the contextual classification of sentences and transcripts. The performance of *Bio-Clinical BERT* on the test set surpassed even its performance on the LOSO train set. This observation aligns with findings from other studies, including both BERT-based (Balagopalan et al., 2020, 2021; Haulcy and Glass, 2021; Syed et al., 2020) and non-BERT-based approaches (Rohanian et al., 2021; Meghanani et al., 2021). Our proposed models, as indicated in Tables 6 and 7, also consistently outperformed on unseen data. This phenomenon can be attributed to the careful stratification of the train and test sets, ensuring a well-balanced distribution of various factors such as age, sex, and class proportions. In addition, we explored the feasibility of using BERT models to directly estimate the MMSE score. However, most models, except *DistilBERT*, yielded poor results. This could be attributed to the fact that the BERT models used in this study were specifically designed for classification tasks and simply replacing the last sigmoid layer with a ReLU layer and changing the loss function to Huber loss does not guarantee acceptable performance in the regression task. The relatively good performance of *DistilBERT* can be attributed to its intrinsic distillation process, which results in fewer parameters and reduces the likelihood of immediate overfitting during fine-tuning.

Our proposed model focuses on the concentration of words spoken during a given task by computing the arithmetic median of these words. To discriminate between transcripts that heavily emphasize a specific group of words and those that exhibit diversity in the words spoken, the representation is normalized by the standard deviation. Previous studies have indicated that individuals with AD often exhibit limited content and coherence in their speech (Riley et al., 2005; Le et al., 2011; Ai and Lu, 2010; Boschi et al., 2017), both of which can be captured to some extent by our proposed model.

While our word2vec-based representation is capable of extracting underlying semantic meanings from the spoken words, it may not fully capture features related to the use of shorter words, increased pronoun and adverb usage, and the presence of words not found in the dictionary. To enhance the generalizability and interpretability of the basic model ($model_{W2V}$), we incorporate linguistic-based features, referred to as LBF. Tables 6, 7, 9, and 10 provide evidence of the improved performance of the LBF-combined model ($model_{W2V+LBF}$) compared to the $model_{W2V}$ across various classifiers and regressors.

The MMSE value is an important indicator of AD progression, and accurate prediction of the MMSE score is essential. To assess the effectiveness of the fine-tuned and proposed regression models in estimating the MMSE score, we employed three correlation-based metrics, along with MAE, MAPE, and RMSE. Among the correlation methods we employed, Pearson's correlation assumes normal distribution for both the original and predicted values, which may not be fully met in many scenarios. Additionally, Pearson's correlation assesses the linearity between the original and predicted values, which may have limited physiological interpretability when it comes to MMSE prediction. On the other hand, non-parametric correlation analyses such as Spearman's rank correlation and Kendall's rank correlation do not require any assumptions regarding the distribution of the data. They focus on the monotonic relationship (rank-wise) between the original and estimated values. Moreover, they are less sensitive to outliers compared to Pearson's correlation.

Contrary to our expectation of a positive alignment between KCC/SCC and lower RMSE, indicating higher correlation and lower error, the SVR model exhibited the highest KCC and SCC values in both the LOSO (Table 9) and unseen data (Table 10), while having a relatively high RMSE. Furthermore, the results show a strong positive linear correlation between PCC values and SCC/KCC values, suggesting that most of the original and predicted values exhibit a linear monotonic relationship.

As shown in Table 11, our proposed classification model outperforms all existing literature on the test set. However, in terms of the regression task, the suggested regression model demonstrates lower performance only compared to the work of Koo et al. (2020), where an ensemble of different neural networks, including TransformerXL, VGGish, and GLoVE, was leveraged for predicting MMSE values. Although each individual model in their ensembled model exhibited poorer performance than our proposed model, the fusion led to higher overall performance in the regression task.

In some previous work, acoustic features extracted from speech were also incorporated (Sarawgi et al., 2020; Syed et al., 2020; Martinc et al., 2021). However, even with the inclusion of these acoustic features, the performance of the models remained lower than the NLP-based models. This highlights the significance of linguistic features extracted from transcripts in revealing certain degrees of AD.

Although our model has demonstrated exceptional performance in both regression and classification tasks, there are several notable limitations that should be acknowledged. Firstly, the embeddings used for feature selection are of high dimensionality, and the choice of feature selection model and its hyperparameters can significantly impact the performance of our proposed technique. Future studies should explore the feasibility of employing different feature selection methods and considering a fusion of various approaches. Secondly, our model does not leverage the contextual representation of words, which may introduce biases and unfair representations of words across the transcript. While training word2vec models on large corpora can mitigate this limitation to some extent, future research should investigate the fusion of contextual models' embeddings with word2vec-based representations for improved performance. Thirdly, as the progression of AD can vary among individuals, it would be valuable to conduct an analysis that groups individuals with AD based on their MMSE levels. This subgroup analysis would allow for evaluating the model's performance within each subset of data, providing more insights into its effectiveness. Moreover, the fine-tuning of large language models (LLMs) like BERT and GPT-2 demands a substantial volume of data. Therefore, we encourage future studies to consider the inclusion of larger datasets, such as DementiaBank,[4] and to merge them with the ADReSS dataset to have an extensive collection of transcripts and further enhance the robustness and generalizability of the developed models. Lastly, it is important to note that our proposed method relies on semi-automatic processes, requiring transcripts for input. In future research, it would be worthwhile to explore automatic approaches where speech is transcribed directly and utilized by the model.

5. Conclusion

In this paper, we presented a novel word2vec-based model for classifying individuals with AD and estimating their corresponding AD score. Our model used a non-contextualized subject-specific embedding representation to capture the concentration of the words within transcripts of speech. Additionally, we investigated the effectiveness of different contextualized models, including several BERT models and GPT-2. Our findings reveal that these contextualized models demonstrate relatively good performance in AD classification, but fall short in terms of MMSE estimation.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Behrad Taghibeyglou reports financial support was provided by AI4PH, a Health Re-search Training Platform funded by the Canadian Institutes of Health Research (CIHR).

Acknowledgements

This research is supported by AI4PH, a Health Research Training Platform funded by the Canadian Institutes of Health Research (CIHR).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.nlp.2023.100046.

References

Abdi, H., 2007. The Kendall rank correlation coefficient. In: Encyclopedia of Measurement and Statistics. Sage, Thousand Oaks, CA. pp. 508–510.

Ai, H., Lu, X., 2010. A web-based system for automatic measurement of lexical complexity. In: 27th Annual Symposium of the Computer-Assisted Language Consortium (CALICO-10). Amherst, MA. June. pp. 8–12.

Alsentzer, E., Murphy, J.R., Boag, W., Weng, W.-H., Jin, D., Naumann, T., McDermott, M., 2019. Publicly available clinical BERT embeddings. arXiv preprint arXiv:1904.03323.

AutoViML, 2020. FeatureWiz. https://github.com/AutoViML/featurewiz.

Balagopalan, A., Eyre, B., Robin, J., Rudzicz, F., Novikova, J., 2021. Comparing pretrained and feature-based models for prediction of Alzheimer's disease based on speech. Front. Aging Neurosci. 13, 635945.

Balagopalan, A., Eyre, B., Rudzicz, F., Novikova, J., 2020. To BERT or not to BERT: comparing speech and language-based approaches for Alzheimer's disease detection. arXiv preprint arXiv:2008.01551.

Bland, J.M., Altman, D., 1986. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 327 (8476), 307–310.

Boschi, V., Catricala, E., Consonni, M., Chesi, C., Moro, A., Cappa, S.F., 2017. Connected speech in neurodegenerative language disorders: a review. Front. Psychol. 8, 269.

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., VanderPlas, J., Joly, A., Holt, B., Varoquaux, G., 2013. API design for machine learning software: experiences from the scikit-learn project. In: ECML PKDD Workshop: Languages for Data Mining and Machine Learning. pp. 108–122.

Chen, J., Ye, J., Tang, F., Zhou, J., 2021. Automatic detection of Alzheimer's disease using spontaneous speech only. In: Interspeech, Vol. 2021. NIH Public Access, p. 3830.

Cockrell, J.R., Folstein, M.F., 2002. Mini-mental state examination. In: Principles and Practice of Geriatric Psychiatry. Wiley Online Library, pp. 140–141.

Cummins, N., Pan, Y., Ren, Z., Fritsch, J., Nallanthighal, V.S., Christensen, H., Blackburn, D., Schuller, B.W., Magimai-Doss, M., Strik, H., et al., 2020. A comparison of acoustic and linguistics methodologies for Alzheimer's dementia recognition. In: Interspeech. ISCA-International Speech Communication Association, pp. 2182–2186.

[4] https://dementia.talkbank.org/access/English/Pitt.html.

de la Fuente Garcia, S., Ritchie, C.W., Luz, S., 2020. Artificial intelligence, speech, and language processing approaches to monitoring Alzheimer's disease: a systematic review. J. Alzheimer's Dis. 78 (4), 1547–1574.

Devlin, J., Chang, M., Lee, K., Toutanova, K., 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. CoRR abs/1810.04805. arXiv:1810.04805.

Farzana, S., Parde, N., 2020. Exploring MMSE score prediction using verbal and non-verbal cues. In: Interspeech. pp. 2207–2211.

Fraser, K.C., Meltzer, J.A., Rudzicz, F., 2016. Linguistic features identify Alzheimer's disease in narrative speech. J. Alzheimer's Dis. 49 (2), 407–422.

Geda, Y.E., Schneider, L.S., Gitlin, L.N., Miller, D.S., Smith, G.S., Bell, J., Evans, J., Lee, M., Porsteinsson, A., Lanctôt, K.L., et al., 2013. Neuropsychiatric symptoms in Alzheimer's disease: past progress and anticipation of the future. Alzheimer's Dement. 9 (5), 602–608.

Goodglass, H., Kaplan, E., Weintraub, S., 2001. BDAE: The Boston Diagnostic Aphasia Examination. Lippincott Williams & Wilkins, Philadelphia, PA.

Guo, Y., Li, C., Roan, C., Pakhomov, S., Cohen, T., 2021. Crossing the "Cookie Theft" corpus chasm: applying what BERT learns from outside data to the ADReSS challenge dementia detection task. Front. Comput. Sci. 3, 642517.

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., Smith, N.A., 2020. Don't stop pretraining: Adapt language models to domains and tasks. In: Proceedings of ACL.

Hauke, J., Kossowski, T., 2011. Comparison of values of Pearson's and Spearman's correlation coefficients on the same sets of data. Quaest. Geogr. 30 (2), 87–93.

Haulcy, R., Glass, J., 2021. Classifying Alzheimer's disease using audio and text-based representations of speech. Front. Psychol. 11, 624137.

Ilias, L., Askounis, D., 2022. Multimodal deep learning models for detecting dementia from speech and transcripts. Front. Aging Neurosci. 14.

Jawahar, G., Sagot, B., Seddah, D., 2019. What does BERT learn about the structure of language? In: ACL 2019-57th Annual Meeting of the Association for Computational Linguistics.

Johnson, A.E., Pollard, T.J., Shen, L., Lehman, L.-w.H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., Mark, R.G., 2016. MIMIC-III, a freely accessible critical care database. Sci. Data 3 (1), 1–9.

König, A., Satt, A., Sorin, A., Hoory, R., Toledo-Ronen, O., Derreumaux, A., Manera, V., Verhey, F., Aalten, P., Robert, P.H., et al., 2015. Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease. Alzheimer's Dement.: Diagn. Assess. Dis. Monit. 1 (1), 112–124.

Koo, J., Lee, J.H., Pyo, J., Jo, Y., Lee, K., 2020. Exploiting multi-modal features from pre-trained networks for Alzheimer's dementia recognition. arXiv preprint arXiv:2009.04070.

Kurlowicz, L., Wallace, M., 1999. The mini-mental state examination (MMSE). J. Gerontol. Nurs. 25 (5), 8–9.

Land, Jr., W.H., Schaffer, J.D., Land, W.H., Schaffer, J.D., 2020. Alzheimer's disease and speech background. In: The Art and Science of Machine Intelligence: With An Innovative Application for Alzheimer's Detection from Speech. Springer, pp. 107–135.

Le, X., Lancashire, I., Hirst, G., Jokel, R., 2011. Longitudinal detection of dementia through lexical and syntactic changes in writing: a case study of three British novelists. Lit. Linguist. Comput. 26 (4), 435–461.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J., 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 36 (4), 1234–1240.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V., 2019. RoBERTa: A robustly optimized BERT pretraining approach. CoRR abs/1907.11692. arXiv:1907.11692.

López-de-Ipiña, K., Solé-Casals, J., Eguiraun, H., Alonso, J.B., Travieso, C.M., Ezeiza, A., Barroso, N., Ecay-Torres, M., Martinez-Lage, P., Beitia, B., 2015. Feature selection for spontaneous speech analysis to aid in Alzheimer's disease diagnosis: A fractal dimension approach. Comput. Speech Lang. 30 (1), 43–60.

Loshchilov, I., Hutter, F., 2017. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101.

Luz, S., Haider, F., de la Fuente, S., Fromm, D., MacWhinney, B., 2020. Alzheimer's dementia recognition through spontaneous speech: The ADReSS challenge. arXiv preprint arXiv:2004.06833.

Luz, S., Haider, F., de la Fuente, S., Fromm, D., MacWhinney, B., 2021. Detecting cognitive decline using speech only: The ADReSSo challenge. arXiv preprint arXiv:2104.09356.

MacWhinney, B., 2014. The CHILDES Project: Tools for Analyzing Talk, Volume I: Transcription Format and Programs. Psychology Press.

MacWhinney, B., 2017a. Tools for analyzing talk part 1: The chat transcription format. Carnegie.[Google Scholar] 16.

MacWhinney, B., 2017b. Tools for analyzing talk part 2: The CLAN program. Talkbank. Org (2000).

Martinc, M., Haider, F., Pollak, S., Luz, S., 2021. Temporal integration of text transcripts and acoustic features for Alzheimer's diagnosis based on spontaneous speech. Front. Aging Neurosci. 13, 642647.

Meghanani, A., Anoop, C., Ramakrishnan, A.G., 2021. Recognition of Alzheimer's dementia from the transcriptions of spontaneous speech using FastText and CNN models. Front. Comput. Sci. 3, 624558.

Mirheidari, B., Blackburn, D., Reuber, M., Walker, T., Christensen, H., 2016. Diagnosing people with dementia using automatic conversation analysis. In: Proceedings of Interspeech. ISCA, pp. 1220–1224.

Mirheidari, B., Blackburn, D., Walker, T., Venneri, A., Reuber, M., Christensen, H., 2018. Detecting signs of dementia using word vector representations. In: Interspeech. pp. 1893–1897.

Nasreddine, Z.S., Phillips, N.A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., Cummings, J.L., Chertkow, H., 2003. Montreal cognitive assessment. Am. J. Geriatr. Psychiatry.

Orimaye, S.O., Wong, J.S.-M., Golden, K.J., 2014. Learning predictive linguistic features for Alzheimer's disease and related dementias using verbal utterances. In: Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality. pp. 78–87.

Pan, Y., Mirheidari, B., Harris, J.M., Thompson, J.C., Jones, M., Snowden, J.S., Blackburn, D., Christensen, H., 2021. Using the outputs of different automatic speech recognition paradigms for acoustic-and BERT-based Alzheimer's dementia detection through spontaneous speech. In: Interspeech. pp. 3810–3814.

Pappagari, R., Cho, J., Joshi, S., Moro-Velázquez, L., Zelasko, P., Villalba, J., Dehak, N., 2021. Automatic detection and assessment of Alzheimer disease using speech and language technologies in low-resource scenarios. In: Interspeech. pp. 3825–3829.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Peng, H., Long, F., Ding, C., 2005. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans. Pattern Anal. Mach. Intell. 27 (8), 1226–1238.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., 2019. Language models are unsupervised multitask learners.

Řehůřek, R., Sojka, P., et al., 2011. Gensim—statistical semantics in python. Retrieved from genism. org.

Rentoumi, V., Raoufian, L., Ahmed, S., de Jager, C.A., Garrard, P., 2014. Features and machine learning classification of connected speech samples from patients with autopsy proven Alzheimer's disease with and without additional vascular pathology. J. Alzheimer's Dis. 42 (s3), S3–S17.

Riley, K.P., Snowdon, D.A., Desrosiers, M.F., Markesbery, W.R., 2005. Early life linguistic ability, late life cognitive function, and neuropathology: findings from the Nun Study. Neurobiol. Aging 26 (3), 341–347.

Rohanian, M., Hough, J., Purver, M., 2021. Multi-modal fusion with gating using audio, lexical and disfluency features for Alzheimer's dementia recognition from spontaneous speech. arXiv preprint arXiv:2106.09668.

Sadeghian, R., Schaffer, J.D., Zahorian, S.A., 2017. Speech processing approach for diagnosing dementia in an early stage. In: Interspeech. pp. 2705–2709.

Sanh, V., Debut, L., Chaumond, J., Wolf, T., 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.

Sarawgi, U., Zulfikar, W., Soliman, N., Maes, P., 2020. Multimodal inductive transfer learning for detection of Alzheimer's dementia and its severity. arXiv preprint arXiv:2009.00700.

Syed, M.S.S., Syed, Z.S., Lech, M., Pirogova, E., 2020. Automated screening for Alzheimer's dementia through spontaneous speech. In: Interspeech, Vol. 2020. pp. 2222–2226.

Taghibeyglou, B., Rudzicz, F., 2023. Who needs context? classical techniques for alzheimer's disease detection. In: Proceedings of the 5th Clinical Natural Language Processing Workshop. pp. 102–107.

Thomas, C., Keselj, V., Cercone, N., Rockwood, K., Asp, E., 2005. Automatic detection and rating of dementia of Alzheimer type through lexical analysis of spontaneous speech. In: IEEE International Conference Mechatronics and Automation, 2005, Vol. 3. IEEE, pp. 1569–1574.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Adv. Neural Inf. Process. Syst. 30.

Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S.J., Brett, M., Wilson, J., Millman, K.J., Mayorov, N., Nelson, A.R.J., Jones, E., Kern, R., Larson, E., Carey, C.J., Polat, İ., Feng, Y., Moore, E.W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E.A., Harris, C.R., Archibald, A.M., Ribeiro, A.H., Pedregosa, F., van Mulbregt, P., SciPy 1.0 Contributors, 2020. SciPy 1.0: Fundamental algorithms for scientific computing in Python. Nature Methods 17, 261–272. http://dx.doi.org/10.1038/s41592-019-0686-2.

Voleti, R., Liss, J.M., Berisha, V., 2019. A review of automated speech and language features for assessment of cognitive and thought disorders. IEEE J. Sel. Top. Sign. Proces. 14 (2), 282–298.

Wankerl, S., Nöth, E., Evert, S., 2017. An N-Gram based approach to the automatic diagnosis of Alzheimer's disease from spoken language. In: Interspeech. pp. 3162–3166.

Weiner, J., Engelbart, M., Schultz, T., 2017. Manual and automatic transcriptions in dementia detection from speech. In: Interspeech. pp. 3117–3121.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., Rush, A., 2020. Transformers: State-of-the-art natural language processing. In:

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Association for Computational Linguistics, Online, pp. 38–45.

Yamada, I., Asai, A., Sakuma, J., Shindo, H., Takeda, H., Takefuji, Y., Matsumoto, Y., 2018. Wikipedia2Vec: An efficient toolkit for learning and visualizing the embeddings of words and entities from Wikipedia. arXiv preprint arXiv:1812.06280.

Yuan, J., Bian, Y., Cai, X., Huang, J., Ye, Z., Church, K., 2020. Disfluencies and fine-tuning pre-trained language models for detection of Alzheimer's disease. In: Interspeech, Vol. 2020. pp. 2162–2166.

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., Fidler, S., 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 19–27.