Who needs context? Classical techniques for Alzheimer's disease detection

Behrad TaghiBeyglou[1,2] and **Frank Rudzicz**[3,4,5]

[1]Institute of Biomedical Engineering, University of Toronto, Toronto, Canada
[2]KITE- Toronto Rehabilitation Institute, University Health Network, Toronto, Canada
behrad.taghibeyglou@mail.utoronto.ca
[3]Faculty of Computer Science, Dalhousie University, Halifax, Canada
[4]Department of Computer Science, University of Toronto, Toronto, Canada
[5]Vector Institute for Artificial Intelligence, Toronto, Canada
frank@dal.ca

Abstract

Natural language processing (NLP) has shown great potential for Alzheimer's disease (AD) detection, particularly due to the adverse effect of AD on spontaneous speech. The current body of literature has directed attention toward context-based models, especially Bidirectional Encoder Representations from Transformers (BERTs), owing to their exceptional abilities to integrate contextual information in a wide range of NLP tasks. This comes at the cost of added model opacity and computational requirements. Taking this into consideration, we propose a Word2Vec-based model for AD detection in 108 age- and sex-matched participants who were asked to describe the Cookie Theft picture. We also investigate the effectiveness of our model by fine-tuning BERT-based sequence classification models, as well as incorporating linguistic features. Our results demonstrate that our lightweight and easy-to-implement model outperforms some of the state-of-the-art models available in the literature, as well as BERT models.

1 Introduction

Alzheimer's disease (AD) is the most prevalent form of dementia, a 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 and 139 million by 2030 and 2050, respectively (Ilias and Askounis, 2022). Symptoms of AD include (but are not limited to) memory decline, disorientation, confusion, and behavioural changes. Importantly, AD progression can lead to loss of independence which significantly impacts patients, their families, and society as a whole (Pappagari et al., 2021). Given that late-stage AD progression is inevitable, early detection of AD through cost-effective and scalable technologies is critical. While most clinical diagnoses of AD rely on neuroimaging, there is a critical need for more accessible and efficient methods of diagnosis.

Accessible evaluation methods for AD include 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 integration with 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). Speech processing and, consequently, natural language processing (NLP) can therefore provide new precision medicine tools for AD diagnosis that deliver objective quantitative analyses and reliable proof, analysis, comparison, and circulation for faster diagnosis.

The Alzheimer's Dementia Recognition through Spontaneous Speech (ADReSS) challenge of INTERSPEECH 2020 is a shared database developed to advance research into automatic AD detection based on spontaneous speech and transcripts (Luz et al., 2020). Participants in the challenge were tasked with describing the Cookie Theft picture in English, which is part of the Boston Diagnostic Aphasia Exam (Guo et al., 2021). The first set of the ADReSS 2020 database comprises speech recordings and CLAN-annotated transcripts of 54 AD patients and 54 sex- and age-matched controls.

Various groups have worked with the ADReSS dataset, approaching the problem from different perspectives and leveraging available information. These studies typically combined speech processing and linguistic feature extraction or NLP-based

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 102–107
July 14, 2023 ©2023 Association for Computational Linguistics

fine-tuning. The literature on speech processing mostly focused on zero-crossing rate, spectral bandwidth, roll-off, and centroids of audio recordings, as well as active data representation cluster-based feature extraction methods including the emobase (Eyben et al., 2010), ComParE (Eyben et al., 2013), and Multi-Resolution Cochleagram (MRCG) (Chen et al., 2014) feature sets. Meanwhile, linguistic features have extracted lexical richness, the proportion of various PoS tags, utterance duration, total utterances, type-token ratio, open-closed class word ratio, and similarity between consecutive utterances. NLP-based methods have comprised from-scratch training or 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), and Bidirectional Encoder Representations from Transformer (BERT) (Balagopalan 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 present a novel approach for detecting AD in the first set of ADReSS dataset by integrating a new Word2Vec-based model and dimension reduction method. We not only implement and compare top-cited and recent state-of-the-art models on the same dataset, but also demonstrate that our approach outperforms these models. Our proposed approach is simple, easy to implement, and highly accurate.

2 Methodology

2.1 Other models

In order to evaluate the performance of our proposed language processing model, we have considered several publicly available models for comparison including:

- **Linguistic-Based Features (LBF)**: In this study, we utilized the CLAN package to extract 34 linguistic-based features (LBFs) from transcripts, including duration, total utterances, mean length of utterance (MLU), type-token ratio, open-closed class word ratio, and percentages of 9 parts of speech. We also incorporated demographic information such as age and sex. To identify the most informative features for classification, we performed correlation and variance analyses on the extracted features using the FeatureWiz package (AutoViML, 2020). We set a correlation threshold of 0.6 and repeated the analyses 5 times with random seeds over all samples. We then selected the top 5 features that appeared in at least 3 iterations for further classification.

- **BERT Models**: Since BERT models have shown promising performance in different applications of NLP, in this study we leveraged some of BERT-based architectures with a maximum length of 512 tokens as a reference for our model. We tested three versions of uncased base BERT (Devlin et al., 2018): one with no extension in the last layers, called *baseBERT1*, another with two fully connected layers at the end ($768 \rightarrow 64$ and $64 \rightarrow 1$), called *baseBERT2*, and the last one with three fully connected layers ($768 \rightarrow 128$, $128 \rightarrow 16$, and $16 \rightarrow 1$), called *baseBERT3*. For *baseBERT2*, we varied the training epochs between 3 and 5. Additionally, we tested *Bio-CLinical BERT* (Alsentzer et al., 2019) with a batch size of 4 and 3 epochs, *DistilBERT* (Sanh et al., 2019) with a batch size of 4 and 3 epochs, and *BioMed-RoBERTa-based* (Gururangan et al., 2020) with a batch size of 4 and 3 epochs. We used binary cross-entropy as the loss function for all models and AdamW (Adam with weight decay) (Loshchilov and Hutter, 2017) as the optimizer with a learning rate of 2×10^{-5}. To address potential issues with local optima, we applied a linear warm-up scheduler. Each transcript is classified as AD if the average of the probabilities (after the sigmoid layer) over all sentences in the transcript is greater than or equal to 0.5; otherwise, it is classified as control.

2.2 Pre-processing

To preprocess the data for our proposed model, we have neglected the first four sentences of each transcript, as the initial speaker is typically a member of the data collection team. Additionally, stop words were removed from each sentence using the Gensim library (Řehřek et al., 2011).

2.3 Proposed model

In this study, we used Wikipedia2Vec (Yamada et al., 2018), a tool that generates embeddings (or vector representations) of words and entities from Wikipedia, to convert tokens to vector embeddings. We used the skip-gram strategy for training, and the embedding dimension of the model was set to 500. We denote this model as $W2V$ throughout this paper. Suppose that each participant's transcript consists of N_k sentences, each comprising m words, where m varies from 1 to M_k (the maximum length among all sentences in the k^{th} transcript). We input each word $\langle w_{i,k}\rangle$ into the $W2V$ model ($W2V(\langle w_{i,k}\rangle)$), which outputs the corresponding embedded vector $\mathbf{x}_{i,k} \in \mathcal{R}^{500}$. All embeddings of the k^{th} transcript form the set X_k. We standardized each 500-dimensional vector across all embeddings of each subject using the following formula:

$$\mathbf{y}_k = \frac{\mathtt{med}(X_k)}{\mathtt{std}(X_k)}, \tag{1}$$

where $\mathtt{med}$ is the median operator applied to each dimension independently, $\mathtt{std}$ is the standard deviation of embeddings, and $\mathbf{y}_k$ denotes the standardized vector for the k^{th} participant. So far, we developed the first framework and leveraged the previously introduced feature selection method by iteratively applying FeatureWiz five times. We then selected features that were chosen at least three times during the process to identify the most informative dimensions for AD detection. This feature selection procedure reduced the dimension from 500 to 64. We refer to this first framework as *model 1*, and Figure 1 illustrates the process. To further enhance our analysis, we concatenate linguistics-based features from the previous section with W2V-based feature vectors and apply feature selection in a similar manner to *model 1*. This second framework, called *model 2*, resulted in the selection of 86 features (out of 537 features). Prior to inputting the features into the classifiers of each model, the zero-mean-unit-variance standardization technique is applied to normalize the features.

2.4 Evaluation and Metrics

All results presented in this study were obtained using the leave-one-subject-out (LOSO) cross-validation technique to evaluate the generalizability of the models. Thus, a total of 104 models were trained per architecture/classifier. For each model, accuracy, sensitivity, specificity, and F1 were reported as performance metrics. For the feature-based models, such as linguist-based features and our proposed frameworks, we employed various classifiers including logistic regression (LR), decision tree (DT), linear and Nu-support vector classification (SVC), linear and quadratic discriminant analysis (LDA and QDA), Gaussian naive Bayes (GNB), extreme gradient boosting (XGBoost), adaptive boosting (AdaBoost), and extra trees classifier.

3 Results

3.1 Other models

We investigated different BERT models for AD classification, and the results are presented in Table 1. As expected, the performance of *Bio-Clinical BERT* and *DistilBERT* models were comparable; however, *Bio-Clinical BERT* showed superior sensitivity and was chosen as the best BERT model in this study. Additionally, as demonstrated in Table 2, integrating linguistic-based features with feature selection and a combination of classifiers achieved an accuracy of 0.81 in AD detection.

Model	E:BS	AC	SP	SE	F1
baseBERT1	3:4	0.80	0.89	0.7	0.78
baseBERT2	3:4	0.79	0.81	0.76	0.78
baseBERT2	5:4	0.79	0.93	0.65	0.77
baseBERT3	3:4	0.78	0.90	0.67	0.77
Bio-CLinical BERT	**3:4**	**0.84**	**0.85**	**0.83**	**0.84**
DistilBERT	3:4	0.84	0.87	0.81	0.84
BioMed-RoBERTa-based	3:4	0.81	0.87	0.76	0.81

Table 1: LOSO performance of other BERT-based models. "E" denotes the number of epochs, "BS" denotes the batch size, and "AC", "SP", and "SE" represent accuracy, specificity, and sensitivity, respectively.

3.2 Proposed frameworks

The performance of our proposed frameworks is presented in Table 3. The best performance was achieved by *model 2* with the help of the GNB classifier, which obtained an accuracy of 0.90. On the other hand, the best performance of *model 1* was achieved by the ExtraTrees classifier.

3.3 Comparison with previous literature

Table 4 compares our proposed model with the existing models in the literature as well as the ones explored in this paper. Our model achieved significantly higher performance, including a 3% improvement in accuracy and an 8% improvement in sensitivity compared to one of the BERT-based

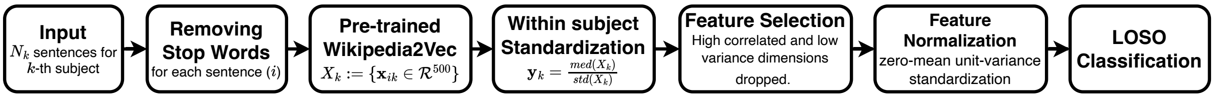
Input
N_k sentences for k-th subject
Removing Stop Words
for each sentence (i)
Pre-trained Wikipedia2Vec
$X_k := \{\mathbf{x}_{ik} \in \mathcal{R}^{500}\}$
Within subject Standardization
$\mathbf{y}_k = \frac{med(X_k)}{std(X_k)}$
Feature Selection
High correlated and low variance dimensions dropped.
Feature Normalization
zero-mean unit-variance standardization
LOSO Classification

Figure 1: Proposed framework for AD classification.

Classifier	AC	SP	SE	F1
LR	0.76	0.81	0.70	0.75
DT	0.69	0.74	0.63	0.68
Linear SVC	0.76	0.80	0.72	0.76
Nu-SVC	0.81	0.83	0.78	0.80
LDA	0.79	**0.89**	0.69	0.78
QDA	**0.81**	0.87	**0.76**	**0.81**
GNB	0.78	0.87	0.69	0.77
XGBoost	0.71	0.70	0.72	0.71
AdaBoost	0.74	0.76	0.72	0.74
ExtraTrees	0.72	0.76	0.69	0.72

Table 2: LOSO performance of the linguist feature-based model, in combination with the proposed feature selection technique.

Classifier	Model	AC	SP	SE	F1
LR	*model 1*	0.74	0.87	0.81	0.84
	model 2	0.74	0.89	0.80	0.84
DT	*model 1*	0.76	0.80	0.72	0.76
	model 2	0.56	0.54	0.57	0.55
Linear SVC	*model 1*	0.81	0.85	0.78	0.81
	model 2	0.80	0.85	0.74	0.79
Nu-SVC	*model 1*	0.85	0.85	0.85	0.85
	model 2	**0.90**	**0.91**	0.89	**0.9**
LDA	*model 1*	0.73	0.74	0.72	0.73
	model 2	0.66	0.69	0.63	0.66
QDA	*model 1*	0.60	0.63	0.57	0.6
	model 2	0.44	0.33	0.56	0.42
GNB	*model 1*	0.87	0.87	0.87	0.87
	model 2	**0.90**	0.89	**0.91**	**0.9**
XGBoost	*model 1*	0.77	0.76	0.78	0.77
	model 2	0.78	0.78	0.78	0.78
AdaBoost	*model 1*	0.81	0.78	0.85	0.81
	model 2	0.82	0.81	0.83	0.82
ExtraTrees	*model 1*	0.88	0.89	0.87	0.88
	model 2	0.89	0.91	0.87	0.89

Table 3: LOSO performance of the linguist feature-based model, in combination with the proposed feature selection technique.

models on the same dataset (Balagopalan et al., 2020, 2021). It is worth noting that our proposed model also outperformed the baseline linguistic model introduced in the ADReSS challenge.

Model	AC	SP	SE	F1
Bio-CLinical BERT	0.84	0.85	0.83	0.84
Best Linguist-based features	0.81	0.87	0.76	0.81
BERT and SVM (Balagopalan et al., 2020, 2021)	0.87	**0.91**	0.83	0.87
Gated LSTM on acoustic and lexical (Rohanian et al., 2021)	0.77	-	-	-
Baseline Linguistic (Luz et al., 2020)	0.77	0.77	0.76	0.77
Best proposed model	**0.90**	0.89	**0.91**	**0.9**

Table 4: LOSO performance comparison of the best proposed model and explored models with some existing models on the same dataset. The best linguist-based features model uses QDA classifier with linguist-based features, and the best proposed model is our proposed *model 2* with GNB classifier.

4 Discussion

By mapping each word into a 500-dimensional space where words with similar context are closer together, the proposed model can identify when all words in a transcript are focused on the same topic with minimal deviations. Coupled with the suggested standardization method, the results demonstrate a significant difference in performance between the proposed model and the only linguist-based model, which prioritizes utterances, pauses, and interactions between text and speech. The BERT models explored in this study are relatively massive and require significant computational resources, and training them requires delicate hyperparameter optimization. In this study, we followed the BERT authors' recommendations to keep the model's trainability on an Nvidia RTX 3080 GPU and to avoid changing the weights of the model by selecting smaller epoch numbers.

5 Conclusions

In this study, we introduced a word2vec-based model that combines pre-trained Wikipedia embeddings with linguistic features. We also employed correlation-based feature selection to reduce the dimensionality of the embeddings. The results demonstrated that our proposed model outperformed existing models on the same dataset. However, as BERT models offer diverse applicability, a potential future direction is to incorporate feature maps extracted from the hidden states of these networks to enhance the performance of our model.

Acknowledgment

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

References

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*.

AutoViML. 2020. featurewiz. `https://github.com/AutoViML/featurewiz`.

Aparna Balagopalan, Benjamin Eyre, Jessica Robin, Frank Rudzicz, and Jekaterina Novikova. 2021. Comparing pre-trained and feature-based models for prediction of alzheimer's disease based on speech. *Frontiers in aging neuroscience*, 13:635945.

Aparna Balagopalan, Benjamin Eyre, Frank Rudzicz, and Jekaterina Novikova. 2020. To bert or not to bert: comparing speech and language-based approaches for alzheimer's disease detection. *arXiv preprint arXiv:2008.01551*.

Jitong Chen, Yuxuan Wang, and DeLiang Wang. 2014. A feature study for classification-based speech separation at low signal-to-noise ratios. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(12):1993–2002.

Jun Chen, Jieping Ye, Fengyi Tang, and Jiayu Zhou. 2021. Automatic detection of alzheimer's disease using spontaneous speech only. In *Interspeech*, volume 2021, page 3830. NIH Public Access.

Nicholas Cummins, Yilin Pan, Zhao Ren, Julian Fritsch, Venkata Srikanth Nallanthighal, Heidi Christensen, Daniel Blackburn, Björn W Schuller, Mathew Magimai-Doss, Helmer Strik, et al. 2020. A comparison of acoustic and linguistics methodologies for alzheimer's dementia recognition. In *Interspeech 2020*, pages 2182–2186. ISCA-International Speech Communication Association.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller. 2013. Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 835–838.

Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462.

Yue Guo, Changye Li, Carol Roan, Serguei Pakhomov, and Trevor Cohen. 2021. Crossing the "cookie theft" corpus chasm: applying what bert learns from outside data to the adress challenge dementia detection task. *Frontiers in Computer Science*, 3:642517.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of ACL*.

Loukas Ilias and Dimitris Askounis. 2022. Multimodal deep learning models for detecting dementia from speech and transcripts. *Frontiers in Aging Neuroscience*, 14.

Junghyun Koo, Jie Hwan Lee, Jaewoo Pyo, Yujin Jo, and Kyogu Lee. 2020. Exploiting multimodal features from pre-trained networks for alzheimer's dementia recognition. *arXiv preprint arXiv:2009.04070*.

Lenore Kurlowicz and Meredith Wallace. 1999. The mini-mental state examination (mmse). *Journal of gerontological nursing*, 25(5):8–9.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Saturnino Luz, Fasih Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney. 2020. Alzheimer's dementia recognition through spontaneous speech: The adress challenge. *arXiv preprint arXiv:2004.06833*.

Saturnino Luz, Fasih Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney. 2021. Detecting cognitive decline using speech only: The adresso challenge. *arXiv preprint arXiv:2104.09356*.

Bahman Mirheidari, Daniel Blackburn, Traci Walker, Annalena Venneri, Markus Reuber, and Heidi Christensen. 2018. Detecting signs of dementia using word vector representations. In *Interspeech*, pages 1893–1897.

Ziad S Nasreddine, Natalie A Phillips, Valérie Bédirian, Simon Charbonneau, Victor Whitehead, Isabelle Collin, Jeffrey L Cummings, and Howard Chertkow. 2003. Montreal cognitive assessment. *The American Journal of Geriatric Psychiatry*.

Yilin Pan, Bahman Mirheidari, Jennifer M Harris, Jennifer C Thompson, Matthew Jones, Julie S Snowden, Daniel Blackburn, and Heidi Christensen. 2021. Using the outputs of different automatic speech recognition paradigms for acoustic-and bert-based alzheimer's dementia detection through spontaneous speech. In *Interspeech*, pages 3810–3814.

Raghavendra Pappagari, Jaejin Cho, Sonal Joshi, Laureano Moro-Velázquez, Piotr Zelasko, Jesús Villalba, and Najim Dehak. 2021. Automatic detection and assessment of alzheimer disease using speech and language technologies in low-resource scenarios. In *Interspeech*, pages 3825–3829.

Radim Řehřek, Petr Sojka, et al. 2011. Gensim—statistical semantics in python. *Retrieved from genism. org*.

Morteza Rohanian, Julian Hough, and Matthew Purver. 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*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Ikuya Yamada, Akari Asai, Jin Sakuma, Hiroyuki Shindo, Hideaki Takeda, Yoshiyasu Takefuji, and Yuji Matsumoto. 2018. Wikipedia2vec: An efficient toolkit for learning and visualizing the embeddings of words and entities from wikipedia. *arXiv preprint arXiv:1812.06280*.