Recognition of Alzheimer's Dementia From the Transcriptions of Spontaneous Speech Using fastText and CNN Models

Amit Meghanani , Anoop C. S.* and Angarai Ganesan Ramakrishnan

MILE Laboratory, Department of Electrical Engineering, Indian Institute of Science, Bengaluru, India

Correspondence*:
Anoop C. S.
anoopcs@iisc.ac.in

ABSTRACT

Alzheimer's dementia (AD) is a type of neurodegenerative disease that is associated with a decline in memory. However, speech and language impairments are also common in Alzheimer's dementia patients. This work is an extension of our previous work, where we had used spontaneous speech for Alzheimer's dementia recognition employing log-Mel spectrogram and Mel frequency cepstral coefficients (MFCC) as inputs to deep neural networks (DNN). In this work, we explore the transcriptions of spontaneous speech for dementia recognition and compare the results with several baseline results. We explore two models for dementia recognition - i) fastText and ii) convolutional neural network (CNN) with a single convolutional layer, to capture the n-gram based linguistic information from the input sentence. The fastText model uses a bag of bigrams and trigrams along with the input text to capture the local word orderings. In the CNN based model, we try to capture different n-grams (we use n = 2,3,4,5) present in the text by adapting the kernel sizes to n. In both fastText and CNN architectures, the word embeddings are initialized using pre-trained GloVe vectors. We use bagging of 21 models in each of these architectures to arrive at the final model using which the performance on the test data is assessed. The best accuracies achieved with CNN and fastText models on the text data are 79.16% and 83.33%, respectively. The best root mean square errors (RMSE) on the prediction of mini-mental state examination (MMSE) score are 4.38 and 4.28 for CNN and fastText, respectively. The results suggest that the n-gram based features are worth pursuing, for the task of AD detection. fastText models have competitive results when compared to several baseline methods. Also, fastText models are shallow in nature and have the advantage of being faster in training and evaluation, by several orders of magnitude, compared to deep models.

Keywords: fastText, CNN, Alzheimer's, dementia, MMSE

1 INTRODUCTION

Dementia is a syndrome characterised by the decline in cognition that is significant enough to interfere with one's independent, daily functioning. Alzheimer's disease contributes to around 60–70% of dementia cases. Towards the final stages of Alzheimer's Dementia (AD), the patients lose control of their physical functions and depend on others for care. As there are no curative treatments for dementia, the early detection is critical to delay or slow down the onset or progression of the disease. The mini-mental state examination

(MMSE) is a widely used test to screen for dementia and to estimate the severity and progression of cognitive impairment.

AD affects the temporal characteristics of spontaneous speech. Changes in the spoken language are evident even in mild AD patients. Subtle language impairments such as difficulties in word finding and comprehension, usage of incorrect words, ambiguous referents, loss of verbal fluency, speaking too much at inappropriate times, talking too loudly, repeating ideas, and digressing from the topic are common in the early stages of AD (Savundranayagam et al., 2005) and they turn extreme in the moderate and severe stages. Szatlóczki et al. (2015) show that AD can be detected with the help of a linguistic analysis more sensitively than with other cognitive examinations. Mueller et al. (2018b) analyzed the connected language samples obtained from simple picture description tasks and found that the speech fluency and the semantic content features declined faster in participants with early mild cognitive impairment. The language profile of AD patients is characterized by “empty speech”, devoid of content words (Nicholas et al., 1985). They tend to use pronouns without proper noun references and indefinite terms like “this”, “that”, “thing” etc., more often (Mueller et al., 2018a). These results motivate us to believe that modeling the transcriptions of the narrative speech in the cookie-theft picture description task using n-gram language models can help in the detection of AD and prediction of MMSE score.

In this work we address the AD detection and MMSE score prediction problems using two natural language processing (NLP) based models - i) fastText and ii) convolutional neural network (CNN). These models have the advantage that they can be easily structured to capture the linguistic cues in the form of n-grams from the transcriptions of the picture description task, provided with the Alzheimer’s Dementia Recognition through Spontaneous Speech (ADReSS) dataset (Luz et al., 2020). CNNs, though originated in computer vision, have become popular for NLP tasks and have achieved great results in sentence classification (Kim, 2014), semantic parsing (tau Yih et al., 2014), search query retrieval (Shen et al., 2014), and other traditional NLP tasks (Collober et al., 2011). Our convolutional neural network model draws inspiration from the work on sentence classification using CNNs (Kim, 2014). The fastText (Joulin et al., 2017) is a simple and efficient model for text classification (eg. tag prediction and sentiment analysis). The fundamental idea in the fastText classifier is to calculate the n-grams of an input sentence and append them to the end of the sentence. Our choice of fastText model is also motivated by its ability to often outperform deep learning classifiers in terms of accuracy and training/evaluation times (Joulin et al., 2017).

The rest of the paper is organised as follows. Section 2 discusses the ADReSS dataset in detail. Section 3 discusses the baseline results in AD detection. Section 4 discusses our proposed NLP based models followed by the listing of results in section 5. Our results and conclusions are discussed in section 6.

2 ADRESS DATASET

The ADReSS dataset (Luz et al., 2020) is designed to provide Alzheimer’s research community with a standard platform for AD detection and MMSE score prediction. The dataset is acoustically pre-processed and balanced in terms of age and gender. It consists of audio recordings and transcriptions (in CHAT format (Macwhinney, 2009)) of the Cookie Theft picture description task, elicited from subjects in the age group of 50-80 years. The training set consists of data from 108 subjects, 54 each from AD and non-AD classes. The test set has data from 48 subjects, again balanced with respect to AD and non-AD classes. More information on the ADReSS dataset can be found in the ADReSS challenge baseline paper (Luz et al., 2020).

3 REVIEW OF BASELINE METHODS

This section provides a brief overview of the various approaches for AD detection and MMSE score prediction on ADReSS dataset. These approaches can be broadly classified into 3 types based on the type of the features used in the problem- i) acoustic feature based, ii) linguistic feature based and iii) a fusion of acoustic and linguistic features. The performance of different approaches on the AD detection and MMSE score prediction tasks are compared using the accuracy and root mean square error (RMSE) measures computed on the ADReSS test set.

$$Accuracy = \frac{TN + TP}{N} \tag{1}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{N}} \tag{2}$$

where N is the total number of subjects involved in the study, TP the number of true positives and TN the number of true negatives. $\hat{y}_i$ and y_i are the estimated and target MMSE scores for i^{th} test sample. The results of different approaches on the ADReSS dataset are summarized in Table 1.

3.1 Acoustic feature-based methods

Luz et al. (2020), explore several acoustic features like extended Geneva minimalistic acoustic parameter set (eGeMAPS) (Eyben et al., 2016), emobase, ComParE-2013 (Eyben et al., 2013), and multi-resolution cochleagram (MRCG) (Chen et al., 2014) feeding the traditional machine learning algorithms like linear discriminant analysis, decision trees, nearest neighbour, random forests and support vector machines. In our previous work (Meghanani et al., 2021), we have used CNN/ResNet + long short-term memory (LSTM) networks and pyramidal bidirectional LSTM + CNN networks trained on log-Mel spectrogram and Mel-frequency cepstral coefficient (MFCC) features extracted from the spontaneous speech. Pompili et al. (2020), exploit the pre-trained models to produce i-vector and x-vector based acoustic feature embeddings. They evaluate x-vector, i-vector, and statistical speech-based functional features. Rhythmic features are proposed in (Campbell et al., 2020), as lower speaking fluency is a common pattern in patients with AD. Koo et al. (2020), use VGGish (Hershey et al., 2017) trained with Audio Set (Gemmeke et al., 2017) for audio classification. They have proposed a modified version of convolutional recurrent neural network (CRNN), where an attention layer is the forefront layer of the network, and fully connected layers follow the recurrent layer.

3.2 Linguistic feature-based methods

Recently, there have been multiple attempts on the AD detection problem based on text based features and models. Searle et al. (2020), use traditional machine learning techniques like support vector machines (SVMs), gradient boosting decision trees (GBDT), and conditional random fields (CRFs). They also try deep learning transformer based models, specifically, bidirectional encoder representations from transformers (BERT) (Devlin et al., 2019), RoBERTa (Liu et al., 2019) and DistilBERT/DistilRoBERTa (Sanh et al., 2019). Pompili et al. (2020), encode each word of the clean transcriptions into 768-dimensional context embedding vector using a frozen English BERT model pre-trained with 12-layers. Three different neural models are trained on top of contextual word embeddings: (i) global maximum pooling, (ii) bidirectional long short-term memory (BLSTM) based recurrent neural networks (RNN) provided with an attention module, and (iii) the second model augmented with part-of-speech (POS) embeddings. In the work of Campbell et al. (2020), authors have used the manual transcripts to extract linguistic information (interventions, vocabulary richness, frequency of verbs, nouns, POS-tagging, etc.) for creating the input features of the classifier. They use another sequential deep learning based classifier, which classifies directly from the sequence of Gobal Vectors (GloVe) based word embeddings. Koo et al. (2020), use transformer (Vaswani et al., 2017) based language models, generative pretraining (GPT) (Radford et al., 2018), RoBERTa (Liu et al., 2019), and transformer-XL (Dai et al., 2020) to get textual features and perform classification and regression tasks using a modified convolutional recurrent neural network based structure.

Graph based representation of word features (Tomás and Radev, 2012), (Cong and Liu, 2014), which have shown promise in classifying texts (De Arruda et al., 2016) are also employed for detection of mild cognitive impairments. Santos et al. (2017) model transcripts as complex networks and enrich them with word embedding to better represent short texts produced in neuro-psychological assessments. They use metrics of topological properties of complex networks in a machine learning classification approach to distinguish between healthy subjects and patients with mild cognitive impairments. Such graph based techniques have also been used in the word sense disambiguation (WSD) tasks to identify the meaning of words in a given context for specific words conveying multiple meanings.(Corra et al., 2018). They suggest that a bipartite network model with local features employed to characterise the context can be useful in improving the semantic characterization of written texts without the use of deep linguistic information.

3.3 Bimodal Methods

Methods with bimodal input features (both acoustic and linguistic) are also used for AD recognition in various studies like (Pompili et al., 2020), (Campbell et al., 2020), (Sarawgi et al., 2020b), (Koo et al., 2020), (Sarawgi et al., 2020a), and (Rohanian et al., 2020). However, in this work, we restrict ourselves to the NLP-based approaches.

4 PROPOSED NLP-BASED METHODS

4.1 Data Preparation

In this work, we explore the linguistic features for AD detection and hence only the textual transcripts in the ADReSS dataset are used. The transcripts contain the conversational content between the participant and the investigator. This include pauses in speech, laughter and discourse markers such as 'um' and 'uh'. Each transcript is considered as a single data point with their corresponding AD label and MMSE score. We create two transcription level datasets after pre-processing the transcripts as in Searle et al. (2020) - 1) PAR: containing the utterances of participant alone, 2) PAR+INV: containing utterances from both the participant and the investigator. In addition to the preprocessing performed in Searle et al. (2020), we keep PAR and INV tags as well in the data (which defines whether the utterance is spoken by the participant or the investigator).

4.2 CNN Model

Language impairments like difficulties in lexical retrieval, loss of verbal fluency, and breakdown in comprehension of higher order written and spoken languages are common in AD patients. Hence the linguistic information like the n-grams present in the input sentence, may provide good cues for AD detection. Any $n \times d$ CNN filter, where n is the number of sequential words looked over by the filter and d is the dimension of word embedding, can be viewed as a feature detector looking for a specific n-gram in the input that can capture the language impairments associated with AD.

We describe the details of the CNN model from the work (Kim, 2014) as follows. Let $z_i \in R^d$ be a d-dimensional word vector corresponding to the i-th word in the sentence. A sentence of length L is represented as $\{z_1, z_2, \dots, z_L\}$. Let $z_{i:i+j}$ represents the concatenation of the words $z_i, z_{i+1}, \dots, z_{i+j}$. A convolution operation involves a filter $w \in R^{nd}$, which is applied to a window of n words to produce a new feature as shown in equation 3, where s_i is generated from a window of words $z_{i:i+n-1}$ by

$$s_i = f(w \cdot z_{i:i+n-1} + b) \tag{3}$$

In the equation 3, f is a non-linear function and b is the bias term. A feature map $\mathcal{S}$ is obtained by applying the filter to all possible windows of words in the sentence $[z_{1:n}, z_{2:n+1}, \ldots, z_{L-n+1:L}]$.

$$\mathcal{S} = [s_1, s_2, ..., s_{L-n+1}] \tag{4}$$

A max-pool over time (Collober et al., 2011) is performed over the feature map to get $s_{max} = \max \mathcal{S}$ as the feature corresponding to that filter. This corresponds to the n-gram that is "most relevant" in the AD recognition task. The weights of the filters, which in turn determine the "most relevant" feature, are learnt using backpropagation. CNNs are trained with just one layer of convolution. Variable length sentences are automatically handled by the pooling scheme. We use pre-trained 100-dimensional GloVe word vectors (Pennington et al., 2014) for word embedding. Multiple kernels of sizes 2×100, 3×100, 4×100 and 5×100 are employed to have a look at the bigrams, trigrams, 4-grams and 5-grams within the text. We use 100 filters each with height 2, 3, 4 and 5. Multiple configurations with filter sizes [2,3,4], [3,4,5] and [2,3,4,5] are applied which are referred to as CNN-bi+tri+4 gram, CNN-tri+4+5 gram, and CNN-bi+tri+4+5 gram in our tables. The outputs of the filter are concatenated together to form a single vector. Dropout with probability $p = 0.5$ is applied on the concatenated filter output and the results are passed through a linear layer for the final prediction task. The linear layer weights up the evidences from each of these n-grams and make a final decision. Fig. 1 shows the basic CNN operation over an example sentence.

4.2.1 Training Details

For the classification task, training is performed for 100 epochs with a batch size of 16. Adam optimizer is used with a learning rate of 0.001. Model with the lowest validation loss is saved and used for prediction. Since AD classification is a two class problem, binary cross entropy with logits loss is used as the loss function. For the MMSE score prediction task, the output layer is a fully connected layer with linear activation function. In the regression task the network is trained for 1500 epochs with the objective to minimize the mean squared error.

We use bootstrap aggregation of models known as bagging Breiman (1996) to predict the final labels/MMSE scores for test samples. Bootstrap aggregation is an ensemble technique to improve the stability and accuracy of machine learning models. It combines the prediction from multiple models. It also reduces variance and helps to avoid overfitting. We fit 21 models and the outputs are combined by a majority voting scheme for final classification. In the regression task, the outputs of these bootstrap models are averaged to arrive at the final MMSE score.

4.3 fastText

fastText based classifiers calculate the n-grams of an input sentence explicitly and append them to the end of the sentence. In this work, we use bigrams and trigrams. We conducted the experiments with 4-grams as well, but the results did not show any improvement over the use of trigrams. This bag of bigrams and trigrams acts as additional features to capture some information about the local word order.

Figure 2 shows the architecture of fastText model. The fastText model has 2 layers, an embedding layer and a linear layer. The embedding layer calculates the word embedding (100-dimensional) for each word. The average of all these word embeddings is calculated and fed through the linear layer for final prediction as described in Fig. 2. fastText models are faster for training and evaluation by many orders of magnitude, compared to the "deep" models. As mentioned in the work (Joulin et al., 2017), fastText can be trained on more than one billion words in less than ten minutes using a standard multicore CPU, and classify half a million sentences among 312K classes in less than a minute.

4.3.1 Training Details

All training details are the same as mentioned in section 4.2.1. The only difference is that dropout is not used in this model. Here also we use 21 bootsrapping models and the outputs are combined as described in section 4.2.1.

5 RESULTS

We have performed 5-fold cross-validation, to estimate the generalization error. One of the folds has 20 validation samples and the remaining four have 22 validation samples. The results of cross-validation on CNN and fastText models trained on PAR and PAR+INV sets are listed in Table 2. The best performing model for classification during the cross validation was fastText with bigrams on the PAR+INV set, which yields an average cross validation accuracy of 86.09%. Among the CNN models, tri+4+5 grams give the best accuracy in both PAR (77.54%) and INV+PAR (81.27%) sets. As far as accuracy is concerned, both the CNN and fastText models seem to benefit with the inclusion of utterances from the investigator. For the prediction of MMSE score, CNN with bi+tri+4+5 grams (RMSE of 4.38) was the best. The fastText models seem to get a clear advantage in RMSE with the addition of the utterances from the investigator. However such a large difference in RMSE is not observable between the CNN models using PAR and INV+PAR sets. The cross-validation results confirmed our belief that the n-grams from the transcriptions of the picture description task could be useful in the detection of AD.

Table 3 lists the classification accuracy and RMSE in the prediction of MMSE score on the test set of the ADReSS corpus. The table also lists the precision, recall and F_1 score for each class. They are computed as precision $\pi = \frac{TP}{TP+FP}$, recall $\rho = \frac{TP}{TP+FN}$, and F_1 score $= \frac{2\pi\rho}{\pi+\rho}$, where TP, FP, TN and FN are the number of true positives, false positives, true negatives and false negatives, respectively. The listed results are obtained after bootstrapping with 21 samples. The best classification accuracy is 83.33% which is achieved using fastText model with appended bigrams and trigrams. The accuracies are similar in both PAR and PAR+INV sets using the fastText model. The maximum accuracy obtained with CNN models is 79.16%, which is achieved on the INV+PAR set using bi+tri+4 grams or tri+4+5 grams. In the detection task, the CNN models seem to get some advantage by the addition of utterances from the investigator. Also the accuracies seem to degrade when bigrams, trigrams, 4-grams and 5-grams are considered together. This behaviour is consistent across the PAR and PAR+INV sets. The best RMSE in the prediction of MMSE score, is 4.28 which is obtained on the PAR+INV set using fastText model employing only bigrams. In the regression task using fastText, the use of bigrams achieve slightly better RMSE compared to the use of both bigrams and trigrams. Also the fastText models seem to benefit from the use of utterances from the investigator. In contrast, CNN models do not seem to get any specific advantage with the inclusion of investigator's utterances. The performance of the CNN models remain almost the same across the use of bi+tri+4, tri+4+5, and bi+tri+4+5 grams.

6 DISCUSSION AND CONCLUSIONS

In this work, we explore two models - CNN with a single convolution layer and fastText, to address the problem of AD classification and prediction of MMSE score from the transcriptions of the picture description task. The choice of these models were based on our initial belief that modeling the transcriptions of the narrative speech in the picture description task using n-grams could give some indication on the status of AD. The chosen models are also shallow. The number of parameters are much less than the usual deep learning architectures and hence they can be trained and evaluated quite fast. Yet, the performance of these models is competitive with the baseline results reported with complex models (refer Table 1). The results suggest that the n-gram based features are worth pursuing, for the task of AD detection.

Among the considered models, fastText model with bigrams and trigrams appended to the input, achieves the best classification accuracy (83.33%). In the regression task, the best results (RMSE of 4.28) are achieved using fastText model with only the bigrams appended to the input. The fastText models have a clear edge over CNN in the classification task. Empirical evidences suggest that fastText models benefit from the inclusion of utterances from the investigator in the regression task, though they do not make much difference in the classification task. The CNN models on the other hand perform better on the PAR+INV sets in the classification task. In the regression task, their performance is similar across the PAR and PAR+INV sets. Bigrams have an edge over bi+tri grams in fastText, when used for prediction of MMSE score. However, the performance of the CNN models remain almost the same across the use of bi+tri+4, tri+4+5, and bi+tri+4+5 grams, in the regression task.

REFERENCES

Breiman, L. (1996). Bagging predictors. *Mach. Learn.* 24, 123–140. doi:10.1007/BF00058655

Campbell, E. L., Docío-Fernández, L., Raboso, J. J., and García-Mateo, C. (2020). Alzheimer's dementia detection from audio and text modalities. *arXiv* 2008.04617

Chen, J., Wang, Y., and Wang, D. (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, 1993–2002. doi:10.1109/TASLP.2014.2359159

Collober, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research* 12, 2493–2537. doi:10.5555/1953048.2078186

Cong, J. and Liu, H. (2014). Approaching human language with complex networks. *Physics of Life Reviews* 11, 598 – 618. doi:https://doi.org/10.1016/j.plrev.2014.04.004

Corra, E. A., Lopes, A. A., and Amancio, D. R. (2018). Word sense disambiguation. *Inf. Sci.* 442, 103–113. doi:10.1016/j.ins.2018.02.047

Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q., and Salakhutdinov, R. (2020). Transformer-XL: Attentive language models beyond a fixed-length context. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* , 2978–2988doi:10.18653/v1/P19-1285

De Arruda, H., Costa, L., and Amancio, D. (2016). Using complex networks for text classification: Discriminating informative and imaginative documents. *EPL* 113. doi:10.1209/0295-5075/113/28007

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *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)* , 4171–4186doi:10.18653/v1/N19-1423

Edwards, E., Dognin, C., Bollepalli, B., and Singh, M. (2020). Multiscale System for Alzheimer's Dementia Recognition Through Spontaneous Speech. *Proc. Interspeech 2020* , 2197–2201doi:10.21437/Interspeech.2020-2781

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., et al. (2016). The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing* 7, 190–202

Eyben, F., Weninger, F., Gross, F., and Schuller, B. (2013). Recent developments in openSMILE, the Munich open-source multimedia feature extractor. *Proceedings of the 2013 ACM Multimedia Conference*

, 835–838doi:10.1145/2502081.2502224

Gemmeke, J., Ellis, D., Freedman, D., Jansen, A., Lawrence, W., Moore, R., et al. (2017). Audio set: An ontology and human-labeled dataset for audio events. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* , 776–780doi:10.1109/ICASSP.2017.7952261

Hershey, S., Chaudhuri, S., Ellis, D., Gemmeke, J., Jansen, A., Moore, R. C., et al. (2017). CNN architectures for large-scale audio classification. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* , 131–135

Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2017). Bag of tricks for efficient text classification. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers* , 427–431

Kim, Y. (2014). Convolutional neural networks for sentence classification. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* , 1746–1751doi:10.3115/v1/D14-1181

Koo, J., Lee, J. H., Pyo, J., Jo, Y., and Lee, K. (2020). Exploiting Multi-Modal Features from Pre-Trained Networks for Alzheimer's Dementia Recognition. *Proc. Interspeech 2020* , 2217–2221doi:10.21437/Interspeech.2020-3153

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., et al. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *ArXiv* abs/1907.11692

Luz, S., Haider, F., de la Fuente, S., Fromm, D., and MacWhinney, B. (2020). Alzheimer's Dementia Recognition Through Spontaneous Speech: The ADReSS Challenge. *Proc. Interspeech 2020* , 2172–2176doi:10.21437/Interspeech.2020-2571

Macwhinney, B. (2009). The CHILDES Project Part 1: The CHAT Transcription Format doi:10.1184/R1/6618440.v1

Meghanani, A., Anoop, C. S., and Ramakrishnan, A. G. (2021). An exploration of log-mel spectrogram and MFCC features for Alzheimer's dementia recognition from spontaneous speech. *Accepted in The 8th IEEE Spoken Language Technology Workshop (SLT), January 19-22, 2021*

Mueller, K. D., Hermann, B., Mecollarib, J., and Turkstra, L. S. (2018a). Connected speech and language in mild cognitive impairment and Alzheimer's disease: A review of picture description tasks. *J Clin Exp Neuropsychol.* 40, 917–939. doi:10.1080/13803395.2018.1446513

Mueller, K. D., Koscik, R. L., Hermann, B., Johnson, S. C., and Turkstra, L. S. (2018b). Declines in connected language are associated with very early mild cognitive impairment: Results from the Wisconsin registry for Alzheimer's prevention. *Frontiers in Aging Neuroscience* 9. doi:10.3389/fnagi.2017.00437

Nicholas, M., Obler, L. K., Albert, M., and Helm-Estabrooks, N. (1985). Empty speech in Alzheimer's disease and fluent aphasia. *Journal of speech and hearing research* 28, 405–410. doi:https://doi.org/10.1044/jshr.2803.405

Pappagari, R., Cho, J., Moro-Velázquez, L., and Dehak, N. (2020). Using State of the Art Speaker Recognition and Natural Language Processing Technologies to Detect Alzheimer's Disease and Assess its Severity. *Proc. Interspeech 2020* , 2177–2181doi:10.21437/Interspeech.2020-2587

Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* , 1532–1543doi:10.3115/v1/d14-1162

Pompili, A., Rolland, T., and Abad, A. (2020). The INESC-ID Multi-Modal System for the ADReSS 2020 Challenge. *Proc. Interspeech 2020* , 2202–2206doi:10.21437/Interspeech.2020-2833

Radford, A., Narasimhan, K., Salimans, T., and Sutskever., I. (2018). Improving language understanding by generative pre-training. *https://www.cs.ubc.ca/ amuham01/LING530/papers/radford2018improving.pdf*

Rohanian, M., Hough, J., and Purver, M. (2020). Multi-Modal Fusion with Gating Using Audio, Lexical and Disfluency Features for Alzheimer's Dementia Recognition from Spontaneous Speech. *Proc. Interspeech 2020* , 2187–2191doi:10.21437/Interspeech.2020-2721

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *ArXiv* abs/1910.01108

Santos, L., Corrêa Júnior, E. A., Oliveira Jr, O., Amancio, D., Mansur, L., and Aluísio, S. (2017). Enriching complex networks with word embeddings for detecting mild cognitive impairment from speech transcripts. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* , 1284–1296doi:10.18653/v1/P17-1118

Sarawgi, U., Zulfikar, W., Khincha, R., and Maes, P. (2020a). Uncertainty-aware multi-modal ensembling for severity prediction of Alzheimer's dementia. *ArXiv* abs/2010.01440

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

Savundranayagam, M., Hummert, M. L., and Montgomery, R. (2005). Investigating the effects of communication problems on caregiver burden. *The journals of gerontology. Series B, Psychological sciences and social sciences* 60 1, S48–55

Searle, T., Ibrahim, Z., and Dobson, R. (2020). Comparing natural language processing techniques for Alzheimer's dementia prediction in spontaneous speech. *Proc. Interspeech 2020* , 2192–2196doi:10.21437/Interspeech.2020-2729

Shen, Y., He, X., Gao, J., li Deng, and Mesnil, G. (2014). Learning semantic representations using convolutional neural networks for web search. *WWW 2014* , 373–374doi:10.1145/2567948.2577348

Syed, M. S. S., Syed, Z. S., Lech, M., and Pirogova, E. (2020). Automated Screening for Alzheimer's Dementia Through Spontaneous Speech. *Proc. Interspeech 2020* , 2222–2226doi:10.21437/Interspeech.2020-3158

Szatlóczki, G., Hoffmann, I., Vincze, V., Kálmán, J., and Pákáski, M. (2015). Speaking in Alzheimer's disease, is that an early sign? Importance of changes in language abilities in Alzheimer's disease. *Frontiers in aging neuroscience* 7. doi:10.3389/fnagi.2015.00195

tau Yih, W., He, X., and Meek, C. (2014). Semantic parsing for single-relation question answering. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* , 643–648doi:10.3115/v1/P14-2105

Tomás, D. R. M. and Radev, D. (2012). Graph-based natural language processing and information retrieval. *Machine Translation* 26, 277–280. doi:10.1007/s10590-011-9122-9

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., et al. (2017). Attention is all you need. *Proceedings of the 31st International Conference on Neural Information Processing Systems* 2017-December, 5999–6009

Yuan, J., Bian, Y., Cai, X., Huang, J., Ye, Z., and Church, K. (2020). Disfluencies and Fine-Tuning Pre-Trained Language Models for Detection of Alzheimer's Disease. *Proc. Interspeech 2020* , 2162–2166doi:10.21437/Interspeech.2020-2516

Table 1. Baseline methods on ADReSS test set

Model	Accuracy	RMSE
(Searle et al., 2020), DistilBERT	81.25%	4.58
(Searle et al., 2020), SVM+CRF	81.25%	5.22
(Pompili et al., 2020), x-vectors SRE	54.17%	–
(Pompili et al., 2020), Sentence embedding	72.92%	–
(Pompili et al., 2020), Fusion of system	81.25%	–
(Luz et al., 2020), linguistic	75.00%	5.20
(Sarawgi et al., 2020b), Ensemble	83.33%	4.60
(Koo et al., 2020), VGGish	72.92%	5.07
(Koo et al., 2020), Transformer-XL	81.25%	4.01
(Koo et al., 2020), VGGish+GloVe	77.08%	4.33
(Koo et al., 2020), VGGish+Transformer-XL	75.00%	3.74
(Koo et al., 2020), Ensembled Output	81.25%	3.77
(Campbell et al., 2020), Fusion II	75.00%	–
(Campbell et al., 2020), Fusion I	72.92%	–
(Campbell et al., 2020), RNN Model	75.00%	–
(Campbell et al., 2020), fluency	60.42%	–
(Campbell et al., 2020), x-vector	54.17%	–
(Sarawgi et al., 2020a), UA Ensemble	–	4.35
(Sarawgi et al., 2020a), UA Ensemble (weighted)	–	3.93
(Pappagari et al., 2020), Acoustic and Transcript	75.00%	5.37
(Rohanian et al., 2020), LSTM (Lexical+Dis)	72.92%	4.88
(Rohanian et al., 2020), LSTM with Gating (Acoustic+Lexical)	77.08%	4.57
(Rohanian et al., 2020), LSTM with Gating (Acoustic+Lexical+Dis)	79.17%	4.54
(Yuan et al., 2020), ERNIE3p	89.58%	–
(Syed et al., 2020)	85.42%	4.30
(Edwards et al., 2020), Phonemes and Audio	79.17%	–
(Meghanani et al., 2021), CNN-LSTM with MFCC	64..58%	6.24
(Meghanani et al., 2021), pBLSTM-CNN with log-Mel	52.08%	5.90
(Meghanani et al., 2021), ResNet-LSTM with log-Mel	62.50%	5.98

Table 2. Average 5-fold cross-validation results for AD classification and RMSE values

Dataset	Model	Accuracy	RMSE
PAR	CNN, bi+tri+4 gram	73.91%	4.55
PAR	CNN, tri+4+5 gram	77.54%	4.41
PAR	CNN, bi+tri+4+5 gram	76.54%	4.65
PAR	fastText, bigram	80.54%	5.43
PAR	fastText, bi+trigram	82.36%	5.40
PAR+INV	CNN, bi+tri+4 gram	80.18%	4.63
PAR+INV	CNN, tri+4+5 gram	81.27%	4.53
PAR+INV	CNN, bi+tri+4+5 gram	80.36%	4.38
PAR+INV	fastText, bigram	86.09%	4.66
PAR+INV	fastText, bi+trigram	85.90%	4.81

Table 3. Results on ADReSS test set

Dataset	Model	Class	Precision	Recall	F1 Score	Accuracy	RMSE
PAR	CNN, bi+tri+4 gram	Non-AD	0.74	0.71	0.72	72.91%	4.38
		AD	0.72	0.75	0.73		
PAR	CNN, tri+4+5 gram	Non-AD	0.76	0.67	0.71	72.91%	4.46
		AD	0.70	0.79	0.75		
PAR	CNN, bi+tri+4+5 gram	Non-AD	0.71	0.71	0.71	70.83%	4.42
		AD	0.71	0.71	0.71		
PAR	fastText, bigram	Non-AD	0.78	0.88	0.82	81.25%	4.51
		AD	0.86	0.75	0.80		
PAR	fastText, bi+trigram	Non-AD	0.81	0.88	0.84	**83.33%**	4.87
		AD	0.86	0.79	0.83		
PAR+INV	CNN, bi+tri+4 gram	Non-AD	0.77	0.83	0.80	79.16%	4.48
		AD	0.82	0.75	0.78		
PAR+INV	CNN, tri+4+5 gram	Non-AD	0.77	0.83	0.80	79.16%	4.47
		AD	0.82	0.75	0.78		
PAR+INV	CNN, bi+tri+4+5 gram	Non-AD	0.74	0.71	0.72	72.91%	4.44
		AD	0.72	0.75	0.73		
PAR+INV	fastText, bigram	Non-AD	0.78	0.88	0.82	81.25%	**4.28**
		AD	0.86	0.75	0.80		
PAR+INV	fastText, bi+trigram	Non-AD	0.79	0.92	0.85	**83.33%**	4.47
		AD	0.90	0.75	0.82		

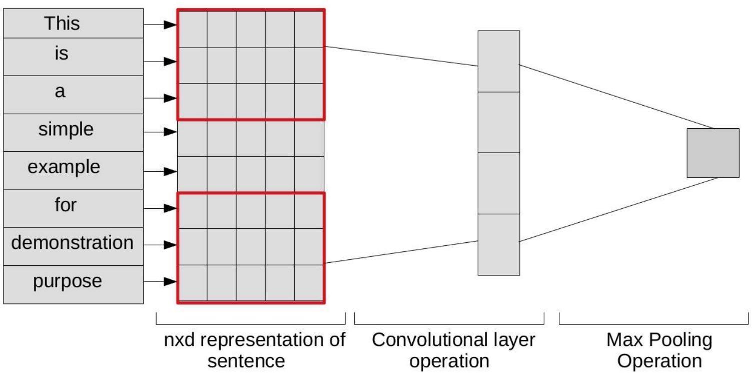

Figure 1. Demonstration of CNN over text for an example sentence.
.

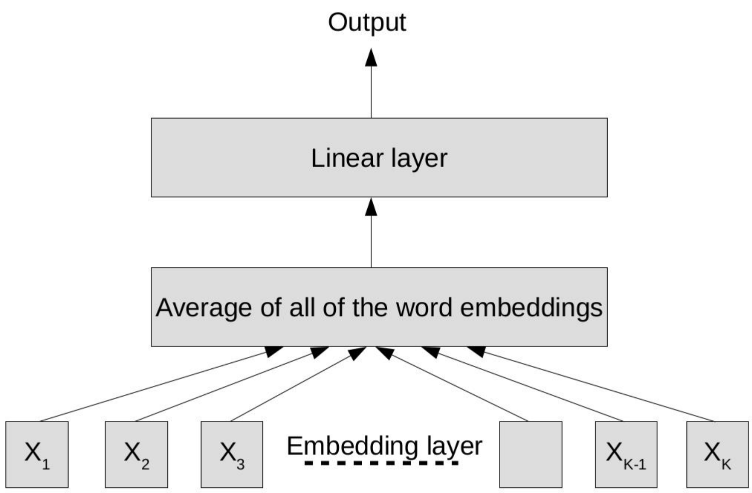

Figure 2. fastText model (Joulin et al., 2017) with appended n-gram features $(X_1, X_2, X_3, ..., X_{K-1}, X_K)$ as input.
.