Accurate Detection of Dementia from Speech Transcripts Using RoBERTa Model

Lovro Matošević, Alan Jović
University of Zagreb, Faculty of Electrical Engineering and Computing
Unska 3, 10000 Zagreb, Croatia
lovromatosev@gmail.com, alan.jovic@fer.hr

***Abstract*—Dementia is a serious disease that is very common in the elderly population. Automatic detection of dementia is a difficult task that may involve the analysis of acoustic features of speech, linguistic features of transcripts, and mental state exams. In this work, we explore the limits of using speech transcripts from doctor-patient conversations to detect dementia. The dataset is prepared from Pitt corpus, which is a part of DementiaBank, a shared database of multimedia interactions for studying communication in dementia. We use a sophisticated natural language processing approach, namely RoBERTa, which addresses the problem by using transformers and self-attention mechanism. We compare RoBERTa with a baseline BERT model. We show that dementia detection using well-prepared speech transcripts alone can lead to detection rates above 90% for RoBERTa model in a near-balanced dataset, outperforming the baseline model.**

***Keywords*—dementia, detection, transformers, RoBERTa, deep learning, Pitt corpus.**

I. Introduction

Dementia is the loss of cognitive functioning – thinking, remembering, and reasoning – to such an extent that it interferes with a person's daily life and activities [1]. It is a serious disease that is usually chronic and progressive, which is a major cause of disability in the elderly population. It poses a significant obstacle not only to those with the disease, but also to their careers and families. More than 50 million people in the world suffer from dementia and it is estimated that the number of cases will triple by 2050 [2].

Dementia is not only a serious global health problem, but also represents a significant economic burden. According to some estimates, the economic impact of dementia is more than $1 trillion, which is more than 1% of the world's gross domestic product [3]. At the time of writing, there is no cure for dementia, nor is there a standardized test to detect dementia. Because dementia is currently incurable, a medical expert's goal is to diagnose dementia as early as possible so that patients can receive appropriate treatment that can alleviate their symptoms and slow cognitive decline, potentially leading to extended life expectancy. All of this forms a motivation for the development of an effective dementia detection tool that could ideally be used in conjunction with some sort of standardized test.

In this work, we explore the limits of using speech transcripts from doctor-patient conversations to detect dementia using a sophisticated natural language processing (NLP) approach, namely RoBERTa, a transformer model that uses a self-attention mechanism. We decided to use the transformer approach, as it has been shown that an architecture using only the attention mechanism can achieve impressive results in certain NLP tasks [4]. Moreover, we use BERT, a precursor of RoBERTa, as a baseline model and compare the results obtained using BERT and RoBERTa. The aforementioned speech transcripts are a part of the Pitt corpus [5], which consists of audio recordings of the Cookie Theft picture test and associated transcripts created by expert linguists. Compared to related work, we achieve better dementia detection results by carefully considering both the optimization of RoBERTa and the information obtainable from transcripts.

In the following sections, we first briefly review related work and their results on dementia detection. Furthermore, we describe our dataset and give a brief introduction to RoBERTa. Finally, we describe our experimental setup, present our results and give a concise conclusion.

II. Related Work

There are a number of studies on the detection of dementia. Although most of them share some similarities, such as the use of the cookie test experiments from the Pitt corpus, they all differ in some ways. In continuation of this section, we explain the different approaches.

A. Feature-Based Approaches

The first approaches to diagnosing dementia from speech were feature-based [6]. A classifier was trained using numerous linguistic features such as verb rate, noun rate, pronoun rate, and others. Further improvements of these feature-based approaches were based on increasing the number of expert-defined features [7], adding acoustic features [8], linking linguistic features to neuropsychological tests [9] and so on. Rather than using features introduced by human experts, some studies focused on clustering pre-trained GloVE embeddings of participants' words to detect dementia [10], [11].

B. Deep Learning-Based Approaches

In recent years, there is an increase in the number of research papers that use deep learning-based approaches instead of feature-based approaches. The first work based on deep learning was by Orimaye et al. [12], in which a deep neural network was trained to predict mild cognitive impairment based on speech. In subsequent studies, convolutional neural

networks (CNNs) [13] and recurrent neural networks with long short-term memory (LSTM) [14] were mainly used. Following the recent success of transformer models, several studies have been published using transformers for dementia detection. Most of them use BERT or RoBERTa methods. We give a brief introduction to RoBERTa and transformer methods in Section IV.

Table I shows a comparison of related works and their results. It includes details such as the dataset, modeling method, and evaluation method used for each of the mentioned studies. In addition, the accuracy and F1 score are given for each work. We would like to caution the reader that all results listed in Table I should be taken with a grain of salt. Namely, it is difficult to directly compare all the studies listed in the table. First, most of the studies reported do not specify exactly how the transcripts provided were preprocessed, although there are some exceptions. Namely, Yancheva and Rudzicz removed all of the filled pauses and repeated speech portions [10]. Karlekar et al. split all of the transcripts into individual utterances and they removed all utterances without POS tags [13]. Ilias and Askounis removed all of the phonological fragments and non-standard forms from transcripts, as well as repeated speech portions and filled pauses [15]. All other mentioned related works do not explicitly state in which way they prepared the data. Furthermore, most of the related works do not specify whether they split the dataset per patient or per sample. The latter would mean that they have a data leak, as it would be possible that the same patient ends up in both the training set and the test set. In addition, not all the studies use the same evaluation method. Finally, some studies have chosen to use samples from participants with mild cognitive impairment, while others have discarded them. Therefore, the results may not be directly comparable. Although RoBERTa was already used for detection of dementia by Jonasson and Wahlforss [16], they used out-of-the-box parameters. We show that optimized parameters as well as longer training can lead to better results. Additionally, our work is the first one to experiment with different ways of preprocessing Pitt corpus' transcripts. The first experiment, similar to the work of Ilias and Askounis [15], removes most of the additional information provided in the transcripts. In our second and third experiments, we add repeated speech portions and filler words, respectively.

III. Dataset

One of the major challenges in training a model to accurately detect dementia using speech is the lack of a large dataset. At the time of writing, the largest dataset available is the Pitt corpus from DementiaBank [5], which we used in our work. The Pitt corpus consists of audio recordings of the Cookie Theft picture test and associated transcripts created by expert linguists. The Cookie Theft picture test is one of many methods used to assess dementia, more specifically to assess a patient's verbal and cognitive abilities. Patients are given the task of describing everything they see in a picture shown in Figure 1. The picture shows a mother washing dishes while children try to steal cookies from a jar, hence the name Cookie Theft. The picture also contains elements and information from different semantic categories. Healthy participants are able to attend to, and perceive, every aspect of the picture. Participants with neurological impairments may present executive neurological function deficits in which a range of cognitive skills including attention, memory, and planning are compromised. These participants may not recall previously describing a part of the scene and may describe it for a second or third time. As a result, their description often contains repetitive language that contributes no new information. If cognitive skills such as planning and organization are compromised, adults with dementia may be unable to convey information in a logical order or present a coherent description of the scene. As a consequence, the description may appear fragmented and disorganized [21].

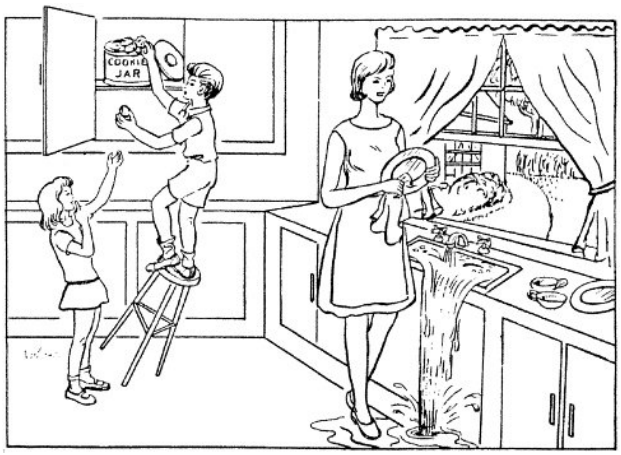

Fig. 1. Cookie Theft picture, adopted from Goodglass et. al. [22]

All of the samples included in the Pitt corpus are from a large study of Alzheimer's dementia conducted between 1983 and 1988. In order to participate in the study, participants had to meet certain criteria, such as having no previous cognitive impairment and not taking medications that affect the central nervous system. A total of 292 participants took part in the study. Of these 292 participants, 98 were in the control group, while the other 194 participants were classified as having MCI (mild cognitive impairment), or definite or probable dementia. After subtracting the participants who were classified as MCI, a total of 177 participants remained in the dementia group. Each participant was tested one to three times with the Cookie Theft picture test. Accordingly, there are 242 samples from patients in the control group and 267 samples in the dementia group, excluding the participants diagnosed with MCI, resulting in a total of 509 audio samples and their associated linguistic transcripts. Detailed dataset statistics are shown in Table II.

For our work, we have decided to use the linguistic transcripts provided. These transcripts are provided in the CHAT format [23]. CHAT is a format used in all TalkBank datasets and contains additional information such as morphological information and grammatical relationships in addition to the

TABLE I
COMPARISON OF RELATED WORK ON DEMENTIA DETECTION

Dataset	Model	Validation	Accuracy, %	F1, %	Reference
ADReSS Challenge [17]	BERT	Stratified 10-fold CV	85.56%	85.43%	Ilias & Askounis [15]
Pitt corpus	CNN-RNN	-	84.9%	-	Karlekar et.al. [13]
Pitt corpus	LSTM	-	83.7%	-	Karlekar et.al. [13]
Pitt corpus	CNN-LSTM	Leave-One-Out CV	85.6%	-	Fritsch et.al. [14]
Pitt corpus	SVM	10-fold CV	78%	82%	Hernandez-Dominguez et.al. [18]
Pitt corpus	RoBERTa (512 maximum input length)	10-fold CV	86.72%	-	Jonasson, Wahlforss [16]
Pitt corpus	S-BERT Large LR	10-fold CV	88.08%	87.23%	Roshanzamir et.al. [19]
ADReSS Challenge [17]	BERT	5-fold CV	80%	74%	Saltz et.al. [20]
Pitt Corpus	BERT	5-fold CV	80%	74%	Saltz et.al. [20]

TABLE II
DATASET STATISTICS

Number of participants	Number of transcripts	Average number of transcripts per patient	Average number of characters per transcript	Average number of words per transcript
275	509	1.85	521.42	107.09

transcribed speech of the participants. For more information on how the dataset samples were used and how the transcripts were processed, see Section V.

IV. ROBERTA

The publication of "Attention Is All You Need" [4], which presented a novel architecture called a "transformer" along with a mechanism called attention, represented a new paradigm for the field of natural language processing. Until that time, recurrent neural networks had been the best way to represent temporal dependencies in sequences. However, it has since been shown that an architecture using only the attention mechanism can achieve impressive results in certain NLP tasks [4]. Transformer is essentially an encoder-decoder architecture. The encoder receives a list of vectors as input. It processes these vectors by passing them into a self-attention layer and then into a feed-forward neural network, which sends the output upward to the next encoder. This process is repeated until the final encoder is reached. The output of the top encoder is then transformed into a set of attention vectors K and V, which are used by each of the decoders to focus on the appropriate locations in the input sequence. The output of each step is passed to the lowest decoder in the next time step, and similarly to the encoders, as they bubble up their decoding results. The last layer is followed by a fully connected layer and a softmax layer that provides an output token. The given input sequence is fed into the encoder stack only once, and the outputs are then fed into the decoder stack at each time step. This process is repeated until the transformer produces an end-of-sequence token. The number of encoders and decoders varies from transformer to transformer. BERT-base and RoBERTa-base models which we use in this work both have 12 encoders/decoders The architecture with six encoders/decoders was used in "Attention Is All You Need". For a more detailed explanation, see [4].

One of the main advantages that came with the introduction of transformers was the ability to train extensive pre-trained models. Through sophisticated pre-training goals and huge model parameters, large-scale pre-trained models can successfully capture knowledge from vast amounts of labeled and unlabeled data. By accumulating knowledge in a huge number of parameters and fine-tuning them for specific tasks, the knowledge implicitly encoded in the model can be useful for a variety of downstream tasks [24]. The first and best known pre-trained model that uses transformers is BERT [25], which was originally pre-trained on the entire English Wikipedia and Brown corpus, outperformed state-of-the-art results on a number of NLP tasks, including all major text classification tasks.

Following the success of BERT, a number of novel architectures using transformers were developed, and the corresponding pre-trained models were made available. One of these models was RoBERTa. In "RoBERTa: A Robustly Optimized BERT Pretraining Approach," it was shown that BERT was significantly undertrained [26]. The authors of RoBERTa modified the BERT training in a number of ways. Firstly, they used longer training times and larger training data - 160 GB in contrast to 16 GB used for the training of BERT. Secondly, they notably increased the batch size from BERT's 256 to 8000. Furthermore, the task of next sentence prediction was removed, while the masked language modeling objective was altered so that the masking was done during training in contrast to BERT performing masking only once at data preparation time. Finally, longer sequences were used as input, but the limitation of 512 tokens was kept. It is important to emphasize that RoBERTa essentially uses the same architecture as BERT.

We believe that a pre-trained model like RoBERTa has many advantages for the NLP task at hand, especially because the Pitt corpus has a dataset that is, by all accounts, quite small compared to the one on which RoBERTa was pre-trained.

V. Experimental Setup

In this section, we describe in detail our experimental setup, so that reproducible results can be attained.

A. Data Preprocessing

As mentioned earlier, the transcripts are written in the CHAT format. The first preprocessing step was to extract all the speech of the participants. Therefore, we discarded the examiners' speech and all information about morphological and grammatical relations in the transcripts.

Next, we had to decide which participant speech information to keep and which to discard. Let us first briefly discuss the content of the transcripts. The transcripts contain not only most of the participant's spoken words, but also additional information, such as simple events which are preceded by: *&=*. So, for example, if *&=coughs* is in a transcript, it means that the participant coughed at that exact moment. Furthermore, the transcripts contain a variety of filled pauses such as *uh*, *um*, *er*, and so on. These pauses are preceded by the ampersand and hyphen mark, *&-*. Finally, the transcripts also contain portions of speech that the participant repeated, retraced or reformulated, which is appropriately indicated by the marks *[/]*, *[//]* and *[///]*.

We decided to create three different experiments. The first experiment would consist of transcripts stripped of all previously mentioned information. To further clarify, the transcripts in our first experiment contain the participant's speech without filled pauses or repeated speech. The second experiment consists of transcripts with only the filled pauses removed. The third and final experiment involves the use of transcripts that contain both the filled pauses and the portions containing all of the repeated, retraced and reformulated speech. Any additional information not mentioned in this paragraph was removed from all transcripts. An example of two preprocessed transcripts can be found in Table III. The red-colored text indicates filled pauses, which were used only in our third experiment. The blue-colored text indicates repeated, retraced or reformulated speech, which was included in both the second and the third experiment. The first experiment did not include any of the colored text from Table III.

B. Training Setup

Because, as mentioned earlier, each patient participated in the Cookie Theft picture test between one and three times, we decided to group the samples per patient. This was done to avoid training and testing the model with the same patients. To further illustrate, let us assume that the patient John Doe participated in the test three times. Consequently, the dataset would contain three samples of John Doe's test. When training or testing our model, we would ensure that all three samples of John Doe are included only in the training or the test set.

In addition, our dataset is not perfectly balanced. There are 242 samples in the control group, which is 47.54% of the total dataset. In addition, the grouping of samples per patient added to the imbalance, as there were only 98 participants in the control group, accounting for 35.64% of all patients. To reduce bias as much as possible, we decided to use stratified 10-fold cross-validation. Therefore, 275 patients were divided into 10 groups, resulting in an average of 27 patients per group.

The basic versions of RoBERTa and BERT, available in the HuggingFace Transformers library [27], were used. We experimented with both a maximum length of 256 and 512 for the RoBERTa and BERT tokenizers. A batch size of 16 was used in all experiments. The implementation of the AdamW optimizer from the HuggingFace transformer library was used along with the following parameters: β_1 = 0.9, β_2 = 0.999, ϵ = 1e-8, and L2 weight decay of 0.01. The learning rate was set to 5e-6. We also experimented with the MADGRAD optimizer [28] in our preliminary experiments. Although it showed promising results in terms of faster convergence than AdamW, we still chose AdamW because there is much more available literature for it. For each fold, a new model was trained for 30 epochs. In addition, accuracy, precision, recall, and F1 score were recorded for each model and later averaged across all models in the 10-fold cross-validation.

VI. Results and Discussion

The results of our experiments can be seen in Tables V and VI. As expected, all of the experiments with RoBERTa outperformed the ones with BERT. Not surprisingly, any RoBERTa model with the maximum length parameter set to 512 performed better than the corresponding RoBERTa model with a maximum length of 256, allowing us to conclude that our models benefit from a larger text span. Another detail of particular interest is that all trained models tend to produce more false negatives than false positives, as can be seen from the fact that all models have higher precision than recall values. We suspect that this is the result of a partial imbalance in the dataset.

To our surprise, the highest accuracy and F1 score were obtained in the second experiment, both by RoBERTa and BERT. From this, we can infer two things. First, repeated speech is characteristic to dementia, it is one of the symptoms of cognitive decline, therefore including it in the training set improves model's accuracy. The results of the first experiment support this hypothesis, as the accuracy was lower when the repeated speech portions from the transcripts were removed. Moreover, from the results of the third experiment we can observe that including filled pauses does not lead to better results in dementia detection. Again, accuracy was noticeably lower after we trained the models with transcripts containing both filled pauses and repeated speech. The results of the first and the third experiments were comparable, with only a slight advantage in accuracy and F1 score for the third experiment.

All of the models failed to correctly classify 20 samples. Of these 20 samples, 13 belong to the dementia group and the other 7 to the control group. In Table IV, we show one sample always misclassified from the dementia group and one from the control group. Again, the blue coloured text indicates parts of the transcript that were used in both the second and third experiments, while the red coloured text indicates parts of the transcript that were used only in the third experiment.

TABLE III

TRANSCRIPT EXAMPLE FOR DIFFERENT PREPROCESSING STEPS IN EACH OF THE EXPERIMENTS. BLUE TEXT, WHICH INDICATES REPEATED, RETRACED OR REFORMULATED SPEECH WAS USED IN THE SECOND AND THE THIRD EXPERIMENT, RED TEXT, WHICH INDICATES FILLED PAUSES WAS USED ONLY IN THE THIRD EXPERIMENT.

Transcript	Label
mhm oh I see a part of the whole kitchen is that all the kitchen or isn't it uh oh I can't read a lady a mother were in her kitchen in her kitchen doing some work I suppose and the uh there's another woman there sharing their pleasures or whatever oh have you have you checked heard of that new game that they started to play after christmas did you is a well it looks like I'd say this is well let's see it looks like oh my mother will beat me by my wife will beat me by a couple rows of this that's that's like the washing would say washing machine or let me see I can't oh that's the son come out of from school maybe or something that's a youngster there well that's just as though they getting ready to go to school or they're just coming out from school and right there he's uh same as back there except for down there in the bottom I think it's uh that's a little	Dementia
okay uh the child's falling off the chair he's taking cookies out of the jar the girl is standing on the floor uh asking for a cookie the door to the cabinet door is open mother is washing dishes the sink is overflowing the water's running uh I don't know if she's dryin em or washin em anyway and the kitchen window has curtains the window's open um it looks like a view of the back there are three dishes on the uh counter	Control

TABLE IV

TWO TRANSCRIPTS WHICH WERE CLASSIFIED INCORRECTLY IN ALL EXPERIMENTS. BLUE TEXT, WHICH INDICATES REPEATED, RETRACED OR REFORMULATED SPEECH WAS USED IN THE SECOND AND THE THIRD EXPERIMENT, RED TEXT ,WHICH INDICATES FILLED PAUSES WAS USED ONLY IN THE THIRD EXPERIMENT.

Transcript	Description
okay I'll start the mother is drying dishes and the sink is over flowing the water is falling onto the floor uh the boy is on his stool uh taking cookies out of a cookie jar and he has one cookie two cookies one in each hand the uh girl is standing reaching up for a cookie with her uh finger over her mouth telling him to be quiet the stool is on one leg uh there's drapes on the window there's a path uh between the grass and the bushes and this little picture is a part of the house and part of the tree in the upper window there are uh uh doors on the uh cabinets in the sink and uh it's daylight um there's two cups and a dish on it on the sink should I describe the two faucets	One of the examples from dementia group which all of the models classified incorrectly
well the boy's trying to get in this cookie jar and the stool overturns and uh the little girl is expecting to hand her a cookie uh the mother is her sink is running over and she's standing in some of the water and uh she's drying a dish or wiping a dish and uh you said everything is happening well the water is still runnin in the sink and I said it's it's overflowing and she's standing in the water and that's I guess look somebody laying in the lawn out there but I can't uh	One of the examples from control group which all of the models classified incorrectly

We would like to point out that the split between training and testing may have significantly affected the reliability of our results, as the data set is small and somewhat unbalanced. Two datasets most commonly used in related studies were the Pitt corpus, which we also use, and the ADReSS challenge dataset [17]. The ADReSS Challenge dataset is a subset of the Pitt corpus and is perfectly balanced with 78 subjects belonging to the dementia group and 78 subjects belonging to the control group. It is difficult to compare our results with those of related studies because, as mentioned earlier, some of them use the ADReSS challenge dataset, whereas most of the others, which use the Pitt corpus, do not specify whether they included participants with MCI and how they preprocessed the transcripts. The results obtained by our RoBERTa model outperform the results which Jonasson and Wahlforss [16] obtained using a RoBERTa model on the same task. They report an accuracy of 86.82% and precision of 90.69%, while our best performing model obtained an accuracy of 90.16% and precision of 92.81%. These differences in scores could be due to a number of things. Firstly, it could be due to them using the out-of-the-box parameters and not performing any optimization. Secondly, since they do not specify in which way they preprocessed the transcripts, their accuracy could be a result of removing the repeated speech portions, for which we have shown that they clearly affect the end results. Finally, as previously indicated, it could be due to the split between training and test sets.

VII. CONCLUSION AND FUTURE WORK

In this study, we have shown that pre-trained transformer models such as RoBERTa can achieve significant results in detecting dementia from speech transcripts. Our best model trained on transcripts that contained repeated parts of speech achieved 90.16% accuracy. These results are promising and motivate the development of a standardized test for dementia detection that could be used in conjunction with a model similar to the one we fine-tuned in this study. Currently, one of the major obstacles to the detection of dementia is the lack of a larger data set. We hope that studies such as this one can encourage the creation of larger, standardized datasets, as this could lead to the development of a truly powerful dementia detection tool that could help in delaying the onset of severe dementia, helping millions of people and significantly extending their lifespan.

TABLE V
EXPERIMENT RESULTS - BERT

	First Experiment	First Experiment	Second Experiment	Second Experiment	Third Experiment	Third Experiment
Model	BERT256	BERT512	BERT256	BERT512	BERT256	BERT512
Precision	**91.86%**	90.17%	88.78%	90.55%	89.84%	90.76%
Recall	80.08%	84.44%	**85.58%**	81.42%	83.27%	82.89%
F1 score	84.99%	86.61%	**86.89%**	85.34%	85.76%	86.36%
Accuracy	85.29%	86.26%	**86.42%**	85.03%	85.22%	86.29%

TABLE VI
EXPERIMENT RESULTS - ROBERTA

	First Experiment	First Experiment	Second Experiment	Second Experiment	Third Experiment	Third Experiment
Model	ROBERTA256	ROBERTA512	ROBERTA256	ROBERTA512	ROBERTA256	ROBERTA512
Precision	**94.26%**	93.46%	90.21%	92.81%	92.88%	91.87%
Recall	80.31%	83.30%	87.27%	**88.60%**	83.46%	86.09%
F1 score	86.31%	87.75%	88.27%	**90.28%**	87.27%	88.49%
Accuracy	86.93%	87.76%	87.74%	**90.16%**	87.22%	88.16%

In the future, we would like to explore a couple of research avenues. First and foremost, we would like to experiment with models that detect dementia directly from patient speech by classifying audio spectrograms. We also want to explore the potential of some newer transformer models such as XLNet [29] and the use of a Longformer [30] that would allow us to process language sequences with more than 512 tokens.

REFERENCES

[1] H. Chertkow, H. H. Feldman, C. Jacova, and F. Massoud, "Definitions of dementia and predementia states in Alzheimer's disease and vascular cognitive impairment: consensus from the Canadian conference on diagnosis of dementia," *Alzheimer's research & therapy*, vol. 5, no. 1, pp. 1–8, 2013.

[2] E. Nichols, J. D. Steinmetz, S. E. Vollset, *et al.*, "Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: An analysis for the global burden of disease study 2019," *The Lancet Public Health*, vol. 7, no. 2, E105–E125, Feb. 2022.

[3] J. Xu, Y. Zhang, C. Qiu, and F. Cheng, "Global and regional economic costs of dementia: A systematic review," *The Lancet*, vol. 390, S47, 2017, SI: The Lancet-CAMS Health Summit, 2017, ISSN: 0140-6736. DOI: https://doi.org/10.1016/S0140-6736(17)33185-9. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0140673617331859.

[4] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is All you Need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, *et al.*, Eds., vol. 30, Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[5] J. T. Becker, F. Boller, O. L. Lopez, J. Saxton, and K. L. McGonigle, "The natural history of Alzheimer's disease. Description of study cohort and accuracy of diagnosis," *Arch Neurol*, vol. 51, no. 6, pp. 585–594, Jun. 1994.

[6] R. S. Bucks, S. Singh, J. M. Cuerden, and G. K. Wilcock, "Analysis of spontaneous, conversational speech in dementia of Alzheimer type: Evaluation of an objective technique for analysing lexical performance," *Aphasiology*, vol. 14, no. 1, pp. 71–91, 2000. DOI: 10.1080/026870300401603. eprint: https://doi.org/10.1080/026870300401603. [Online]. Available: https://doi.org/10.1080/026870300401603.

[7] S. O. Orimaye, J. S.-M. Wong, and K. J. Golden, "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*, Baltimore, Maryland, USA: ACL, Jun. 2014, pp. 78–87. DOI: 10.3115/v1/W14-3210. [Online]. Available: https://aclanthology.org/W14-3210.

[8] A. König, A. Satt, A. Sorin, *et al.*, "Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease," *Alzheimers Dement. (Amst.)*, vol. 1, no. 1, pp. 112–124, Mar. 2015.

[9] E. Eyigoz, S. Mathur, M. Santamaria, G. Cecchi, and M. Naylor, "Linguistic markers predict onset of Alzheimer's disease," *EClinicalMedicine*, vol. 28, no. 100583, p. 100583, Nov. 2020.

[10] M. Yancheva and F. Rudzicz, "Vector-space topic models for detecting Alzheimer's disease," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Berlin, Germany: ACL, Aug. 2016, pp. 2337–2346.

DOI: 10.18653/v1/P16-1221. [Online]. Available: https://aclanthology.org/P16-1221.

[11] K. Sirts, O. Piguet, and M. Johnson, "Idea density for predicting Alzheimer's disease from transcribed speech," in *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, R. Levy and L. Specia, Eds., ACL, 2017, pp. 322–332. DOI: 10.18653/v1/K17-1033.

[12] S. O. Orimaye, J. S.-M. Wong, and J. S. G. Fernandez, "Deep-deep neural network language models for predicting mild cognitive impairment," in *Second international workshop on Advances in Bioinformatics and Artificial Intelligence: Bridging the Gap (BAI 2016)*, A. Baniré Diallo, E. Nguifo, and M. Zaki, Eds., ser. CEUR Workshop Proceedings, BAI 2016, 11-07-2016, Rheinisch-Westfaelische Technische Hochschule Aachen, 2016, pp. 14–20. [Online]. Available: http://ceur-ws.org/Vol-1718/.

[13] S. Karlekar, T. Niu, and M. Bansal, "Detecting Linguistic Characteristics of Alzheimer's Dementia by Interpreting Neural Models," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, New Orleans, Louisiana: ACL, Jun. 2018, pp. 701–707. DOI: 10.18653/v1/N18-2110. [Online]. Available: https://aclanthology.org/N18-2110.

[14] J. Fritsch, S. Wankerl, and E. Nöth, "Automatic Diagnosis of Alzheimer's Disease Using Neural Network Language Models," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 5841–5845. DOI: 10.1109/ICASSP.2019.8682690.

[15] L. Ilias and D. Askounis, "Explainable Identification of Dementia from Transcripts using Transformer Networks," *CoRR*, vol. abs/2109.06980, 2021. arXiv: 2109.06980. [Online]. Available: https://arxiv.org/abs/2109.06980.

[16] A. Aslaksen Jonasson and A. Wahlforss, *Diagnosis of dementia using transformer models*, 2020. [Online]. Available: http://kth.diva-portal.org/smash/record.jsf?dswid=-4861&pid=diva2%3A1458824.

[17] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, *Alzheimer's Dementia Recognition through Spontaneous Speech: The ADReSS Challenge*, 2020. arXiv: 2004.06833 [eess.AS].

[18] L. Hernandez-Dominguez, S. Ratte, G. Sierra-Martinez, and A. Roche-Bergua, "Computer-based evaluation of Alzheimer's disease and mild cognitive impairment patients during a picture description task," *Alzheimers Dement (Amst)*, vol. 10, pp. 260–268, Mar. 2018.

[19] A. Roshanzamir, H. Aghajan, and M. Soleymani, "Transformer-Based Deep Neural Network Language Models for Alzheimer's Disease Detection from Targeted Speech," *BMC Medical Informatics and Decision Making*, Jul. 2020. DOI: 10.21203/rs.3.rs-49267/v1.

[20] P. Saltz, S. Lin, S. Cheng, and D. Si, "Dementia detection using transformer-based deep learning and natural language processing models," in *2021 IEEE 9th International Conference on Healthcare Informatics (ICHI)*, Los Alamitos, CA, USA: IEEE Computer Society, Aug. 2021, pp. 509–510. DOI: 10.1109/ICHI52183.2021.00094. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/ICHI52183.2021.00094.

[21] L. Cummings, "Describing the Cookie Theft picture: Sources of breakdown in Alzheimer's dementia," *Pragmatics and Society*, vol. 10, pp. 151–174, Mar. 2019. DOI: 10.1075/ps.17011.cum.

[22] H. Goodglass and E. Kaplan, *The assessment of aphasia and related disorders*. Lea & Febiger, 1972.

[23] B. MacWhinney, *The CHILDES project: Tools for analyzing talk: Volume I: Transcription format and programs, Volume II: The database*, 3rd ed. Lawrence Erlbaum Associates Publishers, 2000.

[24] X. Han, Z. Zhang, N. Ding, *et al.*, "Pre-trained models: Past, present and future," *CoRR*, vol. abs/2106.07139, 2021. arXiv: 2106.07139. [Online]. Available: https://arxiv.org/abs/2106.07139.

[25] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota: ACL, Jun. 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. [Online]. Available: https://aclanthology.org/N19-1423.

[26] Y. Liu, M. Ott, N. Goyal, *et al.*, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *CoRR*, vol. abs/1907.11692, 2019. arXiv: 1907.11692. [Online]. Available: http://arxiv.org/abs/1907.11692.

[27] T. Wolf, L. Debut, V. Sanh, *et al.*, "Huggingface's transformers: State-of-the-art natural language processing," *CoRR*, vol. abs/1910.03771, 2019. arXiv: 1910.03771. [Online]. Available: http://arxiv.org/abs/1910.03771.

[28] A. Defazio and S. Jelassi, "Adaptivity without compromise: A momentumized, adaptive, dual averaged gradient method for stochastic optimization," *CoRR*, vol. abs/2101.11075, 2021. arXiv: 2101.11075. [Online]. Available: https://arxiv.org/abs/2101.11075.

[29] Z. Yang, Z. Dai, Y. Yang, J. G. Carbonell, R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized Autoregressive Pretraining for Language Understanding," *CoRR*, vol. abs/1906.08237, 2019. arXiv: 1906.08237. [Online]. Available: http://arxiv.org/abs/1906.08237.

[30] I. Beltagy, M. E. Peters, and A. Cohan, *Longformer: The long-document transformer*, 2020. arXiv: 2004.05150 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2004.05150.