INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

To BERT or Not To BERT: Comparing Speech and Language-based Approaches for Alzheimer's Disease Detection

Aparna Balagopalan[1], Benjamin Eyre[1], Frank Rudzicz[2,3,4,5], Jekaterina Novikova[1]

[1]Winterlight Labs Inc, Toronto, ON
[2]University of Toronto, ON
[3] Vector Institute for Artificial Intelligence, Toronto, ON
[4]Li Ka Shing Knowledge Institute, St Michael's Hospital, Toronto, ON
[5]Surgical Safety Technologies, Toronto, ON

aparna@winterlightlabs.com, benjamin@winterlightlabs.com, frank@cs.toronto.edu, jekaterina@winterlightlabs.com

Abstract

Research related to automatically detecting Alzheimer's disease (AD) is important, given the high prevalence of AD and the high cost of traditional methods. Since AD significantly affects the content and acoustics of spontaneous speech, natural language processing and machine learning provide promising techniques for reliably detecting AD. We compare and contrast the performance of two such approaches for AD detection on the recent ADReSS challenge dataset [1]: 1) using domain knowledge-based hand-crafted features that capture linguistic and acoustic phenomena, and 2) fine-tuning Bidirectional Encoder Representations from Transformer (BERT)-based sequence classification models. We also compare multiple feature-based regression models for a neuropsychological score task in the challenge. We observe that fine-tuned BERT models, given the relative importance of linguistics in cognitive impairment detection, outperform feature-based approaches on the AD detection task.

Index Terms: Alzheimer's disease, ADReSS, dementia detection, MMSE regression, BERT, feature engineering, transfer learning.

1. Introduction

Alzheimer's disease (AD) is a progressive neurodegenerative disease that causes problems with memory, thinking, and behaviour. AD affects over 40 million people worldwide with high costs of acute and long-term care [2]. Current forms of diagnosis are both time consuming and expensive [3], which might explain why almost half of those living with AD do not receive a timely diagnosis [4].

Studies have shown that valuable clinical information indicative of cognition can be obtained from spontaneous speech elicited using pictures [5]. Several studies have used speech analysis, natural language processing (NLP), and ML to distinguish between healthy and cognitively impaired speech of participants in picture description datasets [6, 7]. These serve as quick, objective, and non-invasive assessments of an individual's cognitive status. However, although ML methods for automatic AD-detection using such speech datasets achieve high classification performance (between 82%-93% accuracy) [6, 8, 9], the field still lacks publicly-available, balanced, and standardised benchmark datasets. The ongoing ADReSS challenge [1] provides an age/sex-matched balanced speech dataset, which consists of speech from AD and non-AD participants describing a picture. The challenge consists of two key tasks: 1) Speech classification task: classifying speech as AD or non-AD. 2) Neuropsychological score regression task: predicting Mini-Mental State Examination (MMSE) [10] scores from speech.

In this work, we develop ML models to detect AD from speech using picture description data of the demographically-matched ADReSS challenge speech dataset [1], and compare the following training regimes and input representations to detect AD:

1. **Using domain knowledge**: with this approach, we extract linguistic features from transcripts of speech, and acoustic features from corresponding audio files for binary AD vs non-AD classification and MMSE score regression. The features extracted are informed by previous clinical and ML research in the space of cognitive impairment detection [6].
2. **Using transfer learning**: with this approach, we fine-tune pre-trained BERT [11] text classification models at transcript-level. BERT achieved state-of-the-art results on a wide variety of NLP tasks when fine-tuned [11]. Our motivation is to benchmark a similar training procedure on transcripts from a pathological speech dataset, and evaluate the effectiveness of high-level language representations from BERT in detecting AD.

In this paper, we evaluate performance of these two methods on both the ADReSS train dataset, and on the unseen test set. We find that fine-tuned BERT-based text sequence classification models achieve the highest AD detection accuracy with an accuracy of 83.3% on the test set. With the feature-based models, the highest accuracy of 81.3% is achieved by the SVM with RBF kernel model. The lowest root mean squared error obtained for the MMSE prediction task is 4.56, with a feature-based L2 regularized linear regression model.

The main contributions of our paper are as follows:

- We employ a domain knowledge-based approach and compare a number of AD detection and MMSE regression models with an extensive list of pre-defined linguistic and acoustic features as input representations from speech (Section 5 and 6).
- We employ a transfer learning-based approach and benchmark fine-tuned BERT models for the AD vs non-AD classification task (Section 5 and 6).
- We contrast the performance of the two approaches on the classification task, and discuss the reasons for existing differences (Section 7).

http://dx.doi.org/10.21437/Interspeech.2020-2557

Table 1: *Basic characteristics of the patients in each group in the ADReSS challenge dataset are more balanced in comparison to DementiaBank.*

Dataset			Class AD	Class Non-AD
ADReSS	Train	Male	24	24
		Female	30	30
ADReSS	Test	Male	11	11
		Female	13	13
DementiaBank [17]	-	Male	125	83
		Female	197	146

2. Background

2.1. Domain Knowledge-based Approach

Previous work has focused on automatic AD detection from speech using acoustic features (such as zero-crossing rate, Mel-frequency cepstral coefficients) and linguistic features (such as proportions of various part-of-speech (POS) tags [12, 6, 8]) from speech transcripts. Fraser *et al.* [6] extracted 370 linguistic and acoustic features from picture descriptions in the DementiaBank dataset, and obtained an AD detection accuracy of 82% at transcript-level. More recent studies showed the addition of normative data helped increase accuracy up to 93% [8, 13] .

Yancheva *et al.* [14] showed ML models are capable of predicting the MMSE scores from features of speech elicited via picture descriptions, with mean absolute error of 2.91-3.83.

Detecting AD or predicting MMSE scores with engineered features of speech and thereby infusing domain knowledge into the task has several advantages, such as more interpretable model decisions and potentially lower resource requirement when paired with conventional ML models. However, there are also disadvantages, e.g. a time consuming feature engineering process, and a risk of missing highly relevant features.

2.2. Transfer Learning-based Approach

In the recent years, transfer learning in the form of pre-trained language models has become ubiquitous in NLP [15] and has contributed to the state-of-the-art on a wide range of tasks. One of the most popular transfer learning models is BERT [11], which builds on Transformer networks [16] to pre-train bidirectional representations of text by conditioning on both left and right contexts jointly in all layers.

BERT uses powerful attention mechanisms to encode global dependencies between the input and output. This allows it to achieve state-of-the-art results on a suite of benchmarks [11]. Fine-tuning BERT for a few epochs can potentially attain good performance even on small datasets. However, such models are not directly interpretable, unlike feature-based ones.

3. Dataset

We use the ADReSS Challenge dataset [1], which consists of 156 speech samples and associated transcripts from non-AD (N=78) and AD (N=78) English-speaking participants. Speech is elicited from participants through the Cookie Theft picture from the Boston Diagnostic Aphasia exam [5]. In contrast to other speech datasets for AD detection such as DementiaBank's English Pitt Corpus [17], the ADReSS challenge dataset is matched for age and gender (Table 1). The speech dataset is divided into standard train and test sets. MMSE [10] scores are available for all but one of the participants in the train set.

4. Feature Extraction

The speech transcripts in the dataset are manually transcribed as per the CHAT protocol [18], and include speech segments from both the participant and an investigator. We only use the portion of the transcripts corresponding to the participant. Additionally, we combine all participant speech segments corresponding to a single picture description for extracting acoustic features.

We extract 509 manually-engineered features from transcripts and associated audio files (see Appendix A for a list of all features). These features are identified as indicators of cognitive impairment in previous literature, and hence encode domain knowledge. All of them are divided into 3 categories:

1. **Lexico-syntactic features (297):** Frequencies of various production rules from the constituency parsing tree of the transcripts [19], speech-graph based features [20], lexical norm-based features (e.g. average sentiment valence of all words in a transcript, average imageability of all words in a transcript [21]), features indicative of lexical richness. We also extract syntactic features [22] such as the proportion of various POS-tags, and similarity between consecutive utterances.
2. **Acoustic features (187):** Mel-frequency cepstral coefficients (MFCCs), fundamental frequency, statistics related to zero-crossing rate, as well as proportion of various pauses [23] (for example, filled and unfilled pauses, ratio of a number of pauses to a number of words, etc.)
3. **Semantic features based on picture description content (25):** Proportions of various information content units used in the picture, identified as being relevant to memory impairment in prior literature [24].

5. Experiments

5.1. AD vs non-AD Classification

5.1.1. Training Regimes

We benchmark the following training regimes for classification: classifying features extracted at transcript-level and a BERT model fine-tuned on transcripts.

Domain knowledge-based approach: We classify lexicosyntactic, semantic, and acoustic features extracted at transcript-level with four conventional ML models (SVM, neural network (NN), random forest (RF), naïve Bayes (NB)[1].

Hyperparameter tuning: We optimize each model to the best possible hyper-parameter setting using grid-search 10-fold cross-validation (CV). We perform feature selection by choosing top-k number of features, based on ANOVA F-value between label/features. The number of features is jointly optimized with the classification model parameters (see Appendix C for a full list of parameters).

Transfer learning-based approach: To leverage the language information encoded by BERT [11], we add a linear layer mapping representations from the final layer of a pre-trained 12-layer BERT base, uncased model to binary class labels [25] for the AD vs non-AD classification task. The transcript-level input to the model consists of transcribed utterances with corresponding start and separator special tokens for each utterance, following Liu *et al.* [26]. A pooled embedding summarizing information across all tokens in the transcript using the self-attention mechanism in the BERT base is used as the aggregate transcript

[1] https://scikit-learn.org/stable/

representation, and passed to the classification layer [11, 25]. This model is then fine-tuned on training data for AD detection.

Hyperparameter tuning: We optimize the number of epochs to 10 by varying it from 1 to 12 during CV. Adam optimizer [27] and warmup linear learning rate scheduling [28] are used (details in Appendix B).

5.1.2. Evaluation

Cross-validation on ADReSS train set: We use two CV strategies in our work – leave-one-subject-out CV (LOSO CV) and 10-fold CV at transcript level. We report evaluation metrics with LOSO CV for all models except fine-tuned BERT for direct comparison to challenge baselines. Due to computational constraints of GPU memory, we are unable to perform LOSO CV for the BERT model. Hence, we perform 10-fold CV to compare feature-based classification models with fine-tuned BERT. Values of performance metrics for each model are averaged across three runs with different random seeds in all cases.

Predictions on ADReSS test set: We generate three predictions with different seeds from each hyperparameter-optimized classifier trained on the complete train set, and then produce a majority prediction to avoid overfitting. We report performance on the challenge test set, as obtained from the challenge organizers (see Appendix E for more details).

We evaluate performance primarily using accuracy scores, since all train/test sets are known to be balanced. We also report precision, recall, specificity and F1 with respect to the positive class (AD), and compare to the highest challenge baseline (LDA classifier using language outcome measures [1]).

5.2. MMSE Score Regression

5.2.1. Training Regimes

Domain knowledge-based approach: For this task, we benchmark two kinds of regression models, linear and ridge, using pre-engineered features as input. MMSE scores range from 0 to 30, and so predictions are clipped to range between 0 and 30.

Hyperparameter tuning: Each model's performance is optimized using hyperparameters selected via grid-search LOSO CV. We perform feature selection by choosing top-k features, based on F-Scores computed from the correlation of each feature with MMSE score. The number of features is optimized for all models. For ridge regression, the number of features is jointly optimized with the coefficient for L2 regularization, α.

5.2.2. Evaluation

We report root mean squared error (RMSE) and mean absolute error (MAE) for the predictions produced by each of the models on the training set with LOSO CV. In addition, we include the RMSE for two models' predictions on the ADReSS test set. Hyperparameters for these models were selected using grid-search 10-fold cross validation on the training set. We compare regression performance to the best challenge baseline (decision tree regressor using language outcome measures [1]).

6. Results

6.1. AD vs non-AD Classification

In Table 3, the classification performance with all the models evaluated on the train set via 10-fold CV is displayed. We observe that BERT outperforms all domain knowledge-based ML models with respect to all metrics. SVM is the best-performing domain knowledge-based model. However, accuracy of the fine-tuned BERT model is not significantly higher than that of the SVM classifier based on an Kruskal-Wallis H-test ($H = 0.4838, p > 0.05$).

We also report the performance of all our feature classification models with LOSO CV (Table 4), and compare to the highest challenge baseline [1]. Each of our classification models outperforms the challenge baseline, with a +10% accuracy increase with the SVM classifer. Feature selection results in accuracy increase of about 13% for the SVM classifier.

Results on the unseen, held-out ADReSS test set (Table 5) follow the trend of the cross-validated performance in terms of accuracy, with BERT outperforming the best feature-based classification model, SVM, as well as the challenge baseline.

6.2. MMSE Score Regression

Performance of regression models evaluated on both train and test sets is shown in Table 6. Ridge regression with 25 features selected attains the lowest RMSE on the training set amongst our models, with 4.56 RMSE during LOSO-CV, which is 0.18 higher than the challenge baseline. The results show that feature selection can help achieve a decrease of up to 1.5 RMSE points (and up to 0.86 MAE) for a ridge regressor. Furthermore, a ridge regressor is able to achieve an RMSE of 4.56 on the ADReSS test set, a decrease of 0.64 from the baseline.

7. Discussion

7.1. Feature Differentiation Analysis

We extract a large number of features to capture a wide range of linguistic and acoustic phenomena, based on a survey of prior literature in automatic cognitive impairment detection [6, 14, 30, 31]. In order to identify the most differentiating features between AD and non-AD speech, we perform independent t-tests between feature means for each class in the ADReSS training set. 87 features are significantly different between the two groups at $p < 0.05$. 79 of these are text-based lexicosyntactic and semantic features, while 8 are acoustic. These 8 acoustic features include the number of long pauses, pause duration, and mean/skewness/variance-statistics of various MFCC coefficients. However, after Bonferroni correction for multiple testing, we identify that only 13 features are significantly different between AD and non-AD speech at $p < 9e-5$, and none of these features are acoustic (Table 2). This implies that linguistic features are particularly differentiating between the AD/non-AD classes here, which explains why models trained on linguistic features only attain performance well above random chance (see Fig. 1 in Appendix for visualization of class separability).

7.2. Analysing AD Detection Performance Differences

Comparing classification performance across model settings, we observe that BERT outperforms the best domain knowledge-based model in terms of accuracy and F1-score on the train set (10-fold CV; though accuracy is not significantly higher) and on the test set (no significance testing possible since only single set of performance scores are available per model; see Appendix E for procedure for submitting challenge predictions). Based on feature differentiation analysis (Section 7.1), we hypothesize that good performance with a text-focused BERT model on this speech classification task is due to the strong utility of linguistic features on this dataset. BERT captures a range of linguistic phenomena due to its training methodology, potentially encap-

Table 2: *Feature differentiation analysis results based on ADReSS train set. μ_{AD} and μ_{non-AD} show the means of the 13 significantly different features at p<9e-5 (after Bonferroni correction) for the AD and non-AD group respectively. We also show Spearman correlation between MMSE score and features, and regression weights of the features associated with the five greatest and five lowest regression weights from a ridge regressor (25 features, $\alpha = 12$).* next to correlation indicates significance at p<9e-5.*

Feature	**Feature type**	μ_{AD}	μ_{non-AD}	**Correlation**	**Weight**
Average cosine distance between utterances	Semantic	0.91	0.94	-	-
Fraction of pairs of utterances below a similarity threshold (0.5)	Semantic	0.03	0.01	-	-
Average cosine distance between 300-dimensional word2vec [29] utterances and picture content units	Semantic (content units)	0.46	0.38	-0.54*	-1.01
Distinct content units mentioned: total content units	Semantic (content units)	0.27	0.45	0.63*	1.78
Distinct action content units mentioned: total content units	Semantic (content units)	0.15	0.30	0.49*	1.04
Distinct object content units mentioned: total content units	Semantic (content units)	0.28	0.47	0.59*	1.72
Average cosine distance between 50-dimensional GloVe utterances and picture content units	Semantic content units)	-	-	-0.42*	-0.03
Average word length (in letters)	Lexico-syntactic	3.57	3.78	0.45*	1.07
Proportion of pronouns	Lexico-syntactic	0.09	0.06	-	-
Ratio (pronouns):(pronouns+nouns)	Lexico-syntactic	0.35	0.23	-	-
Proportion of personal pronouns	Lexico-syntactic	0.09	0.06	-	-
Proportion of RB adverbs	Lexico-syntactic	0.06	0.04	-0.41*	-0.41
Proportion of ADVP_$->$_RB amongst all rules	Lexico-syntactic	0.02	0.01	-0.37	-0.74
Proportion of non-dictionary words	Lexico-syntactic	0.11	0.08	-	-
Proportion of gerund verbs	Lexico-syntactic	-	-	0.37	1.08
Proportion of words in adverb category	Lexico-syntactic	-	-	-0.4*	-0.49

Table 3: *10-fold CV results averaged across 3 runs with different random seeds on the ADReSS train set. Accuracy for BERT is higher, but not significantly so from SVM ($H = 0.4838, p > 0.05$ Kruskal-Wallis H test). Bold indicates the best result.*

Model	#Features	Accuracy	Precision	Recall	Specificity	F1
SVM	10	0.796	0.81	0.78	0.82	0.79
NN	10	0.762	0.77	0.75	0.77	0.76
RF	50	0.738	0.73	0.76	0.72	0.74
NB	80	0.750	0.76	0.74	0.76	0.75
BERT	-	**0.818**	**0.84**	**0.79**	**0.85**	**0.81**

Table 4: *LOSO-CV results averaged across 3 runs with different random seeds on the ADReSS train set. Accuracy for SVM is significantly higher than NN ($H = 4.50, p = 0.034$ Kruskal-Wallis H test). Bold indicates the best result.*

Model	#Features	Accuracy	Precision	Recall	Specificity	F1
Baseline [1]	-	0.768	0.77	0.76	-	0.77
SVM	509	0.741	0.75	0.72	0.76	0.74
SVM	10	**0.870**	**0.90**	**0.83**	**0.91**	**0.87**
NN	10	0.836	0.86	0.81	0.86	0.83
RF	50	0.778	0.79	0.77	0.79	0.78
NB	80	0.787	0.80	0.76	0.82	0.78

sulating many important lexico-syntactic and semantic features. It is thus able to use information present in the lexicon, syntax, and semantics of transcribed speech after fine-tuning [32].

We see a trend of better performance while increasing the number of folds (see SVM in Table 4 and Table 3) in cross-validation. We postulate that this is due to the small size of the dataset, and hence differences in training set size in each fold.

7.3. Regression Weights

To assess the relative importance of individual input features for MMSE prediction, we report features with the 5 highest and 5 lowest regression weights in Table 2. Each value is the average weight assigned to features selected in each LOSO CV fold using ridge regression. We also present the correlation with MMSE score for these features, as well as their significance.We observe that for each of these highly weighted features, a positive or negative correlation is accompanied by a positive or negative regression weight, respectively. This demonstrates that even in the presence of other regressors, the relationship with MMSE score remains the same for these features. We also note that all 10 of these features are linguistic, further demonstrating that linguistic information is particularly distinguishing when it comes to predicting the severity of a patient's AD.

Table 5: *AD detection results on unseen, held-out ADReSS test set presented in same format as the baseline paper [1]. Bold indicates the best result.*

Model	#Features	Class	Accuracy	Precision	Recall	Specificity	F1
Baseline [1]	-	non-AD	0.750	0.70	**0.87**	-	0.78
		AD		**0.83**	0.62	-	0.71
SVM	10	non-AD	0.813	0.83	0.79	0.83	0.81
		AD		0.80	0.83		0.82
NN	10	non-AD	0.771	0.78	0.75	0.78	0.77
		AD		0.76	0.79		0.78
RF	50	non-AD	0.750	0.71	0.83	0.71	0.77
		AD		0.80	0.67		0.73
NB	80	non-AD	0.729	0.69	0.83	0.69	0.75
		AD		0.79	0.63		0.70
BERT	-	non-AD	**0.833**	**0.86**	0.79	**0.86**	**0.83**
		AD		0.81	**0.88**		**0.84**

Table 6: *LOSO-CV MMSE regression results on the ADReSS train and test sets. Bold indicates the best result.*

Model	#Features	α	RMSE (Train set)	MAE (Train set)	RMSE (Test set)
Baseline [1]	-	-	**4.38**		5.20
LR	15	-	5.37	4.18	4.94
LR	20	-	4.94	3.72	-
Ridge	509	12	6.06	4.36	-
Ridge	35	12	4.87	3.79	**4.56**
Ridge	25	10	4.56	**3.50**	-

8. Conclusions

In this paper, we compare two widely used approaches – explicit features engineering based on domain knowledge, and transfer learning using a fine-tuned BERT [11] classification model. Our results show that pre-trained models that are fine-tuned for the AD classification task are capable of performing well, outperforming hand-crafted feature engineering. In the future, we will experiment with different language representation models, and with different tokenization and encoding strategies for transcript representations. A direction for future work is also developing models that combine representations from language representation models like BERT and hand-crafted features [33]. Such feature-fusion approaches could potentially boost performance on the cognitive impairment detection task.

9. Acknowledgements

This paper benefited greatly from feedback and review from multiple people. Most notably, Dr. Jessica Robin (Winterlight Labs), Jordan Ponn (Winterlight Labs), Liam Kaufman (Winterlight Labs) and Maria Yancheva (Winterlight Labs).
Dr. Frank Rudzicz is supported by a CIFAR Chair in AI.

10. References

[1] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Alzheimer's dementia recognition through spontaneous speech: The adress challenge," in *Proceedings of INTERSPEECH 2020*, Shanghai, China, 2020.

[2] M. Prince, A. Comas-Herrera, M. Knapp, M. Guerchet, and M. Karagiannidou, "World alzheimer report 2016: improving healthcare for people living with dementia: coverage, quality and costs now and in the future," 2016.

[3] G. Prabhakaran, R. Bakshi *et al.*, "Analysis of structure and cost in a longitudinal study of alzheimer's disease," *Journal of Health Care Finance*, 2018.

[4] E. A. Jammeh, B. C. Camille, W. P. Stephen, J. Escudero, A. Anastasiou, P. Zhao, T. Chenore, J. Zajicek, and E. Ifeachor, "Machine-learning based identification of undiagnosed dementia in primary care: a feasibility study," *BJGP Open*, p. bjgpopen18X101589, 2018.

[5] H. Goodglass, E. Kaplan, and B. Barresi, *BDAE-3: Boston Diagnostic Aphasia Examination–Third Edition*. Lippincott Williams & Wilkins Philadelphia, PA, 2001.

[6] K. C. Fraser, J. A. Meltzer, and F. Rudzicz, "Linguistic features identify Alzheimer's disease in narrative speech," *Journal of Alzheimer's Disease*, vol. 49, no. 2, pp. 407–422, 2016.

[7] Z. Zhu, J. Novikova, and F. Rudzicz, "Semi-supervised classification by reaching consensus among modalities," *arXiv preprint arXiv:1805.09366*, 2018.

[8] Z. Noorian, C. Pou-Prom, and F. Rudzicz, "On the importance of normative data in speech-based assessment," *arXiv preprint arXiv:1712.00069*, 2017.

[9] 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: Volume 2 (Short Papers)*, 2018, pp. 701–707.

[10] J. R. Cockrell and M. F. Folstein, "Mini-mental state examination," *Principles and practice of geriatric psychiatry*, pp. 140–141, 2002.

[11] 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: Volume 1 (Long and Short Papers)*, 2019, pp. 4171–4186.

[12] S. O. Orimaye, K. Y. Tai, J. S.-M. Wong, and C. P. Wong, "Learning linguistic biomarkers for predicting mild cognitive impairment using compound skip-grams," *arXiv preprint arXiv:1511.02436*, 2015.

[13] A. Balagopalan, J. Novikova, F. Rudzicz, and M. Ghassemi, "The effect of heterogeneous data for alzheimer's disease detection from speech," *arXiv preprint arXiv:1811.12254*, 2018.

[14] M. Yancheva, K. C. Fraser, and F. Rudzicz, "Using linguistic features longitudinally to predict clinical scores for alzheimer's disease and related dementias," in *Proceedings of SLPAT 2015: 6th Workshop on Speech and Language Processing for Assistive Technologies*, 2015, pp. 134–139.

[15] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing," *IEEE Computational Intelligence Magazine*, vol. 13, no. 3, pp. 55–75, 2018.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez,Ł . Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[17] J. T. Becker, F. Boiler, O. L. Lopez, J. Saxton, and K. L. McGonigle, "The natural history of alzheimer's disease: description of study cohort and accuracy of diagnosis," *Archives of Neurology*, vol. 51, no. 6, pp. 585–594, 1994.

[18] B. MacWhinney, *The CHILDES project: Tools for analyzing talk, Volume I: Transcription format and programs*. Psychology Press, 2014.

[19] J. Chae and A. Nenkova, "Predicting the fluency of text with shallow structural features: Case studies of machine tanslation and human-written text," 2009.

[20] N. B. Mota, N. A. Vasconcelos, N. Lemos, A. C. Pieretti, O. Kinouchi, G. A. Cecchi, M. Copelli, and S. Ribeiro, "Speech graphs provide a quantitative measure of thought disorder in psychosis," *PloS one*, vol. 7, no. 4, 2012.

[21] A. B. Warriner, V. Kuperman, and M. Brysbaert, "Norms of valence, arousal, and dominance for 13,915 english lemmas," *Behavior research methods*, vol. 45, no. 4, pp. 1191–1207, 2013.

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

[23] B. H. Davis and M. Maclagan, "Examining pauses in alzheimer's discourse," *American Journal of Alzheimer's Disease & Other Dementias®*, vol. 24, no. 2, pp. 141–154, 2009.

[24] B. Croisile, B. Ska, M.-J. Brabant, A. Duchene, Y. Lepage, G. Aimard, and M. Trillet, "Comparative study of oral and written picture description in patients with alzheimer's disease," *Brain and language*, vol. 53, no. 1, pp. 1–19, 1996.

[25] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz *et al.*, "Huggingface's transformers: State-of-the-art natural language processing," *ArXiv, abs/1910.03771*, 2019.

[26] Y. Liu and M. Lapata, "Text summarization with pretrained encoders," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 3721–3731.

[27] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[28] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, 2019, pp. 8024–8035.

[29] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[30] C. Pou-Prom and F. Rudzicz, "Learning multiview embeddings for assessing dementia," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 2812–2817.

[31] Z. Zhu, J. Novikova, and F. Rudzicz, "Detecting cognitive impairments by agreeing on interpretations of linguistic features," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019, pp. 1431–1441.

[32] G. Jawahar, B. Sagot, and D. Seddah, "What does bert learn about the structure of language?" in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 3651–3657.

[33] M. Yu, M. R. Gormley, and M. Dredze, "Combining word embeddings and feature embeddings for fine-grained relation extraction," in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics*, 2015, pp. 1374–1379.

[34] X. Lu, "Automatic analysis of syntactic complexity in second language writing," *International journal of corpus linguistics*, vol. 15, no. 4, pp. 474–496, 2010.

[35] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.