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

Comparing Natural Language Processing Techniques for Alzheimer's Dementia Prediction in Spontaneous Speech

Thomas Searle[1], *Zina Ibrahim*[1], *Richard Dobson*[1,2]

[1]Department of Biostatistics and Health Informatics,
Institute of Psychiatry, Psychology and Neuroscience,
King's College London, London, U.K.
[2]Institute of Health Informatics, University College London,
London, London, U.K.
{firstname}.{lastname}@kcl.ac.uk

Abstract

Alzheimer's Dementia (AD) is an incurable, debilitating, and progressive neurodegenerative condition that affects cognitive function. Early diagnosis is important as therapeutics can delay progression and give those diagnosed vital time. Developing models that analyse spontaneous speech could eventually provide an efficient diagnostic modality for earlier diagnosis of AD. The Alzheimer's Dementia Recognition through Spontaneous Speech task offers acoustically pre-processed and balanced datasets for the classification and prediction of AD and associated phenotypes through the modelling of spontaneous speech. We exclusively analyse the supplied textual transcripts of the spontaneous speech dataset, building and comparing performance across numerous models for the classification of AD vs controls and the prediction of Mental Mini State Exam scores. We rigorously train and evaluate Support Vector Machines (SVMs), Gradient Boosting Decision Trees (GBDT), and Conditional Random Fields (CRFs) alongside deep learning Transformer based models. We find our top performing models to be a simple Term Frequency-Inverse Document Frequency (TF-IDF) vectoriser as input into a SVM model and a pre-trained Transformer based model 'DistilBERT' when used as an embedding layer into simple linear models. We demonstrate test set scores of 0.81-0.82 across classification metrics and a RMSE of 4.58.

Index Terms: adress shared task, spontaneous speech classification, alzheimers dementia classification

1. Introduction

Alzheimer's Dementia (AD) is a progressive neurodegenerative condition that largely affects cognitive function. With our globally aging population, conditions such as AD are likely to become more prevalent[1]. Despite there being no cure currently, early diagnosis can offer interventions to slow or delay progression of symptoms[2]. Prior work has used machine learning methods for the prediction of cognitive impairment (CI) conditions, including AD, using patient structured data[3] and medical imaging data[4]. Linguistic phenomenon have also been identified in those already diagnosed with AD[5, 6].

The Alzheimer's Dementia Recognition through Spontaneous Speech (ADReSS) challenge presents two tasks in the modelling of spontaneous speech[7]. Firstly, to classify presence of AD vs controls and secondly, to predict the 'Mental Mini State Exam' score, a common set of questions designed to assess cognitive function[8]. The challenge provides 108 training, 54 AD vs 54 Control samples, and 48 unseen test samples. Spontaneous speech audio and associated transcripts of participants describing the 'Cookie Theft' picture from the Boston Diagnostic Aphasia Exam[9] are provided. Samples are demographically and acoustically balanced and longer in duration than previous clinical studies[7]. The challenge provides an environment for researchers to test competing methods with recommendations for future work. Using machine learning techniques to predict AD from spontaneous speech could potentially offer an efficient early diagnostic modality. For example, audio samples could be collected via a mobile device with results directing individuals to seek more formal medical evaluation.

2. Data Prepossessing

In this work we exclusively focus on the textual transcriptions that are provided alongside the audio samples. Transcripts are supplied using the CHAT transcription format[10]. The transcription schema provides the linguistic content alongside some prosodic content such as: pauses, laughter, discourse markers such as 'um' and 'ah', and abbreviations such as '(be)cause'. We preprocess each transcript before feeding into our model pipelines. All code to re-create the data prepossessing, experiments and analysis is available open-source[1].

The preprocessing parses participant metadata such as age, sex, AD diagnosis and MMSE score. Each transcription line is parsed to remove time duration suffixes, specific speech artifacts such as '[','\]' or '>', '<' and excess white-space such as tabs and newlines. We purposely leave discourse markers such as 'um' 'ah' and other speech artifacts such as '+...', '&=laughs' and '(...)' that indicate various pause types, or laughter in the audio.

2.1. Data Splits and Granularity

We split the transcripts into multiple competing datasets providing the candidate models with greatest opportunity to find adequate signal for the prediction tasks.

2.1.1. Transcript Level Data

Within these datasets each transcript is a single data point with their corresponding AD label and assigned MMSE score. This includes:

1. A dataset with only participant utterances concatenated together into a single paragraph as they appear in the transcript. Denoted **PAR**.

[1] https://github.com/tomolopolis/ADReSS_Challenge

http://dx.doi.org/10.21437/Interspeech.2020-2729

2. A dataset with both participant and interviewer speech concatenated into a single paragraph as they appear in the transcript. Denoted **PAR+INV**.

2.1.2. Utterance Level Data

For these datasets we define each utterance as an individual data point. This provides N=1,476, AD(N=740), controls (N=736). The target labels and MMSE scores are replicated to each utterance. Segments maintain a reference to their source transcript so random shuffling does not produce data leakage between the train and test phases. We only consider participant spoken utterances here as initial experiments indicated including interviewer speech lead to a reduction in performance. This includes:

1. A dataset with only participant utterance as individual classification & regression data points. Denoted **PAR_SPLT**.
2. Further datasets that extend the text based features with the inclusion of temporal and participant demographic features such as: time duration per sentence, time between sentences, average/max/min sentence time denoted **PAR_SPLT+T**, and participant age and sex denoted **PAR_SPLT+T+D**.

3. Methods

The baseline paper accompanying the challenge[7] only creates a single baseline result using only the text transcripts. Therefore, we present a range of models both as a new baseline result for the linguistic features, Section 3.1, alongside our more advanced approaches in Section 3.2.

3.1. Baseline Methods

We make extensive use of Scikit-Learn[11], a python based machine learning framework that provides APIs for common machine learning models, feature extraction, cross validation, hyper parameter optimisation and performance metric calculation. We use the integrated Term-Frequency-Inverse Document Frequency (TF-IDF)[12] 'bag-of-words' vectoriser. With this method text inputs forgo their sequence order and words are counted within and across documents. TF-IDF down-weights the counts of common cross-document terms, and increases weights of rare cross-document but frequent intra-document terms. This embedding method is a common first stage in any textual modelling exercise due to its efficiency and ease of use.

Scikit-learn provides APIs for optimised implementations of common machine learning algorithms such as libsvm[13] for Support Vector Machines(SVM)[14] and XGBoost[15] for Gradient Boosted Decision Trees(GBDT)[16] allowing for fast model fitting. We use both algorithms in the development of our baseline models for the transcript level and utterance level datasets presented in Section 2.1

SVMs and GBDTs are effective techniques to learn non-linear relationships between input features and the decision boundaries for both classification and regression tasks.

3.1.1. Utterance Level Methods

For the segmented speech datasets, PAR_SPLT, PAR_SPLT+T, PAR_SPLT+T+D presented in Section 2.1.2, our modelling approach does not support MMSE prediction so we only report results for AD classification. We train and cross validate TF-IDF/SVM and TF-IDF/GBDT models on each utterance, and feed output prediction probability sequences to a Conditional Random Field (CRF)[17]. CRFs are effective in the modelling of sequential data as input feature representations can depend on previous and future states of the sequence. For the overall classification of the transcript we take the final classification state of the CRF.

3.1.2. Hyper Parameter Optimisation

Table 1 lists the model configuration and associated hyper-parameter spaces we search across during an exhaustive 5-fold cross validation grid-search. As our dataset is fairly small, performing this only took a couple of minutes for each model configuration and each dataset despite the many individual model fits.

Table 1: *Baseline methods hyper-parameter searched and found optimal parameters.* * *values are* $\times 10^3$. † *the parameter spaces are sampled from an exponential probability distributions 15 times with specified* λ

Model	Hyper Parameter	Param Space	Optimal
TF-IDF/GBDT	Max Features	0.1, 0.5, 1, 2, 10$*$	1$*$
TF-IDF/GBDT	Stop Words	english, None	english
TF-IDF/GBDT	Analyser	word, char	word
TF-IDF/GBDT	sublinear TF	True, False	True
TF-IDF/GBDT	N-Estimators	100, 200, 500	100
TF-IDF/GBDT	Max Depth	3, 5, 10	5
TF-IDF/SVM	Max Features	0.1, 0.5, 1, 2, 10$*$	0.1$*$
TF-IDF/SVM	Stop Words	english, None	None
TF-IDF/SVM	Analyser	word, char	word
TF-IDF/SVM	sublinear TF	True, False	True
TF-IDF/SVM	Kernel	rbf, sigmoid	sigmoid
TF-IDF/SVM	C	0.1, 0.5, 1	1
SVM+CRF	c1	$\lambda = 0.5^\dagger$	0.0036
SVM+CRF	c2	$\lambda = 0.05^\dagger$	0.018
GBDT+CRF	c1	$\lambda = 0.5^\dagger$	0.314
GBDT+CRF	c2	$\lambda = 0.05^\dagger$	0.009

3.2. Deep Learning Methods

To converge successfully deep learning (DL) models often require more training data than methods such as SVMs and GBDTs. Training set sizes are often 50 or 100 times larger than available here. Transfer learning presents a compelling option to enable re-use of deep learning models for smaller domain specific data sets. Recently, transfer learning approaches have been successfully applied to a variety of NLP problems[18].

Large pre-trained language models are an example of transfer learning, and can be used to provide semantically rich embedding layers, allowing researchers to re-use knowledge acquired by the model from a prior training process. The language modelling task can be defined as predicting the next word given the sequence of previous words, or formally in Equation 1, modelling the probability distribution of all words w in a vocabulary V conditioned on previous words w_{i-1} to w_1.

$$P(w_i|w_{i-1}, w_{i-2} \cdots w_1) \forall w \in V \qquad (1)$$

The task enables the usage of large corpora of existing texts without any explicit manual annotation, often referred to as self-supervised learning[19]. Each model we use is based upon the Transformer architecture first presented for sequence to sequence problems such as machine translation[20]. The Transformer consists of layers of encoder and decoder blocks of multi-headed self-attention followed by fully connected layers.

Each successive layer learns sophisticated latent representations of the input texts.

We use the 'transformers'[21] library to load, and re-use the BERT[22], RoBERTa[23] and DistilBERT/DistilRoBERTa[24] models as embedding layers for the PAR and PAR+INV datasets.

Running the input transcripts through the pre-trained models produces a fixed size embedding representation for each provided transcript. This is an embedding matrix of size $N \times H$, where N is the number of transcripts and H is the hidden dimension of the pre-trained model. As recommended in prior work we fit simple linear models, Logistic Regression model for AD classification and LASSO Regression for MMSE prediction, to produce our final predictions.

4. Results

Table 2 provides results for average 10 fold cross-validation for hyper parameter selection and best train/development set performance. This attempts to compare model robustness with the available training data, especially for our transcript level datasets where dataset size is limited. Metrics follow the standard definitions as outlined in the baseline work[7] and are averaged between the classes Non-AD / AD for precision, recall and F1. We then pick our 5 best performing models / dataset configurations and run on the unlabelled test dataset, containing 48 samples, sending our AD and MMSE predictions to challenge organisers. Organisers subsequently responded with aggregate results as reported in Section 4.1 for AD classification and Section 4.2 for MMSE predictions.

Table 2: *Average 10-fold CV AD Classification and MMSE prediction results. Results are highlighted if within 0.02 of the highest score.* * *indicates best score for given metric.*

Dataset	Model	Acc	Prec	Recall	F1	RMSE
PAR	GBDT	.82	.84	.82	.81	5.93
PAR	SVM	.86	**.90**	.83	**.86**	6.57
PAR	DistilBERT	**.87**	**.90**	.87	**.87**	**4.49***
PAR	DistilRoBERTa	.84	.86	.85	.82	5.12
PAR	BERT(base)	.84	.86	.85	.82	5.12
PAR	RoBERTa(base)	.75	.79	.72	.74	7.11
PAR	BERT(large)	.77	.80	.77	.76	6.64
PAR	RoBERTa(large)	.77	.81	.73	.76	7.13
PAR+INV	GBDT	.79	.80	.82	.79	5.60
PAR+INV	SVM	**.88**	**.92***	.87	**.87**	6.74
PAR+INV	DistilBERT	**.87**	.89	**.89**	**.88***	4.85
PAR+INV	DistilRoBERTa	.80	.87	.79	.78	7.11
PAR+INV	BERT(base)	.75	.76	.78	.74	7.13
PAR+INV	RoBERTa(base)	.72	.71	.71	.69	5.45
PAR+INV	BERT(large)	.75	.78	.73	.74	7.13
PAR+INV	RoBERTa(large)	.81	.88	.76	.79	6.64
PAR_SPLT	SVM+CRF	.88	.88	**.88**	**.87**	-
PAR_SPLT	GBDT+CRF	.80	.84	.74	.78	-
PAR_SPLT+T	SVM+CRF	**.89***	.87	**.90***	**.88***	-
PAR_SPLT+T	GBDT+CRF	.82	.84	.79	.81	-
PAR_SPLT+T+D	SVM+CRF	.86	.85	.87	**.86**	-
PAR_SPLT+T+D	GBDT+CRF	.83	.86	.79	.81	-

4.1. AD Classification

Table 3 shows our test set results for each metric. We show results for metrics both labels (AD vs No AD) for precision, recall and F1 metrics as defined in baseline work[7].

Table 3: *Test set results for AD classification*

Dataset / Model	Class	Prec	Recall	F1	Acc
PAR / DistilBERT	Non-AD	0.76	0.79	0.78	0.77
	AD	0.783	0.75	0.77	
PAR+INV / DistilBERT	Non-AD	**0.83**	0.79	0.81	**0.81**
	AD	0.80	**0.83**	**0.82**	
PAR / TF-IDF/SVM	Non-AD	0.70	0.83	0.75	0.73
	AD	0.79	0.63	0.70	
PAR_SPLT / SVM+CRF	Non-AD	0.78	0.88	**0.82**	**0.81**
	AD	**0.86**	0.75	0.80	
PAR_SPLT+T / SVM+CRF	Non-AD	0.75	**0.88**	0.81	0.79
	AD	0.85	0.71	0.77	

4.2. MMSE Prediction

Table 4 provides our MMSE prediction results on the provided test set. We observe that the deep learning embedding methods perform best, and in particular the DistilBERT model using only participant sections of the transcript performs best RMSE. Interestingly, the deep learning methods perform well despite having not been trained with regression tasks in mind. Our CRF models do not support regression so we cannot report MMSE prediction scores for those model configurations.

Table 4: *Test set results for MMSE score prediction, 'DBL' indicates our DistilBERT embedding with LASSO linear model.*

Dataset/Model	PAR/DBL	PAR+INV/DBL	PAR/SVM
RMSE Score	5.37	4.58	5.22

4.3. Alternative Configurations

The supplied transcripts included temporal metadata for each sentence for participants and interviewers. We experimented with these time time based features alongside the text features at the transcript level, i.e. a PAR+TIME dataset. This included features: participant average / minimum / maximum and median sentence times, and time between sentences. Intuitively, we assumed that AD subjects would exhibit distinctly different time based features due to their impaired cognitive function. However, this dataset (PAR+TIME) performed poorly across the modelling approaches so we do not include in our results. We suggest this is due to the aggregation in the transcript level dataset removing any signal to that could be detected by the modelling approaches. PAR_SPLT+T does include temporal level features but does not perform as as well as linguistic features only.

5. Discussion

We discuss our results in context of model complexity, model generalisability and potential utility as a diagnostic modality. Our most effective models are DistilBERT with PAR+INV and SVM+CRF with PAR_SPLT. Both models perform similarly for the AD classification task, but the deep learning approach can also output MMSE score predictions. The DL methods will likely generalise better as the majority of the modelling is accomplished by the embedding layer. Both models could be deployed to mobile devices for a potentially ubiquitous early diagnostic tool.

In a potential diagnostic scenario, models would seek to balance recall and precision. A true-positive label of AD would prompt the user to seek a formal evaluation, under medical supervision, potentially leading to an earlier diagnosis allowing for slower progression of the disorder. However, ensuring the false-positive rate is low would minimise unnecessary anxiety during the following formal clinical evaluation.

5.1. Baseline Approaches

The SVM models report higher performance than the GBDT models across all metrics and tasks. They are also computationally faster to fit and cross validate. It is unclear if these models will generalise to alternative or larger datasets. The models have captured correlations in frequency of appearances of 'key words' as identified by TF-IDF vectoriser. Further datasets may result in variations in performance as the frequencies of 'key words' change providing insufficient signal for accurate modelling of the decision boundaries necessary for prediction.

Despite offering the worst performance across all configurations and datasets, GBDTs do provide good model interpretability. For example, we find the top 20 most informative word level features from the TF-IDF vectoriser contain discourse markers such as 'oh', 'uh' and 'um' for both tasks. This indicates the model has found occurrences of these words are useful in making predictions for both tasks. However, prior work has suggested more informative features are more complex[25].

5.2. Deep Learning Approaches

Our results are inline with previous work that has empirically shown that large, pre-trained, DL Transformer based, language models are an effective embedding layer that capture a variety of linguistic phenomenon[26]. As models are pre-trained we incur no model fit expense to use them. Training from scratch requires days or weeks with specialised hardware and large data sets. The simple LR or LASSO models that are fit on top of the fixed size output embeddings are as efficient to fit as the baseline SVM models.

BERT and RoBERTa models are available in their 'base' and 'large' varieties. In prior work, 'large' often performs better due to the increased parameter space and longer training time[22]. However, we observe in our experiments 'large' models are often equivalent or worse performing. We also observe this trend with DistilBERT / DistilRoBERTa that have further reduced parameters compared to 'base' varieties that broadly produce better results in our experiments although prior work would suggest the contrary.

6. Future Work

6.1. Further NLP Modelling

For future work we would look to replicate findings with larger datasets to demonstrate model robustness. We also currently use the models 'out-of-the-box' so they have only been trained with large corpora of prepared speech (Wikipedia and the Toronto Book Corpus[22]). For future work we would also look to fine-tune the deep learning embedding models specifically to spontaneous speech as fine-tuning to domain specific data often boosts performance[27]. Spontaneous speech corpora would likely show a difference in lexicon and grammar as well subtle prosodic differences such as rhythm, tempo that are often captured within spontaneous speech transcripts. An example of such a corpus is the 'The British National Corpus[28]'. A large corpora of informal spontaneous speech containing 1251 recordings and ∼11 million words from 668 speakers. We have cleaned and prepared the corpus using a sliding sentence window producing a dataset of ∼767k 'documents'. We successfully begun the fine-tuning process observing a reduction in training loss. However, due to extenuating circumstances our GPU resource became unavailable and we were unable to complete the fine-tuning. We make the data pre-processing, and language model fine-tuning scripts available open-source[2].

6.2. Feature Combinations and Model Ensembling

Combining features or model ensembling that incorporated the acoustic data (i.e. prosodic/articulatory features) may provide further gains in performance. Audible phenomenon such as changes in pitch, intonation, stress and subtle changes in tempo would only be available in the audio dataset and have shown to be useful during prior work[29, 25]. We leave the investigation and ensembling of these features to future work.

7. Conclusions

We have presented a range of NLP techniques applied to the ADReSS challenge dataset, a shared task for the prediction of AD and MMSE scores of AD patients and controls. Each dataset and model configuration is rigorously optimised and tested. We observe promising results, above published baselines, for machine learning techniques such as SVMs and Deep Learning approaches. We highlight that the Deep Learning approaches are particularly effective when used as embedding layers for both the AD classification and MMSE score prediction tasks even despite the lack of domain and task specific fine-tuning.

8. Acknowledgements

RD's work is supported by 1.National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. 2. Health Data Research UK, which is funded by the UK Medical Research Council, Engineering and Physical Sciences Research Council, Economic and Social Research Council, Department of Health and Social Care (England), Chief Scientist Office of the Scottish Government Health and Social Care Directorates, Health and Social Care Research and Development Division (Welsh Government), Public Health Agency (Northern Ireland), British Heart Foundation and Wellcome Trust. 3. The National Institute for Health Research University College London Hospitals Biomedical Research Centre. This paper represents independent research part funded by the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. The views expressed are those of the author(s) and not necessarily those of the NHS, MRC, NIHR or the Department of Health and Social Care.

[2] https://github.com/tomolopolis/ADReSS_Challenge/blob/master/Fine-Tune-LanguageModel.ipynb

9. References

[1] R. Mayeux and Y. Stern, "Epidemiology of alzheimer disease," Cold Spring Harb. Perspect. Med., vol. 2, no. 8, Aug. 2012.

[2] J. Rasmussen and H. Langerman, "Alzheimer's disease - why we need early diagnosis," Degener. Neurol. Neuromuscul. Dis., vol. 9, pp. 123–130, Dec. 2019.

[3] M. J. Kang, S. Y. Kim, D. L. Na, B. C. Kim, D. W. Yang, E.-J. Kim, H. R. Na, H. J. Han, J.-H. Lee, J. H. Kim, K. H. Park, K. W. Park, S.-H. Han, S. Y. Kim, S. J. Yoon, B. Yoon, S. W. Seo, S. Y. Moon, Y. Yang, Y. S. Shim, M. J. Baek, J. H. Jeong, S. H. Choi, and Y. C. Youn, "Prediction of cognitive impairment via deep learning trained with multi-center neuropsychological test data," BMC Med. Inform. Decis. Mak., vol. 19, no. 1, p. 231, Nov. 2019.

[4] C. M. Stonnington, C. Chu, S. Klöppel, C. R. Jack, Jr, J. Ashburner, R. S. J. Frackowiak, and Alzheimer Disease Neuroimaging Initiative, "Predicting clinical scores from magnetic resonance scans in alzheimer's disease," Neuroimage, vol. 51, no. 4, pp. 1405–1413, Jul. 2010.

[5] Z. Guo, Z. Ling, and Y. Li, "Detecting alzheimer's disease from continuous speech using language models," J. Alzheimers. Dis., vol. 70, no. 4, pp. 1163–1174, 2019.

[6] L. Zhou, K. C. Fraser, and F. Rudzicz, "Speech recognition in alzheimer's disease and in its assessment," in Interspeech, 2016, pp. 1948–1952.

[7] 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.

[8] C. de Boer, F. Mattace-Raso, J. van der Steen, and J. J. M. Pel, "Mini-Mental state examination subscores indicate visuomotor deficits in alzheimer's disease patients: A cross-sectional study in a dutch population," Geriatr. Gerontol. Int., vol. 14, no. 4, pp. 880–885, Oct. 2014.

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

[10] B. MacWhinney, "Tools for analyzing talk part 1: The chat transcription format," 2017.

[11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine learning in python," J. Mach. Learn. Res., vol. 12, no. Oct, pp. 2825–2830, 2011.

[12] H. P. Luhn, "A statistical approach to mechanized encoding and searching of literary information," IBM J. Res. Dev., vol. 1, no. 4, pp. 309–317, Oct. 1957.

[13] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," May 2011.

[14] C. Cortes and V. Vapnik, "Support-vector networks," Mach. Learn., vol. 20, no. 3, pp. 273–297, Sep. 1995.

[15] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ser. KDD '16. New York, NY, USA: Association for Computing Machinery, Aug. 2016, pp. 785–794.

[16] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," Ann. Stat., vol. 29, no. 5, pp. 1189–1232, 2001.

[17] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in Proceedings of the Eighteenth International Conference on Machine Learning, ser. ICML '01. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., Jun. 2001, pp. 282–289.

[18] S. Ruder, M. E. Peters, S. Swayamdipta, and T. Wolf, "Transfer learning in natural language processing," in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials, 2019, pp. 15–18.

[19] 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 26, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2013, pp. 3111–3119.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. U. Kaiser, and I. Polosukhin, "Attention is all you need," in Advances in Neural Information Processing Systems 30, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 5998–6008.

[21] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, and Others, "Huggingface's transformers: State-of-the-art natural language processing," ArXiv, abs/1910. 03771, 2019.

[22] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pretraining of deep bidirectional transformers for language understanding," Oct. 2018.

[23] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," Jul. 2019.

[24] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," Oct. 2019.

[25] 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.

[26] 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. Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 3651–3657.

[27] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," Jan. 2018.

[28] B. N. C. Consortium and Others, "The british national corpus, version 3 (BNC XML edition)," Distributed by Oxford University Computing Services on behalf of the BNC Consortium, vol. 5, no. 65, p. 6, 2007.

[29] A. König, A. Satt, A. Sorin, R. Hoory, O. Toledo-Ronen, A. Derreumaux, V. Manera, F. Verhey, P. Aalten, P. H. Robert, and R. David, "Automatic speech analysis for the assessment of patients with predementia and alzheimer's disease," Alzheimers. Dement., vol. 1, no. 1, pp. 112–124, Mar. 2015.