2021 IEEE 9th International Conference on Healthcare Informatics (ICHI) | 978-1-6654-0132-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICHI52183.2021.00094

Dementia Detection using Transformer-Based Deep Learning and Natural Language Processing Models

Ploypaphat Saltz
Dept. of Computing & SW Sys.
UW Bothell, WA, USA
saltzp@uw.edu

Shih Yin Lin
NYU Rory Meyers
College of Nursing
sl199@nyu.edu

Sunny Chieh Cheng
School of Nursing &
Healthcare Leadership
UW Tacoma, WA, USA
ccsunny@uw.edu

Dong Si*
Dept. of Computing & SW Sys.
UW Bothell, WA, USA
dongsi@uw.edu

Abstract—**Dementia is a disease characterized by cognitive impairment that leads to incoherent or illogical thoughts and speech. There are attempts to identify dementia through speech analyses, but there is a dearth of research on casual conversation analysis. This work examined communication impairment detection of people with early-stage memory loss, including mild dementia and mild cognitive impairment. The data sets included semi-structured interviews from two studies conducted at the University of Washington (UW), the DementiaBank's Pitt Corpus, and the ADReSS Challenge at INTERSPEECH 2020. We applied Transformer-based deep learning models to automatically extract linguistic features for identifying individuals with dementia. Our results showed the models' abilities on detecting linguistic deficits with the best mean F1-score of 76% on the Pitt Corpus, 84% on the ADReSS, 90% on the augmented ADReSS, and 74% on the UW transcripts. The results suggest the potential possibility of a more flexible examination setting, casual semi-structured individual or group interview, for detecting incoherent or illogical thoughts and speech in patients with dementia.**

Index Terms—**Dementia, Text Classification, Transformer, Natural Language Processing, Transfer Learning**

I. Introduction

Dementia is a debilitating chronic condition affecting approximately 88 million people worldwide [1]. The disease is characterized by progressive declines in one or more cognitive domains (e.g., complex attention, memory, language) severe enough to interfere with independence in everyday activities. There is a transitional stage between normal aging and dementia called mild cognitive impairment (MCI). Different from dementia, people with MCI retain independence in everyday activities despite their cognitive deficits. Because the conversion of MCI to dementia is common, researching language impairment in MCI also has implications for the early detection of dementia. Early detection of dementia is important because it allows the patients and families to plan and obtain access to relevant resources and support.

Linguistic features of language impairment, such as vocabulary richness and size, have been used for detection, classification, and staging of cognitive impairment, including dementia and MCI. Several works have applied Natural Language Processing (NLP) and Deep Learning (DL) techniques to construct automatic text classifier models for identifying dementia. [2], [3]. Some works leveraged pre-trained Transformer-based NLP models to analyze and identify dementia [3]–[5]. However, most studies experimented on structured examinations, and not many studies utilize semi-structured interviews. Semi-structured interviews are less restrictive and closer to conversations occurring during clinical visits. Therefore, detecting language impairment in dementia/MCI using semi-structured interviews may have greater implications for real-world clinical practice.

This work extends a study on the detection of dementia/MCI by comparing several pre-trained Transformer-based language learning models' efficiencies using both semi-structured interview and structured examination transcripts. We focused on the binary classifications utilizing the five pre-trained Transformer-based models: BERT, ALBERT, XLNet, RoBERTa, and ELECTRA. We classified the UW semi-structured interviews, the DementiaBank's Pitt (Cookie)[1], and the ADReSS Challenge at INTERSPEECH 2020 [6].

II. Methods

For manual linguistic analysis, we measured the vocabulary size produced by the participants attending the individual UW interviews using a Type-Token Ratio(TTR). The transcripts were composed of the group and individual interviews corpora. However, we excluded the group type because we anticipated that ambient communication during group interviews could interfere with personal vocabulary generation.

For the selected NLP models, we used their Transformer-based pre-trained models. Each model reserved 20% of the input train set and used it as the validation set during tuning. We used an 80:20 ratio for the train and test sets of the UW transcripts and Pitt Corpus. For the ADReSS, we used its initial pre-determined train and test data. For the augmented ADReSS, we augmented the training using the Random Deletion (RD) of the Easy Data Augmentation (EDA) [7] (n_{aug}= 4, α = 0.05). The motivations of testing these datasets are to compare performance when experimenting with semi-structured and structured interview transcripts. In addition to that, we were interested in observing the impact of data augmentation. The metrics to evaluate are the best accuracy during validation ($\mathrm{V_{Acc}}$) and prediction ($\mathrm{P_{Acc}}$), precision (Prec), recall (Rec), F1-score, and Matthews Correlation Coefficient (Mcc). We labeled

[1] https://talkbank.org/DementiaBank/

978-1-6654-0132-6/21/$31.00 ©2021 IEEE
DOI 10.1109/ICHI52183.2021.00094

Pathological Group (PG), early memory loss/Mild Cognitive Impairment (MCI), and Dementia Patient (DP), *Positive*; and Healthy Control (HC) *Negative*.

III. Results and discussion

To differentiate between groups using vocabulary size, we computed an unpaired t-test to compare the mean TTR of three subject pairs: HC/PG, HC/MCI, and HC/DP. The two-tailed P-values associated with the pairwise evaluations were 0.7516, 0.4717, and 0.8391, respectively. The results showed that they are not statistically significant (none of them less than 0.05).

Table I reports the mean of 5-fold cross-validation. We included other works' results that examined the same dataset using the same models. Note that other work used different settings and validation methods. The classification results of the UW corpus show that the rankings between models are not consistent, but all of them could remain strong classification records; classification scores are greater than .51. When classifying the Pitt Corpus and the ADReSS data using our models, the overall performance is generally strong but still weaker than other works. Furthermore, it is noticeable that classifying structured interviews yields slightly better results than when classifying the semi-structured interview transcripts (using the same models and settings). When applying augmenting to the ADReSS data, which is also balanced (labeled ADReSS2020_{Aug}), we can notice some performance gains.

TABLE I
The **mean** of 5-fold evaluation of the models against datasets

Data	Model	V_{Acc}	P_{Acc}	Prec	Rec	F1	Mcc
UW HC/PG	BERT	.85	.63	.70	.60	.60	.31
	ALBERT	.66	**.70**	.84	.67	**.69**	.44
	XLNet	.69	.57	.63	.52	.52	.50
	RoBERTa	.69	.59	.83	.42	.51	.27
	ELECTRA	.74	.61	.76	.50	.56	.30
UW HC/MCI	BERT	.90	.76	.83	.75	.72	.55
	ALBERT	.85	.75	.86	.69	.71	.44
	XLNet	.88	**.77**	.90	.64	**.74**	.59
	RoBERTa	.85	.71	.91	.55	.65	.48
	ELECTRA	.92	**.78**	.96	.61	**.74**	.62
UW HC/DP	BERT	.87	.73	.82	.70	.69	.49
	ALBERT	.83	.72	.85	.63	.68	.50
	XLNet	.86	**.75**	.85	.65	**.72**	.55
	RoBERTa	.83	.68	.88	.56	.62	.44
	ELECTRA	.88	.71	.73	.82	.68	.51
Pitt Cookie HC/PG	BERT	.74	.80	.73	.73	.74	.50
	ALBERT	.73	.72	.85	.61	.72	.48
	XLNet	.80	**.76**	.89	.67	**.76**	.56
	RoBERTa	.74	.74	.88	.63	.73	.53
	ELECTRA	.76	.75	.88	.67	.74	.54
	BERT [3]	-	.84	.86	.85	.82	-
	RoBERTa [3]	-	.75	.79	.72	.74	-
	BERT [8]	-	.84	.90	.76	.82	-
	XLNet [8]	-	.80	.83	.74	.78	-
ADReSS 2020 HC/PG	BERT	.74	.80	.73	.73	.74	.50
	ALBERT	.73	.72	.85	.61	.72	.48
	XLNet	.80	**.76**	.89	.67	**.76**	.56
	RoBERTa	.74	.74	.88	.63	.73	.53
	ELECTRA	.76	.75	.88	.67	.74	.54
	BERT [4]	-	.83	.81	.88	.84	-
	RoBERTa [2]	-	.92	-	-	.92	-
ADReSS 2020_{Aug} HC/PG	BERT	.83	**.90**	.88	.92	**.90**	.78
	ALBERT	.93	.85	1.0	.71	.84	.74
	XLNet	.77	.82	.94	.71	.83	.68
	RoBERTa	.85	.81	.76	.92	.81	.63
	ELECTRA	.78	.85	.95	.75	.85	.72

IV. Conclusion

The TTR evaluation of the UW individual interviews corpus could not differentiate the dementia patients from the healthy subjects efficiently. On the other hand, the selected pre-trained transformer-based models show that they can identify dementia. Testing against the semi-structured transcripts, BERT, XLNet, and ELECTRA often took the lead, and XLNet took the lead most frequently. For the ranking inconsistency across experiments of our models, we anticipate that randomizing and splitting the training and validating datasets during tuning can cause the altered overall performance since randomizing was involved. However, the results confirm the capability of the selected NLP in detecting crucial linguistic features necessary to identify dementia. Compared to other work, the results suggest that our design can be optimized. For example, the RD augmentation on the ADReSS training set enhanced the performance slightly. Even though the performance improvement after applying the RD was slim, the output recommended that the augmentation has boosted the performance on a small dataset. This notion is valuable since a small sample size is a typical limitation of dementia study.

Our overall experimental results suggest that these pre-trained Transformer-based models can automatically detect language impairments from both structured (i.e., diagnostic assessment) and semi-structured interviews. Regarding the limitation of this work, although the UW interview data was subjective, the strength of this study is that the data included patients with early-stage memory and mild cognitive impairment. Thus, our findings could be more generalized to the broader population with dementia than the data collected based on restrictive dementia and MCI eligibility criteria.

References

[1] Alzheimer's Association, "2020 alzheimer's disease facts and figures," *Alzheimer's & Dementia*, vol. 16, no. 3, pp. 391–460, 2020.

[2] J. Koo, J. H. Lee, J. Pyo, Y. Jo, and K. Lee, "Exploiting multi-modal features from pre-trained networks for alzheimer's dementia recognition," 2020.

[3] T. Searle, Z. Ibrahim, and R. Dobson, "Comparing natural language processing techniques for alzheimer's dementia prediction in spontaneous speech," *arXiv preprint arXiv:2006.07358*, 2020.

[4] A. Balagopalan, B. Eyre, F. Rudzicz, and J. Novikova, "To bert or not to bert: Comparing speech and language-based approaches for alzheimer's disease detection," 2020.

[5] A. Roshanzamir, H. Aghajan, and M. S. Baghshah, "Transformer-based deep neural network language models for alzheimer's disease detection from targeted speech," *learning*, vol. 1, p. 2, 2020.

[6] 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. [Online]. Available: https://arxiv.org/abs/2004.06833

[7] J. Wei and K. Zou, "Eda: Easy data augmentation techniques for boosting performance on text classification tasks," *arXiv preprint arXiv:1901.11196*, 2019.

[8] A. Roshanzamir, H. Aghajan, and M. S. Baghshah, "Transformer-based deep neural network language models for alzheimer's disease risk assessment from targeted speech," *BMC Medical Informatics and Decision Making*, vol. 21, no. 1, pp. 1–14, 2021.