Robustness and Sensitivity of BERT Models Predicting Alzheimer's Disease from Text

Jekaterina Novikova
Winterlight Labs / Toronto, Canada
jekaterina@winterlightlabs.com

Abstract

Understanding robustness and sensitivity of BERT models predicting Alzheimer's disease from text is important for both developing better classification models and for understanding their capabilities and limitations. In this paper, we analyze how a controlled amount of desired and undesired text alterations impacts performance of BERT. We show that BERT is robust to natural linguistic variations in text. On the other hand, we show that BERT is not sensitive to removing clinically important information from text.

1 Introduction

Alzheimer's disease (AD) is a prevalent neurodegerative condition that inhibits cognitive abilities and impacts one's language abilities. For example, cognitively impaired people tend to use more pronouns instead of nouns, and pause more often between sentences in narrative speech (Roark et al., 2011). This insight makes automatic detection possible. Machine learning (ML) classifiers can detect cognitive impairments given descriptive linguistic features or using pre-trained large language models (Balagopalan et al., 2018; Zhu et al., 2019).

BERT is a model that achieves promising performance on a variety of tasks, including AD prediction from speech and language (Searle et al., 2020; Yuan et al., 2020). However, this promising performance may be fallacious, i.e. deep neural language models may learn pseudo patterns from training data to attain high performance on test sets (Goyal et al., 2019; Gururangan et al., 2018; Glockner et al., 2018; Tsuchiya, 2018; Geva et al., 2019). Therefore, in order to be confident in the outcomes of BERT models classifying AD it is important to assess whether these models are robust to some natural noise that may be introduced in language. It is also important to know if BERT models are sensitive to the aspects that are considered to be important for recognizing cognitive impairment from human language.

In this paper, we analyze robustness and sensitivity of BERT models in their ability to classify AD from text by analysing the effect of noise, introduced from artificial text perturbations, on the performance of the model. Some previous research was conducted on the impact of ASR-related noise on dementia detection (Balagopalan et al., 2020b), as well as the effect of artificial text alterations on AD classification (Novikova et al., 2019). However, these previous studies only focus on conventional classification models, such as Random Forest and SVM. To the best of our knowledge, we are the first to analyse how the noise introduced by texts perturbations impact BERT models, in the domain of AD classification.

2 Methodology

2.1 Data

We use the ADReSS Challenge dataset (Luz et al., 2020), 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 (Goodglass et al., 2001). In contrast to other datasets for AD detection such as DementiaBank's English Pitt Corpus (Becker et al., 1994), the ADReSS challenge dataset is well balanced in terms of age and gender (Table 1). Another benefit of this dataset is its division into standard train and test sets that makes it easy to directly compare to the previous research in the area.

2.2 Model

Multiple recent studies showed that BERT is a promising model achieving strong enough performance in detecting Alzheimer's disease from transcribed speech (Searle et al., 2020; Yuan et al., 2020; Balagopalan et al., 2020a, 2021). Motivated

Table 1: Basic characteristics of the patients in each group in the ADReSS challenge dataset.

Dataset			Class	
			AD	Non-AD
ADReSS	Train	Male	24	24
		Female	30	30
ADReSS	Test	Male	11	11
		Female	13	13

by these results, we use a fine-tuned BERT (Devlin et al., 2019) model in this work. To leverage the language information encoded by BERT (Devlin et al., 2019), we add a linear layer mapping representations from the final layer of a pre-trained 12-layer BERT base[1] for the AD vs non-AD binary 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.* (Liu and Lapata, 2019). A pooled embedding summarizing information across all tokens in the transcript is used as the aggregate transcript representation, and passed to the classification layer (Devlin et al., 2019; Wolf et al., 2019). This model is then finetuned on training data for AD detection. For hyperparameter tuning, we optimize the number of epochs to 10 by varying it from 1 to 12 during cross-validation. Adam optimizer (Kingma and Ba, 2014) and warmup linear learning rate scheduling (Paszke et al., 2019) are used, based on prior work on fine-tuning BERT (Devlin et al., 2019; Wolf et al., 2019).

2.3 Perturbation Approaches

We used a variety of word-based augmentation approaches with the help of the nlpaug[2] library to generate perturbed versions of the test set of the ADReSS dataset for the experiments.

Back-translation: this augmentation technique proposed by Sennrich et al. (2016) leverages two translation models, one translating the source text from English to German and the other translating it back to English. Back-translated texts should maintain the semantics and basic syntactic structure of original texts and as such, robust model's performance should not decrease because of this augmentation.

Word substitution with synonyms: following Niu and Bansal (2018), we substitute a controlled varying amount of words in the transcript (10-90%) with their synonyms in order to maintain semantic meaning of the utterances. Synonyms are extracted from the NLTK WordNet corpus[3]. Replacing words with their synonyms should not affect the ability of a robust model to accurately distinguish between healthy and AD classes.

Embedding-based word substitution: following Alzantot et al. (2018); Wang and Yang (2015), we use pre-trained word2vec embeddings to perform a KNN with cosine similarity search to find the similar word for replacement. We then substitute a varying subsets (from 10 to 90%) of the original transcripts with these replacements. We hypothesize that model performance can be affected by such augmentation stronger than by synonym replacement, although this effect should not be significant for a robust AD prediction model.

Removal of filled pauses: we remove all the filled pauses (transcribed as *um* and *uh*) from the original texts. Previous literature highlights the importance of pauses in Alzheimer's disease detection from speech (Calley et al., 2010; Mack et al., 2013; Seifart et al., 2018). Several authors report increases in AD detection performance by extracting acoustic features such as filled pause counts (Eyre et al., 2020; Tóth et al., 2015, 2018; Pistono et al., 2016). Removal of such information should make it more difficult for a model to accurately detect AD-related samples of text.

Removal of information units: multiple studies of AD narratives in picture description tasks have reported the importance of information units in detecting cognitive impairment (Fraser et al., 2016; Croisile et al., 1996). Following (Croisile et al., 1996), we define four key categories of information units - subjects, locations, objects, and actions - and delete them from the original transcripts to generate perturbed versions of the test set. Such a removal should make it more difficult for a model to distinguish between healthy and AD samples.

3 Results

The results of testing the fine-tuned BERT model on the variety of perturbed versions of the ADReSS test set show that the performance changes differently depending on different types of text alterations (Table 2). Removing tokens of filled pauses does not change the performance at all. Removing

[1] https://huggingface.co/bert-base-uncased

[2] https://github.com/makcedward/nlpaug , the Python library for generating synthetic textual and speech data.

[3] https://www.nltk.org/howto/wordnet.html

Type of perturbation	Level of perturbation	Acc	F1	Prec	Rec	Spec	W_1
Original transcript	NA	0.83	0.83	0.86	0.79	0.88	NA
Deleting filled pauses	All	0.83	0.83	0.86	0.79	0.88	2.40
Deleting information units	All	0.81	0.84	0.74	0.96	0.67	2.87
	Action	0.77	0.80	0.71	0.92	0.63	2.87
	Location	0.77	0.78	0.74	0.83	0.71	0.77
	Object	0.75	0.79	0.69	0.92	0.58	2.45
	Subject	0.79	0.79	0.79	0.79	0.79	2.13
Back translation	Eng <->DE	0.75	0.75	0.75	0.75	0.75	6.02
Substituting with the most similar word (via word2vec embeddings)	10%	0.83	0.84	0.81	0.88	0.79	5.63
	20%	0.81	0.82	0.78	0.88	0.75	6.23
	30%	0.81	0.82	0.80	0.83	0.79	6.23
	40%	0.81	0.82	0.78	0.88	0.75	6.32
	50%	0.81	0.80	0.86	0.75	0.88	6.20
	60%	0.83	0.84	0.81	0.88	0.79	6.10
	70%	0.71	0.70	0.73	0.67	0.75	6.32
	80%	0.75	0.75	0.75	0.75	0.75	6.35
	90%	0.77	0.78	0.76	0.79	0.75	6.26
Substituting synonyms (via WordNet)	10%	0.77	0.79	0.72	0.88	0.67	3.55
	20%	0.75	0.75	0.75	0.75	0.75	3.97
	30%	0.77	0.78	0.76	0.79	0.75	4.12
	40%	0.75	0.77	0.71	0.83	0.67	4.01
	50%	0.75	0.75	0.75	0.75	0.75	3.86
	60%	0.71	0.70	0.73	0.67	0.75	4.33
	70%	0.81	0.82	0.80	0.83	0.79	3.98
	80%	0.81	0.82	0.78	0.88	0.75	4.08
	90%	0.77	0.78	0.74	0.83	0.71	4.16

Table 2: Performance of the fine-tuned BERT model and similarity between original and perturbed texts (Wasserstein distance W_1).

information units, however, decreases the accuracy of the model by 4-8%, depending on the type of the information unit. Back translation and substitutions of words with their synonyms or otherwise similar words also negatively affect performance of the model, although have the opposite effect on recall vs specificity.

4 Discussion

4.1 Change in Classification Performance

Undesired change: As we have mentioned in Section 2.3, some types of text alterations, such as back translation, synonym substitution and embedding-based substitution, represent natural noise that can occur in user-generated texts. Changes in classification performance are not desired in this case because we want the model to be robust towards multiple paraphrases and use of synonyms. Our fine-tuned BERT model behaves in a robust way in terms of F1 and accuracy scores when up to 40% of words are substituted with similar words based on word2vec embeddings (the F1 score decreases by 1% and accuracy - by 2%, both changes not significant with McNemar's test p >0.65) or synonyms (a decrease in 4-8% in F1 and accuracy, both changes not significant with McNemar's test p >0.15).

Type of perturbation	Correlation between W_1 and Acc	F1	Prec	Rec	Spec
Deleting informational units	0.24	0.56	-0.07	0.52	-0.21
Substituting with the most similar word (via word2vec embeddings)	0.77	0.89	0.89	0.87	-0.91
Deep representation (VGG-16)	-0.26	-0.37	0.10	-0.45	0.41

Table 3: Correlation between similarity and performance metrics.

Interestingly, recall and specificity values seem to show the opposite results here - specificity decreases by up to 21%, while recall stays on the same level or even increases by 4-9%. Substituting words with their similar alternatives or synonyms may be understood as increasing the level of lexical complexity, i.e. the model is introduced with multiple, maybe less usual, ways to express the same meaning. It is known that lexical complexity is one of the prominent ways that allow detecting cognitive impairment from language. Thus such an implicit way to change the lexical complexity of texts seems to help the BERT model in reducing the amount of true positive errors while detecting AD. However, more than 40% of such substitutions may make the original texts less realistic, which, as we see from the results, substantially reduces model performance, including reducing recall level.

Desired change: Other types of text alterations, such as removal of information units or tokens representing filled pauses, are not considered to be natural noise. As these characteristics of language are clinically important in detecting cognitive impairment, the models should be sensitive to such changes in language. Our results show that the fine-tuned BERT model ignores completely removal of filled pauses. Performance of the model decreases by 3-5% of F1 as a reaction to deleting different types of information units but this change is not significant (McNemar's test p >0.18). This change in performance is similar to the change caused by synonym substitution and shows that the model is not sensitive enough to removal of clinically relevant information.

This leads us to inspect how each type of alterations affects distributional shift from the original text and whether there is a relation between the shift and model's performance.

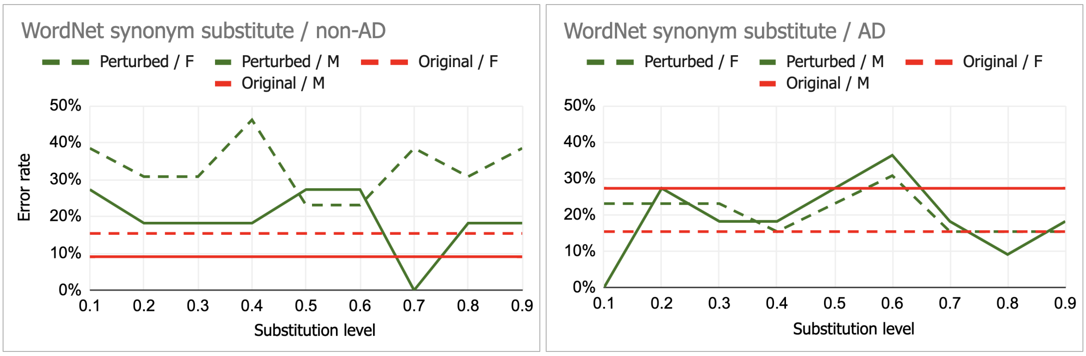

Figure 1: Differences in error rate between genders, by class. Here, M means 'male' and F means 'female'.

4.2 Correlation with Distributional Shift

Inspired by (Lee et al., 2018) where hidden activations were used to detect out-of-distribution samples for images and by Rychener et al. (2020) applied this method to text, we used sentence embeddings produced by BERT to quantify the distributional shift among the original test set and its perturbed versions. To understand the level of dissimilarity among the versions of test sets, we calculated the 1-Wasserstein distance ("earth mover distance", W_1), since it measures the minimum cost to turn one probability distribution into another (Table 2).

W_1 values show that deletion of information units and filled pauses has the lowest effect on the original text, while word2vec-based substitution shifts the distribution further away from the original. Correlation between performance metrics and W_1 is not consistent across different types of text alterations (Table 3): it is strongly positive between F1 and accuracy scores in case of embedding-based substitutions (0.77 and 0.89), positive but less strong (0.24 and 0.56) in case of deleting clinically relevant information, and negative in case of synonym substitution (-0.26 and -0.37). These inconsistencies imply that the lack of sensitivity in BERT models is caused by intrinsic model reasons rather than distributional shift of test data.

4.3 Differences Based on Gender

In order to understand if BERT performance is biased towards any gender, we analyse the rate of error within each gender group and how the error rate is changing with additional amount of text alterations. The results of this analysis do not reveal any differences between males and females within the class of AD data samples. However when it comes to the non-AD class, BERT tends to misclassify the text samples produced by female subjects significantly more often than those produced by males, across all types of text alterations. The effect is pronounced the most in the case of synonym substitution (see Figure 1), where the error rate of classifying female-produced samples is 14% higher on average than that of male-produced samples[4]. Given that both training and test sets of the dataset are well balanced, such a difference implies the pre-trained BERT model is gender-biased and this bias is not eliminated during fine-tuning.

5 Limitations and Conclusions

In this work, we analysed how the controlled amount of desired and undesired text alterations impacts BERT classification performance in the domain of AD detection. We showed that BERT is robust enough to the natural linguistic noise, although the model is biased towards text samples of non-AD females. On the other hand, BERT is not sensitive enough to removal of clinically relevant information. This lack of sensitivity is not directly influenced by distributional shift.

This work is a first step towards investigating BERT models' robustness and sensitivity in the domain of AD detection from text, and we only report empirical results of one BERT model fine-tuned and tested on one dataset. More work should be done in this area to ensure the results are widely generalizable within the domain. Textual data used in our experiment represent transcribed conversational speech and as such, may be quite different from other types of texts, e.g. written text. Future work is necessary to see if the effect of text alterations remain the same with other types of text.

[4] Also significantly different based on t-test, $p < 0.005$.

References

Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. 2018. Generating natural language adversarial examples. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2890–2896, Brussels, Belgium. Association for Computational Linguistics.

Aparna Balagopalan, Benjamin Eyre, Jessica Robin, Frank Rudzicz, and Jekaterina Novikova. 2021. Comparing pre-trained and feature-based models for prediction of alzheimer's disease based on speech. *Frontiers in aging neuroscience*, 13:189.

Aparna Balagopalan, Benjamin Eyre, Frank Rudzicz, and Jekaterina Novikova. 2020a. To BERT or not to BERT: Comparing Speech and Language-Based Approaches for Alzheimer's Disease Detection. In *Proc. Interspeech 2020*, pages 2167–2171.

Aparna Balagopalan, Jekaterina Novikova, Frank Rudzicz, and Marzyeh Ghassemi. 2018. The effect of heterogeneous data for Alzheimer's disease detection from speech. *arXiv preprint arXiv:1811.12254*.

Aparna Balagopalan, Ksenia Shkaruta, and Jekaterina Novikova. 2020b. Impact of asr on alzheimer's disease detection: All errors are equal, but deletions are more equal than others. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 159–164.

James T Becker, François Boiler, Oscar L Lopez, Judith Saxton, and Karen L McGonigle. 1994. The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis. *Archives of Neurology*, 51(6):585–594.

Clifford S Calley, Gail D Tillman, Kyle Womack, Patricia Moore, John Hart Jr, and Michael A Kraut. 2010. Subjective report of word-finding and memory deficits in normal aging and dementia. *Cognitive and behavioral neurology: official journal of the Society for Behavioral and Cognitive Neurology*, 23(3):185.

Bernard Croisile, Bernadette Ska, Marie-Josee Brabant, Annick Duchene, Yves Lepage, Gilbert Aimard, and Marc Trillet. 1996. Comparative study of oral and written picture description in patients with alzheimer's disease. *Brain and language*, 53(1):1–19.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186.

Ben Eyre, Aparna Balagopalan, and Jekaterina Novikova. 2020. Fantastic features and where to find them: Detecting cognitive impairment with a subsequence classification guided approach. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 193–199, Online. Association for Computational Linguistics.

Kathleen C Fraser, Jed A Meltzer, and Frank Rudzicz. 2016. Linguistic features identify alzheimer's disease in narrative speech. *Journal of Alzheimer's Disease*, 49(2):407–422.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. 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)*, pages 1161–1166, Hong Kong, China. Association for Computational Linguistics.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking NLI systems with sentences that require simple lexical inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–655, Melbourne, Australia. Association for Computational Linguistics.

Harold Goodglass, Edith Kaplan, and Barbara Barresi. 2001. *BDAE-3: Boston Diagnostic Aphasia Examination–Third Edition*. Lippincott Williams & Wilkins Philadelphia, PA.

Yash Goyal, Tejas Khot, Aishwarya Agrawal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2019. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. *International Journal of Computer Vision*, 127(4):398–414.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. 2018. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 7167–7177.

Yang Liu and Mirella Lapata. 2019. 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), pages 3721–3731.

Saturnino Luz, Fasih Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney. 2020. Alzheimer's Dementia Recognition through Spontaneous Speech: The ADReSS Challenge.

Jennifer Mack, Aya Meltzer-Asscher, Sarah Dove, Sandra Weintraub, Marsel Mesulam, and Cynthia K Thompson. 2013. Word-finding pauses in primary progressive aphasia (ppa): Effects of lexical category.

Tong Niu and Mohit Bansal. 2018. Adversarial oversensitivity and over-stability strategies for dialogue models. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 486–496.

Jekaterina Novikova, Aparna Balagopalan, Ksenia Shkaruta, and Frank Rudzicz. 2019. Lexical features are more vulnerable, syntactic features have more predictive power. *W-NUT 2019*, page 431.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035.

Aurélie Pistono, Mélanie Jucla, Emmanuel J Barbeau, Laure Saint-Aubert, Béatrice Lemesle, Benjamin Calvet, Barbara Köpke, Michèle Puel, and Jérémie Pariente. 2016. Pauses during autobiographical discourse reflect episodic memory processes in early alzheimer's disease. *Journal of Alzheimer's Disease*, 50(3):687–698.

Brian Roark, Margaret Mitchell, John-Paul Hosom, Kristy Hollingshead, and Jeffrey Kaye. 2011. Spoken language derived measures for detecting mild cognitive impairment. *IEEE transactions on audio, speech, and language processing*, 19(7):2081–2090.

Yves Rychener, Xavier Renard, Djamé Seddah, Pascal Frossard, and Marcin Detyniecki. 2020. Sentence-based model agnostic nlp interpretability. *arXiv preprint arXiv:2012.13189*.

Thomas Searle, Zina Ibrahim, and Richard Dobson. 2020. Comparing Natural Language Processing Techniques for Alzheimer's Dementia Prediction in Spontaneous Speech. In *Proc. Interspeech 2020*, pages 2192–2196.

Frank Seifart, Jan Strunk, Swintha Danielsen, Iren Hartmann, Brigitte Pakendorf, Søren Wichmann, Alena Witzlack-Makarevich, Nivja H de Jong, and Balthasar Bickel. 2018. Nouns slow down speech across structurally and culturally diverse languages. *Proceedings of the National Academy of Sciences*, 115(22):5720–5725.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Laszló Tóth, Gábor Gosztolya, Veronika Vincze, Ildikó Hoffmann, Gréta Szatlóczki, Edit Biró, Fruzsina Zsura, Magdolna Pákáski, and János Kálmán. 2015. Automatic detection of mild cognitive impairment from spontaneous speech using asr. In *Sixteenth Annual Conference of the International Speech Communication Association*.

László Tóth, Ildikó Hoffmann, Gábor Gosztolya, Veronika Vincze, Gréta Szatlóczki, Zoltán Bánréti, Magdolna Pákáski, and János Kálmán. 2018. A speech recognition-based solution for the automatic detection of mild cognitive impairment from spontaneous speech. *Current Alzheimer Research*, 15(2):130–138.

Masatoshi Tsuchiya. 2018. Performance impact caused by hidden bias of training data for recognizing textual entailment. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

William Yang Wang and Diyi Yang. 2015. That's so annoying!!!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using #petpeeve tweets. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2557–2563, Lisbon, Portugal. Association for Computational Linguistics.

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

Jiahong Yuan, Yuchen Bian, Xingyu Cai, Jiaji Huang, Zheng Ye, and Kenneth Church. 2020. Disfluencies and Fine-Tuning Pre-Trained Language Models for Detection of Alzheimer's Disease. In *Proc. Interspeech 2020*, pages 2162–2166.

Zining Zhu, Jekaterina Novikova, and Frank Rudzicz. 2019. 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, Volume 1 (Long and Short Papers)*, pages 1431–1441.