



Crossing the “Cookie Theft” Corpus Chasm: Applying What BERT Learns From Outside Data to the ADReSS Challenge Dementia Detection Task

Yue Guo^{1*}, Changye Li², Carol Roan³, Serguei Pakhomov² and Trevor Cohen¹

¹ Department of Biomedical Informatics and Medical Education, University of Washington, Seattle, WA, United States,

² Pharmaceutical Care and Health Systems, University of Minnesota, Minneapolis, MN, United States, ³ Department of Sociology, University of Wisconsin-Madison, Madison, WI, United States

Large amounts of labeled data are a prerequisite to training accurate and reliable machine learning models. However, in the medical domain in particular, this is also a stumbling block as accurately labeled data are hard to obtain. DementiaBank, a publicly available corpus of spontaneous speech samples from a picture description task widely used to study Alzheimer’s disease (AD) patients’ language characteristics and for training classification models to distinguish patients with AD from healthy controls, is relatively small—a limitation that is further exacerbated when restricting to the balanced subset used in the Alzheimer’s Dementia Recognition through Spontaneous Speech (ADReSS) challenge. We build on previous work showing that the performance of traditional machine learning models on DementiaBank can be improved by the addition of normative data from other sources, evaluating the utility of such extrinsic data to further improve the performance of state-of-the-art deep learning based methods on the ADReSS challenge dementia detection task. To this end, we developed a new corpus of professionally transcribed recordings from the Wisconsin Longitudinal Study (WLS), resulting in 1366 additional Cookie Theft Task transcripts, increasing the available training data by an order of magnitude. Using these data in conjunction with DementiaBank is challenging because the WLS metadata corresponding to these transcripts do not contain dementia diagnoses. However, cognitive status of WLS participants can be inferred from results of several cognitive tests including semantic verbal fluency available in WLS data. In this work, we evaluate the utility of using the WLS ‘controls’ (participants without indications of abnormal cognitive status), and these data in conjunction with inferred ‘cases’ (participants with such indications) for training deep learning models to discriminate between language produced by patients with dementia and healthy controls. We find that incorporating WLS data during training a BERT model on ADReSS data improves its performance on the ADReSS dementia detection task, supporting the hypothesis that incorporating WLS data adds value in this context. We also demonstrate that weighted cost functions and additional prediction targets may be effective ways to address issues arising from class imbalance and confounding effects due to data provenance.

Keywords: dementia diagnosis, Alzheimer’s disease, natural language processing, BERT, machine learning

OPEN ACCESS

Edited by:

Saturnino Luz,
University of Edinburgh,
United Kingdom

Reviewed by:

Maria Koutsombogera,
Trinity College Dublin, Ireland
Heidi Christensen,
The University of Sheffield,
United Kingdom

*Correspondence:

Yue Guo
yguo50@uw.edu

Specialty section:

This article was submitted to
Human-Media Interaction,
a section of the journal
Frontiers in Computer Science

Received: 16 December 2020

Accepted: 23 March 2021

Published: 16 April 2021

Citation:

Guo Y, Li C, Roan C, Pakhomov S and
Cohen T (2021) Crossing the “Cookie
Theft” Corpus Chasm: Applying What
BERT Learns From Outside Data to
the ADReSS Challenge Dementia
Detection Task.
Front. Comput. Sci. 3:642517.
doi: 10.3389/fcomp.2021.642517

1. INTRODUCTION

Alzheimer's Dementia (AD) is a debilitating condition with few symptomatic treatments and no known cure. According to the Alzheimer's Association, in 2018 an estimated 5.8 million Americans were living with AD (Association, 2019). By 2050, these numbers are projected to increase to 14 million people with AD at a cost of \$1.1 trillion per year (Association, 2019). Diagnosis of this condition is often missed or delayed (Bradford et al., 2009), and delays may occur over an extended period with cognitive changes anticipating future dementia preceding clinical diagnosis by as many as 18 years (Rajan et al., 2015; Aguirre-Acevedo et al., 2016). Earlier diagnosis of AD has the potential to ease the burden of disease on patients and caregivers by reducing family conflict and providing more time for financial and care planning (Boise et al., 1999; Bond et al., 2005; Stokes et al., 2015). Delayed diagnosis of this condition also contributes substantively to the cost of care of this disease on account of a high utilization of emergency rather than routine care, amongst other factors—it is estimated that early and accurate diagnosis can help save an estimated \$7.9 trillion in medical and care costs (Association, 2018). Furthermore, survey findings show the vast majority (~80%) would prefer to know if their unexplained symptoms of confusion or memory loss were due to AD dementia in a formal clinical evaluation (Blendon et al., 2011).

One path to earlier diagnosis of AD involves the application of machine learning methods to transcribed speech, with the publicly available DementiaBank corpus (Becker et al., 1994) providing a focal point for research in this area. The majority of this prior work has involved the application of supervised machine learning methods (see e.g., Orimaye et al., 2014, 2017, 2018; Fraser et al., 2016; Yancheva and Rudzicz, 2016; Karlekar et al., 2018; Cohen and Pakhomov, 2020) to classify groups of transcripts, specific transcripts or even individual utterances as to whether or not the participants producing them were clinically diagnosed with dementia. While many of the methods developed during the course of this research exhibited promising performance, their performance is not strictly comparable on account of differences in units of analysis, restrictions on the inclusion of participants, evaluation metrics and cross-validation strategies. Furthermore, the DementiaBank dataset was constructed without case/control matching, resulting in statistically significant differences in age and level of education across the AD and control groups. Consequently there is a danger that diagnostic performance of classifiers trained and evaluated on this set may be overestimated on account of their ability to learn to recognize these differences, rather than linguistic indicators of AD.

2. BACKGROUND

The ADReSS challenge reference set was deliberately constructed to remediate some of these issues with the original data (Luz et al., 2020). This dataset represents a subset of the DementiaBank data, matched for age and gender, with enhancement of the accompanying audio data, and containing only a single transcript for each participant (as opposed to the multiple transcripts corresponding to multiple study visits per participant available

in the original set). As has been noted by the developers of the ADReSS dataset, it has the potential to advance the field by providing a standardized set for comparison between methods, which is a welcome advance on account of previously published work in this area often using different subsets of DementiaBank, as well as different cross-validation strategies and performance metrics. The ADReSS set and the accompanying challenge task present a standardized approach to evaluation on two tasks—AD recognition and Mini-mental State Exam (MMSE) prediction—for comparative evaluation moving forward. However, it is also true that this subset is even smaller in size than the original DementiaBank set, with only 108 training examples and 54 test examples, both split equally between healthy controls and participants with AD dementia.

In previous work, Noorian et al. (2017) demonstrated that the performance of machine learning approaches in the context of the DementiaBank set can be improved by providing the models concerned with additional “Cookie Theft” transcripts derived from other datasets. In this work, the authors introduced two additional sets of transcripts: Talk2Me and WLS. The former is an internal collection, while the latter is drawn from the Wisconsin Longitudinal Study (Herd et al., 2014), an extended study of a sample of students graduating from high school in Wisconsin 1957 born between 1938 and 1940 (initial $n = 10,317$), with some participants performing the “Cookie Theft” picture description task in a subsequent 2011 survey, aged in their early seventies. The authors report the availability of an additional 305 and 1,366 transcripts from participants without AD in the Talk2Me and WLS sets, respectively. In both cases, only recordings were available for analysis—text features were extracted using the Kaldi open source Automated Speech Recognition (ASR) engine (Povey et al., 2011), with an estimated word error rate of ~12.5% on the Talk2Me data, and none provided for the WLS set. As the additional data were considered as controls, the ADASYN (He et al., 2008) synthetic sampling method was used to oversample the minority “dementia” class. On a random 80/20 train/test split of the DementiaBank data, the authors report a considerable advantage in performance with the addition of the WLS controls in particular, with improvements of over 10% (absolute) in macro-averaged F-measure across a range of machine learning methods trained on a set of 567 manually engineered features, with oversampling offering an advantage over training without balancing the set in some but not all methods.

In this paper we evaluate the extent to which the performance of contemporary deep learning architectures can benefit from the addition of data from the WLS set. After attaining the relevant institutional approvals, we obtained all available “Cookie Theft” recordings from the WLS collection, as well as professional transcriptions of these recordings, to obviate the need to consider ASR error in our subsequent analyses. We evaluate the utility of the resulting transcripts as a means to improve performance of transfer learning using pre-trained Transformer-based architectures (Vaswani et al., 2017), focusing on the widely-used Bidirectional Encoder Representations from Transformers (BERT) model (Devlin et al., 2018) that has been shown to outperform other machine learning methods on the ADReSS challenge in recent work (Balagopalan et al., 2020).

Combining text corpora drawn from different sources to train NLP models should be approached with caution. Recent NLP research has identified and attempted to address the potentially deleterious role of confounding variables in text classification (Landeiro and Culotta, 2018). A confounding variable is a variable that can influence both a predictor and an outcome of a predictive model. One manifestation of the issue of confounding in NLP concerns a scenario in which data are drawn from different sources (Howell et al., 2020), each with different underlying class distributions. The WLS and ADReSS sets exemplify this problem. The ADReSS set is balanced by design, with an equal number of case and control transcripts. However, while some indication of cognitive impairment can be inferred from the metadata that accompanied the WLS recordings, the control transcripts vastly outnumber the cases in which data from cognitive tasks indicates such impairment. Consequently, if differences in language use across the populations from which these datasets are drawn were to permit a machine learning model to distinguish between the two sets, such a model may approach its optimization objective of accurate classification by simply learning to label all WLS examples as controls. In this context, the provenance of a transcript serves as a confounding variable, because it influences both the intended predictors (words in the transcript) and the outcome of interest (whether or not the transcript was produced by a healthy control). In the context of deep neural networks for image recognition, it has been proposed that the problem of confounding can be addressed by introducing confounding variables of interest as additional model outputs (Zhong and Ettinger, 2017). The authors of this work argue that including confounding variables as secondary prediction objectives will influence model weights via backpropagation, resulting in models with better generalizability and overall performance. This argument is supported by empirical results demonstrating improved performance on an image classification task when potential confounding variables indicating position and orientation are incorporated as secondary targets for prediction. Motivated by this argument, we evaluate the utility of treating the provenance of a transcript (ADReSS vs. WLS) as a secondary target for prediction on overall model performance, with the secondary objective of determining the extent to which deep neural networks can learn to distinguish between unseen transcripts from each of these corpora. This secondary objective is of interest because accurate classification of unseen transcripts would confirm that there is systematic difference between transcripts from each corpus that has the potential to bias machine learning models, despite this not being immediately apparent upon qualitative evaluation of randomly selected transcripts during the process of data preparation.

A second concern with combining transcripts in this manner is that it introduces a class imbalance, where transcripts from healthy “controls” greatly outnumber those from patients with dementia. Previous work with WLS data used oversampling of the minority class to address this imbalance, which was effective with some but not all models (Noorian et al., 2017). As recent work with BERT suggests cost-sensitive learning is an effective alternative to address class imbalance (Madabushi et al., 2019),

we evaluate the utility of this method also. Cost-sensitive learning involves adjusting the loss function of a model such that changes in performance on one class are weighted more heavily. In this case this involves proportionally weighting the loss function as an inverse function of the class distribution, such that the model learns to avoid misclassifying transcripts from dementia patients more assiduously than it learns not to misclassify those from healthy controls. Finally, we note that unlike the ADReSS set, the WLS transcripts do not come with diagnostic labels. However, the metadata accompanying these transcripts do include results of verbal fluency tests, as well as metadata indicative of clinical diagnoses other than dementia. A straightforward way to use these metadata involves developing an exclusion criterion, such that transcripts from participants with verbal fluency scores suggestive of diminished cognitive function are not treated as controls. In an additional effort to address the class imbalance introduced by the WLS data, we also experiment with treating the below-threshold fluency scores appended to these excluded transcripts as “noisy labels” (Natarajan et al., 2013) for the presence of dementia.

Thus, our research aims to answer the following key questions:

1. Does the performance of contemporary deep learning models on the ADReSS challenge diagnosis task benefit from the introduction of additional normative data comprising of “Cookie Theft” recordings from outside the ADReSS (or DementiaBank) set?
2. Does the addition of auxiliary outputs, or the incorporation of a cost-sensitive weight function, provide a way to compensate for the potential confounding effects and class imbalance introduced by these additional normative data, respectively?
3. Can verbal fluency scores be used to derive “noisy labels” to produce additional “case” training examples that are of value for performance on this task?
4. Are the two corpora sufficiently different that a deep learning model might learn to distinguish between them, during the course of the classification procedure?

Our main contributions can be summarized as follows:

1. We introduce a new professionally transcribed data set of 1,366 transcripts of the “Cookie Theft” task
2. We use associated metadata to infer noisy “case” and “control” labels for each transcript
3. We evaluate the utility of these additional data with and without inferred labels to improve the performance of transfer learning approaches on the ADReSS challenge classification task
4. We compare a set of loss function alternatives as a means to further improve performance.

3. MATERIALS AND METHODS

3.1. Dataset

3.1.1. ADReSS

The ADReSS dataset, derived from the DementiaBank dataset, consists of 156 speech transcriptions from AD and non-AD

patients which are matched for age and gender. Transcripts are English language responses to the “Cookie Theft” task of the Boston Diagnostic Aphasia Exam, and are classified as “AD” or “control” on the basis of clinical and/or pathological examination. We downloaded the ADReSS dataset from the *Alzheimer’s Dementia Recognition through Spontaneous Speech: The ADReSS Challenge* website¹.

3.1.2. Wisconsin Longitudinal Study

The Wisconsin Longitudinal Study (WLS) is a longitudinal study of a random sample of 10,317 graduates from Wisconsin high schools in 1957. The study also includes a randomly selected sibling of graduates, and spouses of graduates and siblings. WLS participants were interviewed up to six times across 60-years between 1957 and 2011. Beginning in 1993, during the fourth round of interviews, the WLS included cognitive evaluations. The “Cookie Theft” task was administered in the sixth-round of the survey in 2011 survey (see Herd et al., 2014 for details). In July of 2019 the ongoing seventh round of data collection began.

3.2. Experiments

3.2.1. Dataset Construction

All audio samples in the WLS dataset were transcribed near-verbatim by a professional service. The resulting near-verbatim transcripts include filled pauses (um’s and ah’s) and tags for unintelligible speech. The transcriptionists also separated the speech of the examiner (containing task instructions and task-final comments) from the participant’s response to the task. For the purposes of the current study, we removed filled pauses and unintelligible speech segments as well as the text corresponding to the examiner’s speech.

The metadata of WLS do not currently provide dementia-related diagnoses; however, they do provide a limited set of cognitive test scores, and answers to questions about some health conditions. Of relevance to the current research, WLS participants underwent two category verbal fluency cognitive tests in which they were asked to name all words that belonged to a category (animals, food) in 1 min. The semantic (category) verbal fluency task has been previously shown to be highly sensitive (albeit not specific) to manifestations of AD dementia (Henry et al., 2004) with an unadjusted for age and education cutoff of 15 on the animal category recommended for use as a screening instrument in a clinical setting (Duff-Canning et al., 2004).

In order to identify a subgroup of healthy controls in the WLS dataset comparable to controls in the ADReSS dataset we used the verbal fluency scores and an answer of “yes” to the question “Have you ever been diagnosed with mental illness?” as inclusion/exclusion criteria as follows. We classified transcripts of participants as cases (as opposed to healthy controls) if (1) the participants had evidence of impairment in semantic verbal fluency, or (2) have been diagnosed with a mental illness².

Prior work on verbal fluency performance in participants with AD established that animal fluency scores <15 are 20 times more likely in a patient with AD than in a healthy individual and were found to discriminate between these two groups with sensitivity of 0.88 and specificity of 0.96 (Duff-Canning et al., 2004). Recognizing the fact that verbal fluency performance does vary slightly by age and education (Tombaugh et al., 1999; Marceaux et al., 2019), we used statistically determined age and education-adjusted thresholds of 16, 14, and 12 for participants in <60, 60–79, and >79 age ranges, respectively. We did not have normative data available for the food category; however, since the distributions of semantic verbal fluency scores on the “animal” category and “food” category were very similar, we applied the same cutoffs for the food category as for the animal category.

The initial set of 1,366 WLS participants was reduced to 1,165 by removing those with extremely long and short transcripts whose length was beyond one standard deviation around the mean length of a WLS transcript. Of the remaining WLS participants with a “Cookie Theft” picture description task transcription, 954 participants also had a category semantic verbal fluency score or indicated a mental illness diagnosis. Of these participants, 839 had a verbal fluency score above the normative threshold and did not have a mental illness diagnosis. These were labeled as “controls.” Of the remaining 115 participants, 98 had a verbal fluency score below the threshold and 20 had a mental illness diagnosis. These 115 participants were labeled as “cases.”

Descriptive statistics for the ADReSS and WLS datasets are shown in **Table 1**. The mean ages of WLS controls and cases at the point of data collection are lower than those of participants whose transcripts make up the ADReSS dataset. Upon analysis of the differences in age of participants between the two corpora, we found that while there was no statistically significant difference [$t(1108) = 4.3, p = 1.96$] in the overall age of ADReSS ($M = 65.6, SD = 6.6$) and WLS participants ($M = 63.9, SD = 4.1$), nor in the age of controls [$t(915) = 1.3, p = 0.19$]³, there was a significant difference between the ages of AD cases in the ADReSS set and inferred WLS “cases” [$t(191) = 4.6, p < 0.001$]. While statistically significant, this difference in mean ages is relatively small (2.3 years) and may be of limited practical significance. Gender distributions among these two datasets are similar. In both the WLS and ADReSS sets, a larger proportion of the control group attained post-high-school education.

3.2.2. Model

Bidirectional Encoder Representations from Transformers (BERT Devlin et al., 2018) provides a pretrained deep neural network for researchers and practitioners to fine tune on specific tasks by adding just one additional output layer (Liu and Lapata, 2019). BERT exemplifies the “transfer learning” approach that has been used to improve performance across a range

is related to cognitive impairment. However, in the absence of specific metadata related to the presence of dementia, we decided it would be better to exclude these participants from the control set also.

³T-test results are reported in APA style: $t(\text{degrees of freedom}) = t$ statistic, $p = p$ -value. The abbreviations M and SD stand for mean and standard deviation, respectively.

¹<http://www.homepages.ed.ac.uk/sluzfil/ADReSS/>

²We use a generic term “cases” for participants with potential cognitive impairment and mental illness only as a way to distinguish them from controls, as we expect their language production on the picture description task to differ from that of controls. We do not in any way imply that a mental illness diagnosis

TABLE 1 | Dataset description.

n		ADReSS			WLS		
		Control 78	Case 78	P-value	Control 839	Case 115	P-value
Age, mean (SD)		65.0 (7)	66.3 (7)	0.226	63.9 (5)	63.9 (4)	0.902
Gender, n (%)	Female	43 (55)	43 (55)	1	295 (35)	32 (28)	0.995
	Male	35 (45)	35 (45)		213 (25)	23 (20)	
	Refused/Missing	0	0		331 (40)	60(52)	
Education, n (%)	≤12 years	34 (44)	52 (67)	0.002	401 (48)	78 (68)	<0.001
	>12 years	43 (55)	22 (28)		438 (52)	37 (32)	
	Refused/Missing	1 (1)	4 (5)		0	0	

Two-sample t-tests were used to evaluate the p-value for continuous variables, and Chi-squared was used for categorical variables.

of classification tasks in image and text processing in recent years. Essentially, transfer learning allows for the application of information learned while training a model on one task, to a different one. In the case of BERT for text classification, the initial task involves predicting held out (“masked”) words or sentences in a large corpus of otherwise unlabeled text. The general information about word distribution and relative position learned in this manner can then be applied to a downstream classification task, with or without fine-tuning the weights of BERT in addition to a classification layer that is appended to this pretrained deep neural network model. Unlike previous recurrent neural network approaches, BERT allows the model to process words in relation to all other words in a passage in parallel rather than sequentially, enhancing the scalability of the pre-training procedure. An important feature of BERT is its use of attention modules (Vaswani et al., 2017), which take into account other words in a unit of text when generating a word representation during pre-training and subsequent tasks. BERT can therefore take the broader context of a word into consideration, with the capacity to resolve ambiguities in contextual word meaning. Most importantly, the information acquired during the pre-training process enables BERT to perform well even when only small amounts of annotated data are available for fine tuning. Following previous work, we modified BERT by adding a classification layer, to obtain binary class labels corresponding to “cases” and “controls” in the ADReSS dataset.

3.2.3. Loss Functions

We evaluated the utility of several variants of the BERT loss function as a means to compensate for class imbalance, and potential confounding effects. The standard loss function for categorization with BERT (as implemented in the widely used Huggingface Transformers library⁴) is the `CrossEntropy` loss, which combines a softmax function with the standard Cross Entropy loss. This encourages a model to choose one of a set of possible classes in a text categorization class, by converting model outputs into a series of probabilities across classes, which sum to one across all classes, before calculating the loss. For multi-label

classification, where more than one label can be assigned (in our case, diagnosis = [case|control], source = [WLS|ADReSS]), a reasonable alternative is to use the `BCEwithLogits` (BCE) loss function, which does not require probabilities as inputs. As this loss function also provides a convenient means to weight classes, we retained it for our experiments with cost-weighting as a means to compensate for class imbalance by applying a weight of $\frac{n}{c}$ for each class, where n is the number of transcripts in the set, and c is the number of transcripts of the class of interest. Less frequent classes (the “dementia” class when WLS is used) will have more influence on the cost function, as they will have a smaller denominator. In order to isolate the effects of this loss function from the multilabel and weighted configurations of it, we also report results with an unweighted edition of the `BCEwithLogits` loss, as well as the standard loss function.

3.2.4. Methods and Evaluation

To evaluate the effect of adding more data, the WLS control and WLS total sets (case and control) were added to the ADReSS training set separately. We used the single unique ADReSS test set as the testing set for all models, and evaluated the models by accuracy and area under the receiver-operator curve (AUC). We also performed cross-validation (CV) on the training set.

We report evaluation metrics with 5-fold CV (rather than the leave-one-subject-out protocol used in some prior work) due to memory and time constraints. In this case, values of evaluation metrics were averaged across CV folds. To evaluate performance on the test set, we generated 10 instantiations of each model using different random seeds to determine the initialization of classifier weights for each instantiation. We trained each of these models on the training set (\pm the WLS components) and reported the mean and standard error across these ten runs. For two class label prediction, we evaluated models with the standard loss function, a weighted BCE loss function, and an unweighted BCE loss function. Finally, we evaluated a multi-label classification model (AD, not AD, ADReSS, WLS), using an unweighted BCE loss function.

3.2.5. Training Details

All experiments were conducted with the 12-layer `bert-base-uncased` model. Experiments using cross-validation on the training set were run on a single NVIDIA Tesla

⁴<https://github.com/huggingface/transformers>

P-40 GPU, while experiments with evaluation on the test set were run on a single NVIDIA Tesla V-100 GPU. All models were developed using Python 3.7 and PyTorch 1.2.0. We used the Transformers library to implement BERT in PyTorch (Wolf et al., 2019), permitting fine-tuning of BERT model weights in addition to tuning of the classification layer. The maximum sentence length was set to the maximum length of the current training set, and the batch size was set to 8. The learning rate was set to 1×10^{-5} . All models were run for 20 epochs. We adopted the Adam optimizer (Kingma and Ba, 2014) with linear scheduling (Paszke et al., 2019) of the learning rate. For the BCE loss function, nn.BCEWithLogitsLoss was used. Other hyper-parameters were set to their default values.

4. RESULTS

The results of our 5-fold cross-validation experiments are shown in Table 2. When interpreting this table it is important to bear in mind that the cross-validation splits in the WLS control and WLS total scenarios include examples from the respective WLS sets also. Thus, they are not comparable to one another, nor are they comparable to the results shown with the ADReSS set only. However, it is informative to compare the results within each panel in turn (aside from the ADReSS-only result, which provides an indication of the robustness of the results from the train-test split used in the challenge). It is also important to note that the standard error of the mean (indicated with \pm) is calculated across the five cross-validation folds, and consequently are indicative of the differences between the validation sets in these folds, rather than differences emerging from stochastic initialization of the classification layer of the BERT models concerned (these were initialized with the same random seed).

Both the WLS control and WLS total results suggest a trend toward an advantage for the loss function variants under consideration, as compared with the standard loss function, with unweighted and weighted variants of the BCE loss function generally outperforming the standard loss function. In addition, the best results in most cases are attained by the multilabel model. This suggests that augmentation of the model with additional targets for prediction may be helpful to reduce the confounding effect of the provenance of the transcripts concerned, when transcripts from both sources are included in the validation set. However, we note that one exception to this finding is the AUC in the WLS total set—in this configuration, the standard loss function performs best. The relatively poor performance with the addition of the “WLS total” set in 5-fold CV may result from discrepancies between the noisily labeled WLS cases and the clinically determined ADReSS AD dementia cases.

Results on the held-out ADReSS challenge test set are shown in Table 3, with the model trained on the ADReSS training set only and using the standard loss function taken as a baseline (these baseline results are largely consistent with the 5-fold cross-validation results on this set, suggesting the test set is representative of the data set as a whole). When comparing results from the three models trained with a standard loss function to evaluate the impact of the WLS data on a standard

TABLE 2 | Five-fold cross-validation results on training set.

Data	Loss function	% Accuracy	% AUC
ADReSS	Standard	80.5 \pm 4.0	88.2 \pm 3.1
ADReSS + WLS control	Standard	96.5 \pm 0.3	98.7 \pm 0.3
	Weighted BCE	97.4 \pm 0.3	98.9 \pm 0.2
	Unweighted BCE	97.4 \pm 0.3	98.8 \pm 0.4
	Multilabel BCE	97.9 \pm 0.5	99.2 \pm 0.1
ADReSS + WLS total	Standard	83.3 \pm 1.2	68.8 \pm 0.6
	Weighted BCE	83.7 \pm 1.4	61.2 \pm 2.8
	Unweighted BCE	83.7 \pm 1.4	65.7 \pm 2.8
	Multilabel BCE	84.8 \pm 1.3	66.1 \pm 1.6

Results shown are the mean across the 5-folds \pm the standard error. Best results in panels showing multiple models are in boldface.

TABLE 3 | Results on ADReSS test set.

Data	Loss function	% Accuracy	% AUC
ADReSS	Standard	79.8 \pm 0.9	88.3 \pm 0.5
ADReSS + WLS control	Standard	81.2 \pm 1.1	90.6 \pm 0.9
	Weighted BCE	82.1 \pm 1.0	92.3 \pm 0.4*
	Unweighted BCE	80.8 \pm 1.1	91.6 \pm 0.3*
	Multilabel BCE	81.2 \pm 0.5	90.6 \pm 0.5*
ADReSS + WLS total	Standard	81.9 \pm 1.1	91.2 \pm 0.9*
	Weighted BCE	80.8 \pm 0.6	89.3 \pm 0.9
	Unweighted BCE	80.8 \pm 1.1	88.9 \pm 0.5
	Multilabel BCE	80.4 \pm 0.9	91.2 \pm 0.4*

Results shown are the mean across ten iterations \pm the standard error. *Indicates statistically significant difference from the baseline, as estimated by a paired t-test (as each repeated train/test evaluation was initialized with the same random seed across models). Best results in panels showing multiple models are in boldface.

BERT classifier, we find both incorporating additional WLS controls, and the WLS total data (with controls and noisy labels for cases) leads to improvements over the baseline model. On account of the small number of test cases, only the advantages in AUC are statistically significant—presumably on account of accuracy generally having higher variance across runs than the AUC (as a small change in the predicted probability of an example may lead to a larger change in accuracy if this crosses the classification boundary and leads to error). Nonetheless, the general trend supports the hypothesis that the additional normative data will improve the performance of BERT on the ADReSS challenge diagnosis task.

When comparing the loss function variants, we observe that those models trained on the ADReSS set with the addition of WLS controls only using the weighted BCE function achieves the best AUC and accuracy amongst all the models, suggesting that weighting the loss function is an effective way of compensating for the class imbalance that results from these additional “control” data points. More importantly, this model significantly improves AUC compared to the baseline model in the test set. Unlike the 5-fold CV scenario, the multi-label loss function does

not lead to better performance than the standard loss function—which is perhaps unsurprising given the total absence of WLS data in the test set, obviating the need to resolve confounding effects emerging from data provenance. That the unweighted BCE function also does not improve performance over the standard function supports the hypotheses that it is indeed the weighting of this function that is responsible for its advantages in performance.

An additional finding from these experiments is that the multilabel models correctly identified the provenance of the ADReSS-derived examples in the test set with perfect accuracy in nine of 10 runs, and ~98% accuracy on the remaining run. These results strongly support the hypothesis that a deep learning model trained on data from both corpora would learn to distinguish between them. This finding is further supported by a perfect accuracy in distinguishing between these corpora in the held-out validation split (including both WLS- and ADReSS-derived examples) demonstrated in a subsequent run.

The results with the inclusion of “noisy” WLS cases differ from those with controls alone. With the standard loss function, the addition of these data improves performance beyond that attained by adding WLS controls alone. However, performance does not match the best of the “control only” models, and is not improved further with the addition of variant loss functions. One explanation for the latter finding may be that class imbalance effects are already obviated through the introduction of additional cases, increasing the positively labeled training examples from ~5 to ~16% of the data available for training.

5. DISCUSSION

In this paper, we evaluated the utility of the incorporation of additional “Cookie Theft” transcripts drawn from the Wisconsin Longitudinal Study as a means to improve the performance of a BERT-based classifier on the ADReSS challenge diagnosis task. Our aims in doing so were primarily to evaluate whether or not these data would improve performance, but also to establish the extent to which weighting the cost function of the model and representing corpus provenance as additional targets for prediction could compensate for the issues of class imbalance and corpus-specific confounding effects, respectively. Finally, we wished to determine whether or not a model could learn to distinguish between the two corpora, to determine if our concern about such corpus-specific confounding effects was justified.

We found that incorporation of WLS data improved performance over that of a model trained on ADReSS data alone, and that these improvements were present both when only WLS “controls” (transcripts from participants with verbal fluency scores in the normal range for their age) were added, and when these were combined with noisily-labeled WLS “cases” (transcripts from participants with low verbal fluency scores, or reported diagnoses related to mental illness). When only controls were added, further improvements in performance were obtained when weighting the cost function to compensate for class imbalance, resulting in the best-performing models on the ADReSS challenge test set, with a mean accuracy of 82.1% and

mean AUC of 92.3% across ten repeated instantiations of the model, as opposed to a mean accuracy and AUC of 79.8 and 88.3%, respectively, without the addition of WLS data.

While we note that Balagopalan et al. (2020) report an accuracy of 83.3% with a BERT-based classifier on this task when trained on the AD set alone, these experiments did not include repeated model instantiations to determine the effects of stochastic initialization of classifier weights on performance, that our baseline AD-only BERT model attained an accuracy of 83.3 or higher on two of ten such iterations, and that the best performance of the cost-weighted model across iterations resulted in an accuracy of 89.6% (with an AUC of 94.8%). This difference in baseline performance may be attributable to differences in stochastic initialization, or an unspecified difference in model architecture (e.g., BERT-base vs. BERT-large) or hyperparameter settings, and we do not believe it detracts from the strength of our conclusions.

While most of the work with the ADReSS challenge data has focused on multimodal analyses of acoustic and transcript data simultaneously, the paper introducing this data set provides some baseline results with language-only models, which were trained on a set of thirty-four linguistic outcome measures (such as total number of utterances, and part-of-speech percentage) (Luz et al., 2020). Test set classification accuracy is generally lower than results attained using BERT trained on raw text (even without the addition of WLS data), ranging from 0.625 to 0.792 across algorithms. Best performance was attained using a Support Vector Machine, though this configuration performed worse than other algorithms in cross-validation experiments. This suggests that BERT is able to automatically extract predictive features that outperform handcrafted features. However, it should be noted that BERT has considerably more trainable parameters than the models evaluated in this prior work, and that a fair comparison between BERT-based and engineered features would require the ascertainment of BERT’s performance with freezing of all layers aside from the classification layer. Other work focusing on linguistic features explores the utility of using terms as features directly. Searle et al. (2020) compared machine learning models applied to word-level features to the DistilBERT architecture, reporting tied best accuracies of 81% with DistilBERT and an utterance-level combination of a Support Vector Machine and Conditional Random Field classifier. Additional work suggests that incorporation of acoustic features may offer further advantages in performance. Syed et al. (2020) demonstrated an accuracy of 85.4% with a multimodal learning system that incorporated both audio signals and transcripts. BERT and RoBERTa were included in the multimodal framework. These results suggest that incorporating additional information from auditory features may suggest a path toward further improving the performance of our models, although there are technical challenges concerning the differences in recording instruments across data sets that would need to be addressed in order to explore this.

In the context of 5-fold cross-validation experiments, where both WLS- and ADReSS-derived examples were present in the validation splits, adding transcript provenance as an additional target for prediction in a multi-label setting resulted in best

performance, supporting the hypotheses that this may be an effective way to address corpus-specific confounding effects, which are an important concern in biomedical machine learning when there is a need to assemble a data set from smaller constituents that may have been collected at different institutions. Of particular interest for future work, these models also learned to classify the provenance of the data sets concerned with perfect or near-perfect accuracy, suggesting systematic differences between the source corpora that were not apparent upon informal inspection of word usage and lexical patterns. Further research is required to determine the cues used by the models to make these distinctions.

The results presented in this paper should be interpreted in light of several limitations. First, the ADReSS dataset is relatively small. The results reported here need to be replicated on larger datasets to determine their generalizability. Second, while the WLS dataset contains a very rich set of participant characteristics, these characteristics do not include those that can be used directly for characterization of AD dementia. Thus, the results pertaining to WLS cases should be interpreted with caution. In particular, while utilization of mental health diagnoses to exclude transcripts from the analysis is readily justifiable, using these diagnoses to derive noisy labels for cases may exceed the bounds of noise that our models can tolerate. In future work we will evaluate the extent to which using verbal fluency derived criterion only leads to noisy labels with greater downstream utility. We note also that efforts are currently underway to interview WLS participants in order to obtain clinical diagnoses of dementia. We anticipate this measure will be available for all eligible participants within a few years of the time of this writing, which will further enhance the utility of our transcripts for future research on the linguistic manifestations of dementia. Third, cases in both datasets are significantly less educated than the controls which may result in language use artifacts that have not been accounted for. These potential differences in language use should be further investigated. There are also some methodological alternatives that we did not fully consider in the current work. Our study did not consider the acoustic components of the available data, and depended upon manual transcriptions of speech data. Further research is needed to determine the utility of incorporating acoustic features, as well as the model's robustness to errors that may be introduced during the process of automated speech recognition. Furthermore, we did not formally evaluate oversampling strategies. Preliminary experiments with random oversampling suggested this would not be a fruitful strategy, and to our knowledge BERT-based strategies for similarity-based oversampling have yet to be developed. In addition we have yet to evaluate the combination of auxiliary prediction targets with weighting of the cost function, which may be a productive direction to pursue in future work on account of their individual utility when transcripts from both corpora are present at the point of validation. Finally, the utility of auxiliary targets as a means to obviate for confounding effects may be more readily apparent when the distribution of positive cases across corpora is different at test time than at training time (Landeiro and Culotta, 2018). Establishing whether or not this is the case would require additional evaluation involving

validation sets in which these distributions are artificially modified.

6. CONCLUSION

In this paper, we evaluated the utility of using additional “Cookie Theft” picture description transcripts from the Wisconsin Longitudinal Study, as a means to improve the performance of a BERT-based classification approach on the dementia detection task of the ADReSS challenge. Our results indicate that training on these additional data leads to improved performance on this task, both when using all available transcripts as normative data regardless of cognitive status and subsets of the data extracted based on cognitive status inferred from available metadata (i.e., verbal fluency and mental health status). In the former case in particular, we find that weighted cost functions are an effective way to compensate for the class imbalance introduced by the addition of more “control” transcripts. Furthermore, results from our cross-validation studies suggest that introducing dataset provenance as an auxiliary target for prediction shows potential as a means to address different case/control distributions when combining datasets drawn from different sources. As such, our results suggest that our professionally transcribed WLS “Cookie Theft” transcripts are a valuable resource for the development of models to detect linguistic anomalies in dementia. These transcripts are available upon request from wls@ssc.wisc.edu.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found at: Alzheimer's Dementia Recognition through Spontaneous Speech (ADReSS), <http://www.homepages.ed.ac.uk/sluzfil/ADReSS/>. Requests to access our professional transcriptions of the Wisconsin Longitudinal Study (WLS) data should be directed to wls@ssc.wisc.edu.

AUTHOR CONTRIBUTIONS

YG, CL, TC, and SP conceived of the study design. YG and TC performed the experiments, with analysis of the results conducted in collaboration with authors SP and CL. CR provided the access to the WLS recordings and metadata, as well as guidance in the analysis and interpretation of these metadata. YG wrote the initial draft of the manuscript, with input from TC. All authors read and reviewed the manuscript, providing edits and suggestions for improvement where appropriate.

FUNDING

This research was supported by Administrative Supplement R01 LM011563 S1 from the National Library of Medicine, and R21 AG069792 from the National Institute of Aging. Since 1991, the WLS has been supported principally by the National Institute on Aging (AG-9775, AG-21079, and AG-033285), with additional support from the Vilas Estate Trust, the National Science Foundation, the Spencer Foundation, and the Graduate School of the University of Wisconsin-Madison.

REFERENCES

- Aguirre-Acevedo, D. C., Lopera, F., Henao, E., Tirado, V., Muñoz, C., Giraldo, M., et al. (2016). Cognitive decline in a colombian kindred with autosomal dominant alzheimer disease: a retrospective cohort study. *JAMA Neurol.* 73, 431–438. doi: 10.1001/jamaneurol.2015.4851
- Association, A. (2018). 2018 Alzheimer's disease facts and figures. *Alzheimers Dement.* 14, 367–429. doi: 10.1016/j.jalz.2018.02.001
- Association, A. (2019). 2019 Alzheimer's disease facts and figures. *Alzheimers Dement.* 15, 321–387. doi: 10.1016/j.jalz.2019.01.010
- Balagopalan, A., Eyre, B., Rudzicz, F., and Novikova, J. (2020). To bert or not to bert: comparing speech and language-based approaches for Alzheimer's disease detection. *arXiv [Preprint] arXiv:2008.01551*. doi: 10.21437/Interspeech.2020-2557
- Becker, J. T., Boiler, F., Lopez, O. L., Saxton, J., and McGonigle, K. L. (1994). The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis. *Archiv. Neurol.* 51, 585–594. doi: 10.1001/archneur.1994.00540180063015
- Blendon, R., Benson, J., Wikler, E., Weldon, K., Baumgart, M., Jansen, S., et al. (2011). Five-country survey of public experiences, attitudes and beliefs concerning Alzheimer's disease and the value of a diagnosis. *Alzheimers Dement.* 7:e50. doi: 10.1016/j.jalz.2011.09.209
- Boise, L., Camicioli, R., Morgan, D. L., Rose, J. H., and Congleton, L. (1999). Diagnosing dementia: perspectives of primary care physicians. *Gerontologist* 39, 457–464. doi: 10.1093/geront/39.4.457
- Bond, J., Stave, C., Sganga, A., Vincenzino, O., O'connell, B., and Stanley, R. (2005). Inequalities in dementia care across europe: key findings of the facing dementia survey. *Int. J. Clin. Pract.* 59, 8–14. doi: 10.1111/j.1368-504X.2005.00480.x
- Bradford, A., Kunik, M. E., Schulz, P., Williams, S. P., and Singh, H. (2009). Missed and delayed diagnosis of dementia in primary care: prevalence and contributing factors. *Alzheimer Dis. Assoc. Disord.* 23:306. doi: 10.1097/WAD.0b013e3181a6bbebc
- Cohen, T., and Pakhomov, S. (2020). "A tale of two Q15 perplexities: sensitivity of neural language models to lexical retrieval deficits in dementia of the Alzheimer's type," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (Association for Computational Linguistics)*, 1946–1957. doi: 10.18653/v1/2020.acl-main.176
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: pre-training of deep bidirectional transformers for language understanding. *arXiv [Preprint] arXiv:1712.00069*.
- Duff-Canning, S., Leach, L., Stuss, D., Ngo, L., and Black, S. (2004). Diagnostic utility of abbreviated fluency measures in Alzheimer disease and vascular dementia. *Neurology* 62, 556–562. doi: 10.1212/WNL.62.4.556
- Fraser, K. C., Meltzer, J. A., and Rudzicz, F. (2016). Linguistic features identify Alzheimer's disease in narrative speech. *J. Alzheimers Dis.* 49, 407–422. doi: 10.3233/JAD-150520
- He, H., Bai, Y., Garcia, E. A., and Li, S. (2008). "Adasyn: Adaptive synthetic sampling approach for imbalanced learning," in *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)* (Hong Kong: IEEE), 1322–1328.
- Henry, J. D., Crawford, J. R., and Phillips, L. H. (2004). Verbal fluency performance in dementia of the Alzheimer's type: a meta-analysis. *Neuropsychologia* 42, 1212–1222. doi: 10.1016/j.neuropsychologia.2004.02.001
- Herd, P., Carr, D., and Roan, C. (2014). Cohort Profile: Wisconsin longitudinal study (WLS). *Int. J. Epidemiol.* 43, 34–41. doi: 10.1093/ije/dys194
- Howell, K., Barnes, M., Curtis, J. R., Engelberg, R. A., Lee, R. Y., Lober, W. B., et al. (2020). "Controlling for confounding variables: accounting for dataset bias in classifying patient-provider interactions," in *Explainable AI in Healthcare and Medicine*, eds A. Shaban-Nejad, M. Michalowski, and D. L. Buckeridge (New York, NY: Springer), 271–282. doi: 10.1007/978-3-030-53352-6_25
- Karlekar, S., Niu, T., and Bansal, M. (2018). Detecting linguistic characteristics of Alzheimer's dementia by interpreting neural models. *arXiv [Preprint] arXiv:2006.07358*. doi: 10.18653/v1/N18-2110
- Kingma, D. P., and Ba, J. (2014). Adam: a method for stochastic optimization. *arXiv [Preprint] arXiv:1412.6980*.
- Landeiro, V., and Culotta, A. (2018). Robust text classification under confounding shift. *J. Artif. Intell. Res.* 63, 391–419. doi: 10.1613/jair.1.11248
- Liu, Y., and Lapata, M. (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)* (Hong Kong), 3721–3731. doi: 10.18653/v1/D19-1387
- Luz, S., Haider, F., de la Fuente, S., Fromm, D., and MacWhinney, B. (2020). Alzheimer's dementia recognition through spontaneous speech: the address challenge. *arXiv [Preprint] arXiv:2004.06833*. doi: 10.21437/Interspeech.2020-2571
- Madabushi, H. T., Kochkina, E., and Castelle, M. (2019). "Cost-sensitive bert for generalisable sentence classification with imbalanced data," in *EMNLP-IJCNLP 2019* (Hong Kong), 125. doi: 10.18653/v1/D19-5018
- Marceaux, J. C., Prose, M. A., McClure, L. A., Kana, B., Crowe, M., Kissela, B., et al. (2019). Verbal fluency in a national sample: telephone administration methods. *Int. J. Geriatr. Psychiatry* 34, 578–587. doi: 10.1002/gps.5054
- Natarajan, N., Dhillon, I. S., Ravikumar, P. K., and Tewari, A. (2013). *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013*, eds C. J. C. Burges, L. Bottou, Z. Ghahramani, and K. Q. Weinberger.
- Noorian, Z., Pou-Prom, C., and Rudzicz, F. (2017). On the importance of normative data in speech-based assessment. *arXiv [Preprint] arXiv:1712.00069*.
- Orimaye, S. O., Wong, J. S., Golden, K. J., Wong, C. P., and Soyiri, I. N. (2017). Predicting probable Alzheimer's disease using linguistic deficits and biomarkers. *BMC Bioinformatics*, (Baltimore, MD) 18:34. doi: 10.1186/s12859-016-1456-0
- Orimaye, S. O., Wong, J. S.-M., and Golden, K. J. (2014). "Learning predictive linguistic features for Alzheimer's disease and related dementias using verbal utterances," in *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 78–87. doi: 10.3115/v1/W14-3210
- Orimaye, S. O., Wong, J. S.-M., and Wong, C. P. (2018). Deep language space neural network for classifying mild cognitive impairment and Alzheimer-type dementia. *PLoS ONE* 13:e0205636. doi: 10.1371/journal.pone.0205636
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. (2019). *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019*, eds H. M. Wallach, H. Larochelle, A. Beygelzimer, Florence d'Alché-Buc and E. B. Fox, and R. Garnett (Vancouver, BC).
- Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., et al. (2011). "The Kaldi speech recognition toolkit" in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, Number CONF* (Barcelona: IEEE Signal Processing Society).
- Rajan, K. B., Wilson, R. S., Weuve, J., Barnes, L. L., and Evans, D. A. (2015). Cognitive impairment 18 years before clinical diagnosis of Alzheimer disease dementia. *Neurology* 85, 898–904. doi: 10.1212/WNL.0000000000001774
- Searle, T., Ibrahim, Z., and Dobson, R. (2020). Comparing natural language processing techniques for Alzheimer's dementia prediction in spontaneous speech. *arXiv [Preprint] arXiv:2006.07358*. doi: 10.21437/Interspeech.2020-2729
- Stokes, L., Combes, H., and Stokes, G. (2015). The dementia diagnosis: a literature review of information, understanding, and attributions. *Psychogeriatrics* 15, 218–225. doi: 10.1111/psyg.12095
- Syed, M. S. S., Syed, Z. S., Lech, M., and Pirogova, E. (2020). "Automated screening for Alzheimer's dementia through spontaneous speech," in *INTERSPEECH*, 1–5. doi: 10.21437/Interspeech.2020-3158
- Tombaugh, T. N., Kozak, J., and Rees, L. (1999). Normative data stratified by age and education for two measures of verbal fluency: Fas and animal naming. *Archiv. Clin. Neuropsychol.* 14, 167–177. doi: 10.1016/S0887-6177(97)00095-4
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, eds I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett (Long Beach, CA).

- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., et al. (2019). Huggingface's transformers: state-of-the-art natural language processing. *arXiv abs/1910.03771*. doi: 10.18653/v1/2020.emnlp-demos.6
- Yancheva, M., and Rudzicz, F. (2016). "Vector-space topic models for detecting Alzheimer's disease" in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Berlin), Vol. 1, 2337–2346. doi: 10.18653/v1/P16-1221
- Zhong, Y., and Ettinger, G. (2017). "Enlightening deep neural networks with knowledge of confounding factors," in *Proceedings of the IEEE International Conference on Computer Vision Workshops (Venice)*, 1077–1086. doi: 10.1109/ICCVW.2017.131

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2021 Guo, Li, Roan, Pakhomov and Cohen. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.