Training Models on Oversampled Data and a Novel Multi-class Annotation Scheme for Dementia Detection

Nadine Abdelhalim, Ingy Abdelhalim and **Riza Batista-Navarro**
The University of Manchester, UK
nadine.abdelhalim@student.manchester.ac.uk
ingy.abdelhalim@student.manchester.ac.uk
riza.batista@manchester.ac.uk

Abstract

This work introduces a novel three-class annotation scheme for text-based dementia classification in patients, based on their recorded visit interactions. Multiple models were developed utilising BERT, RoBERTa and DistilBERT. Two approaches were employed to improve the representation of dementia samples: oversampling the underrepresented data points in the original Pitt dataset and combining the Pitt with the Holland and Kempler datasets. The DistilBERT models trained on either an oversampled Pitt dataset or the combined dataset performed best in classifying the dementia class. Specifically, the model trained on the oversampled Pitt dataset and the one trained on the combined dataset obtained state-of-the-art performance with 98.8% overall accuracy and 98.6% macro-averaged F1-score, respectively. The models' outputs were manually inspected through saliency highlighting, using Local Interpretable Model-agnostic Explanations (LIME), to provide a better understanding of its predictions.

1 Introduction

Dementia is a condition characterised by impaired memory, thinking or decision-making ability that interferes with daily activities (Gale et al., 2018). This global issue affects approximately 50 million individuals, with projections suggesting that the number will increase to 139 million by 2050 (World Health Organization, 2021). While no known cure for dementia currently exists, early diagnosis is essential, as it enables patients to access interventions that can help manage symptoms, prevent further degeneration and improve their quality of life.

Recent research suggests that language changes and a decline in episodic memory may serve as an essential signal for early diagnosis of dementia, with language impairments reported in both preclinical dementia and severe cases (Mueller et al., 2018; Yuan et al., 2020).

Methods for natural language processing (NLP) can help in detecting dementia through the analysis of the language used by a patient of interest. Indeed, previous research cast dementia detection as a binary text classification task, categorising a patient as exhibiting dementia or not, based on their language use (Roshanzamir et al., 2021; Matošević and Jović, 2022; Wahlforss and Jonasson, 2020; Orimaye et al., 2014; Yuan et al., 2020). However, thus far, no studies have investigated the classification of patient conversation transcripts into more than two classes. Our study aims to address this gap and seeks to analyse patients according to three classes: *Healthy Control (HC)*, *Early Stage or Mild Cognitive Impairment (MCI)* and *Dementia*. The goal is to provide medical professionals with a tool (that can be used in conjunction with standardised tests) for identifying patients exhibiting early-stage dementia symptoms. Such a tool can be useful in organisations where there is a lack of expertise among personnel responsible for screening patients, for the purposes of identifying those who could benefit from interventions that might potentially slow the progression of the disease.

Our approach involves analysing speech transcripts from doctor-patient conversations, with participants categorised into the three aforementioned classes. This task is a multi-class classification problem, which we address by developing models that are capable of classifying text (i.e., the transcripts) according to three classes. In particular, we developed models based on the transformer architecture (Vaswani et al., 2017), considering that transformers have demonstrated state-of-the-art performance in many clinical text classification tasks (Yogarajan et al., 2021). Additionally, we utilised explainability techniques to identify words that are indicative of dementia and may be used as features in the diagnostic process.

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 118–124

Model	Validation	Accuracy	F1	Reference
RoBERTa	Stratified 10-fold CV	90.60%	90.28%	Matošević and Jović (2022)
ERNIE+3Pause	LOO CV	89.6%	88.9%	Yuan et al. (2020)
BERT Large	10-fold CV	88.08%	87.23%	Roshanzamir et al. (2021)
DistilBERT+LR	Grid search and CV	88%	87%	Liu et al. (2022)
RoBERTa	10-fold CV	86.75%	86.82	Wahlforss and Jonasson (2020)

Table 1: Recent work on dementia detection using the Pitt corpus, excluding some models with slightly weaker performance. ERNIE+3Pause, which also uses audio, is based on the ERNIE 2.0 transformer architecture (Sun et al., 2020) with three types of pauses. Key: LR = logistic regression, CV = cross validation, LOO = leave one out.

2 Related Work

Recent work on dementia detection has been underpinned by text classification models based on transformer architectures. Table 1 highlights the most relevant and recent models developed using the Pitt Corpus from DementiaBank (Becker et al., 1994). The work by Matošević and Jović (2022), which was based on a RoBERTa model, has thus far achieved the state-of-the-art binary classification accuracy of 90.60%. Our own work similarly employed transformer-based models, i.e., BERT, RoBERTa and DistilBERT, while investigating the conversion of binary classification into a multi-class classification task for dementia severity. It is important to note that no previous work has been conducted on multi-class classification for dementia using text; thus, the performance of such models was previously unknown.

3 Methodology

This study employed two distinct approaches to developing dementia classification models. The first approach aimed to ensure comparability with previous research by solely utilising the Pitt dataset. However, the original Pitt dataset was highly imbalanced (with 259 HC, 127 MCI and 24 Dementia samples in the whole dataset), containing a limited number of confirmed dementia cases, necessitating oversampling to address this limitation. Specifically, we oversampled the MCI and Dementia classes to allow for a more balanced representation of these classes in the training set. Utilising stratified 10-fold cross-validation (CV) in our experiments, the resulting training dataset for each fold included original HC samples, MCI samples duplicated thrice, and Dementia samples duplicated 16 times. On the other hand, the test set (for each fold) was left unaltered.

The second approach involved combining the Pitt, Holland, and Kempler datasets to increase the representation of naturally occurring dementia in the dataset, thus eliminating the need for oversampling. This approach enabled us to assess the performance of the models with unique dementia data samples and a wider range of discussion topics. Table 4 in Appendix B presents the number of samples in the datasets that we have utilised in our experiments.

3.1 Data Pre-processing

The dataset was originally in the CHAT transcription format (MacWhinney, 2009), requiring conversion to plain text and subsequent pre-processing to eliminate extraneous punctuation and retain only participants' speech. The transcripts not only capture participants' spoken words but also provide additional information about their actions. The participants' actions were represented by symbols such as `&=coughs` for coughing or `&=clear` for clearing their throat. Pauses in the speech were indicated by bracketed full stops at the beginning of a sentence, with the number of full stops indicating the length of the pause. While most of the participants' actions and unnecessary punctuation were removed during pre-processing, pauses were retained due to their potential diagnostic value, as they are considered to be an important linguistic indication of cognitive decline in dementia patients (Sluis et al., 2020).

Following the pre-processing of the transcripts, each transcript was mapped to its corresponding Diagnostic ID by utilising its corresponding participant's ID. Based on these Diagnostic IDs, the transcripts were classified into three categories: *Healthy Control (HC)*, *Early Stage or Mild Cognitive Impairment (MCI)*, and *Dementia*. These labels were one-hot encoded: *[1,0,0] for HC, [0,1,0] for MCI and [0,0,1] for Dementia*. Transcripts with a Diagnostic ID corresponding to probable or possible dementia were excluded from the dataset. The resulting dataset was saved in a comma-separated values (CSV) file for ease of use in our experiments.

3.2 Model Training

We developed six bidirectional transformer-based models, specifically, the base variants of BERT, RoBERTa and DistilBERT: BERT-base (Devlin et al., 2018), RoBERTa-base (Liu et al., 2019) and DistilBERT-base (Sanh et al., 2019). The architectures of all six multi-class models were nearly identical, with the dataset, pre-trained layer and tokeniser being the primary distinguishing factors. Figure 1 provides an illustration of the architecture for the DistilBERT model. Additionally, a binary classification model was developed using RoBERTa to replicate results reported by Matošević and Jović (2022), using the same hyper-parameters described in their paper.

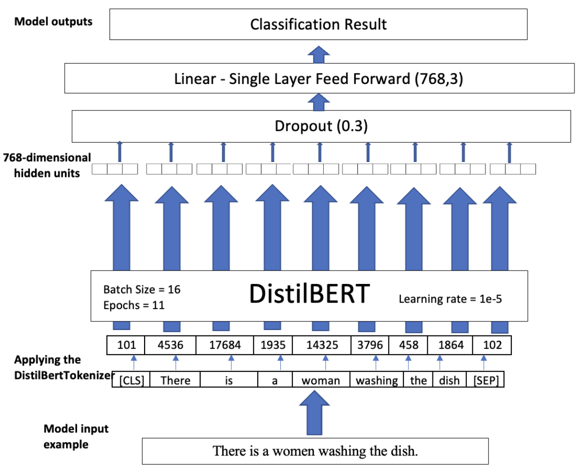

Figure 1: Model architecture. Image adapted from Liu et al. (2022).

3.3 Hyper-parameter optimisation

In order to optimise the performance of the models, hyper-parameter optimisation was performed for each pre-trained model type (BERT-base, RoBERTa-base and DistilBERT-base) and each dataset (Pitt, and the combined Pitt, Kempler and Holland dataset). Specifically, we explored different epochs ranging from 1 to 15 and different learning rates: 5e-5, 4e-5, 3e-5, 2e-5 and 1e-5. The optimal number of epochs varied for each model, but all models had an optimal learning rate of 1e-5. Stratified 10-fold CV was conducted to evaluate the average performance of each model.

3.4 Explainability

Explainability is crucial for NLP models, especially those that are intended for use in healthcare. By providing insight into a model's decision-making process, explanations can enhance the trust and confidence placed in the model's outputs. Furthermore, it can help to identify any potential biases or errors. To this end, we investigated the use of Local Interpretable Model-Agnostic Explanations (LIME) to explain the outputs of each of our models (Ribeiro et al., 2016).

4 Evaluation and Results

The objective of the experiments conducted was to test two fundamental hypotheses. Firstly, it was hypothesised that utilising the novel three-class labelling system would improve classification performance by enabling a more refined classification that can distinguish between more nuanced differences in the data. Secondly, it was hypothesised that models developed utilising the combination of datasets would exhibit superior performance to those developed using solely the Pitt dataset. The rationale behind this hypothesis was that the combined dataset would provide a more diverse and representative range of data, ultimately improving the generalisability of the models.

As described above, to test these hypotheses, three models were created using BERT, RoBERTa, and DistiBERT for each approach. The performance of the models was then evaluated using stratified 10-fold CV, with the performance metrics being accuracy, micro- and macro-averaged F1 scores, and, importantly, precision for the Dementia class. The lattermost metric is crucial in a medical diagnosis scenario: false positives for the Dementia class should be minimised as they could lead to unnecessary interventions or distress.

Table 2 presents a summary of the performance of the developed models. In terms of accuracy, the best performing model is the three-class DistilBERT model utilising the Pitt dataset. Meanwhile, the model that obtained the highest macro-averaged F1 score is the three-class DistilBERT model trained on the combined Pitt, Holland, and Kempler datasets. Appendix A includes an example of saliency highlighting performed by the LIME model.

Model	Dataset	Epochs	Accuracy	Macro F1	Precision
Binary (baseline) - RoBERTa	Pitt	-	90.3%	89.0%	-
3-class - BERT	Pitt - O	11	95.4%	93.0%	100%
3-class - RoBERTa	Pitt - O	11	96.5%	97.6%	100%
3-class - DistilBERT	Pitt - O	11	**98.8%**	97.6%	100%
3-class - BERT	Combined P+H+K	8	92.7%	91.4%	96.0%
3-class - RoBERTa	Combined P+H+K	30	94.4%	97.5%	100%
3-class - DistilBERT	Combined P+H+K	11	98.5%	**98.6%**	100%

Table 2: Performance of all models on the oversampled Pitt dataset (Pitt - O) and the combined Pitt, Holland and Kempler (P+H+K) dataset based on stratified 10-fold cross-validation. Precision is reported only for the Dementia class. The metric values for the baseline model were reproduced from the original paper by Matošević and Jović (2022). All models had an optimal batch size of 16.

5 Discussion

The three-class annotation scheme improves classification performance. As can be seen in Tables 1 and 2, using a three-class labelling system improved the performance of all models. The improved performance is likely due to the finer-grained system allowing for a more nuanced classification, distinguishing between cognitive impairment levels.

The results demonstrate that almost all three-class models achieved an average precision of 100% for the Dementia class. The best models were able to correctly identify positive cases without generating any false positives, making them valuable in medical diagnosis. In order to provide a more comprehensive evaluation of the class-level performance of our top-performing model in terms of F1-Score, the DistilBERT model trained on the combined dataset, a detailed breakdown of its performance table is presented in Table 3. It shows that for every class, the model performs well in terms of both precision and recall.

Oversampling is a viable method to improving a dementia detection model's accuracy and macro-averaged F1-score. As can be seen in Table 2, the DistilBERT model trained on the oversampled Pitt dataset obtained the highest accuracy of all the models created. This is very promising for any future work where combining multiple datasets or having a larger dataset is not an option.

The addition of a small number of dementia samples from outside the Pitt dataset significantly improves macro-averaged F1-score and accuracy. The best-performing model, in terms of macro-averaged F1-score, is the DistilBERT model generated using the combined dataset; this shows that a model using the three-class labelling system can exhibit optimal performance simply with the addition of a small number of dementia samples.

Although the work by Matošević and Jović (2022) did not provide any detailed performance breakdown for each class that would facilitate straightforward comparisons, the observed improvement in the overall performance of our DistilBERT model can be presumed to extend to the model's class-level performance.

Class	Precision	Recall	F1-Score
Healthy Control : [1,0,0]	0.97	1.00	0.99
Mild Cognitive Impairment : [0,1,0]	1.00	0.97	0.98
Dementia : [0,0,1]	1.00	0.91	0.95

Table 3: Performance of the 3-class DistilBERT model trained on the combined dataset.

6 Conclusion

This study proposes a novel three-class labelling system for classifying dementia in patients based on conversation transcripts. The proposed labelling system includes three classes: Healthy Control (HC), Early Stage or Mild Cognitive Impairment (MCI) and Dementia. Multiple models were developed utilising BERT, RoBERTa, and DistilBERT. To improve the representation of dementia data, we experimented with oversampling the Pitt dataset as well as combining the Pitt dataset with the Holland and Kempler datasets to increase the number of dementia-classified data samples. The best-performing models were built upon DistilBERT and trained on either the oversampled Pitt dataset or the newly combined dataset. Through hyper-parameter tuning, we achieved state-of-the-art performance, including an accuracy of 98.8%, a macro-averaged F1-score of 98.6% and a precision of 100% for the Dementia class. Additionally, LIME was employed to explain the outputs of the model and highlight the features of interest.

Future research could explore applying the model to more recently collected data, in line with current medical practices, to evaluate its effectiveness in real-world medical applications. Furthermore, since the DementiaBank database contains transcripts in multiple languages, such as German and Mandarin, further research could be done to develop a multi-lingual dementia classifier to extend the benefits of these models globally.

Acknowledgements

We thank the Clinical NLP 2023 Workshop reviewers for their feedback. We would also like to acknowledge grants NIA AG03705 and AG05133 for supporting the development of the DementiaBank Pitt Corpus.

References

Ayimnisagul Ablimit, Catarina Botelho, Alberto Abad, Tanja Schultz, and Isabel Trancoso. 2022. Exploring Dementia Detection from Speech: Cross Corpus Analysis. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6472–6476. IEEE.

Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.

Alzheimer's Association. n.d. What is dementia? https://www.alz.org/alzheimers-dementia/what-is-dementia.

American Psychological Association. 1983. *Publications Manual*. American Psychological Association, Washington, DC.

Rie Kubota Ando and Tong Zhang. 2005. A Framework for Learning Predictive Structures from Multiple Tasks and Unlabeled Data. *Journal of Machine Learning Research*, 6:1817–1853.

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

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.

Mondher Bouazizi, Chuheng Zheng, and Tomoaki Ohtsuki. 2022. Dementia Detection Using Language Models and Transfer Learning. In *2022 The 5th International Conference on Software Engineering and Information Management (ICSIM)*, pages 152–157.

Andrea Bradford, Mark E Kunik, Paul Schulz, Susan P Williams, and Hardeep Singh. 2009. Missed and Delayed Diagnosis of Dementia in Primary Care: Prevalence and Contributing Factors. *Alzheimer Disease & Associated Disorders*, 23(4):306–314.

Centers for Disease Control and Prevention. n.d. Dementia. https://www.cdc.gov/aging/dementia/index.html.

Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 1981. Alternation. *Journal of the Association for Computing Machinery*, 28(1):114–133.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. SMOTE: Synthetic Minority over-Sampling Technique. *J. Artif. Int. Res.*, 16(1):321–357.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.

Seth A. Gale, Diler Acar, and Kirk R. Daffner. 2018. Dementia. *The American Journal of Medicine*, 131(10):1161–1169.

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.

Sweta Karlekar, Tong Niu, and Mohit Bansal. 2018. "Detecting Linguistic Characteristics of Alzheimer's Dementia by Interpreting Neural Models". In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 701–707, New Orleans, Louisiana. Association for Computational Linguistics.

Daniel Kempler. 1995. Language changes in dementia of the Alzheimer type. *Dementia and communication*, 1:98–114.

Blanka Klimova, Petra Maresova, Martin Valis, Jakub Hort, and Kamil Kuca. 2015. Alzheimer's disease and language impairments: social intervention and medical treatment. *Clinical interventions in aging*, pages 1401–1408.

M Rupesh Kumar, Susmitha Vekkot, S Lalitha, Deepa Gupta, Varasiddhi Jayasuryaa Govindraj, Kamran Shaukat, Yousef Ajami Alotaibi, and Mohammed Zakariah. 2022. Dementia Detection from Speech Using Machine Learning and Deep Learning Architectures. *Sensors*, 22(23):9311.

Eunchan Lee, Changhyeon Lee, and Sangtae Ahn. 2022. Comparative study of multiclass text classification in research proposals using pretrained language models. *Applied Sciences*, 12(9):4522.

Hali Lindsay, Johannes Tröger, and Alexandra König. 2021. Language impairment in Alzheimer's disease—robust and explainable evidence for ad-related deterioration of spontaneous speech through multilingual machine learning. *Frontiers in aging neuroscience*, page 228.

Ning Liu, Kexue Luo, Zhenming Yuan, and Yan Chen. 2022. A Transfer Learning Method for Detecting Alzheimer's Disease Based on Speech and Natural Language Processing. *Frontiers in Public Health*, 10.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Saturnino Luz, Fasih Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney. 2020. Alzheimer's dementia recognition through spontaneous speech: The ADReSS challenge. *arXiv preprint arXiv:2004.06833*.

Brian MacWhinney. 2009. The CHILDES project part 1: The CHAT transcription format.

Lovro Matošević and Alan Jović. 2022. Accurate Detection of Dementia from Speech Transcripts Using RoBERTa Model. In *2022 45th Jubilee International Convention on Information, Communication and Electronic Technology (MIPRO)*, pages 1478–1484. IEEE.

Kimberly D Mueller, Rebecca L Koscik, Bruce P Hermann, Sterling C Johnson, and Lyn S Turkstra. 2018. Declines in connected language are associated with very early mild cognitive impairment: Results from the Wisconsin Registry for Alzheimer's Prevention. *Frontiers in Aging Neuroscience*, 9:437.

Margaret A Naeser, Carol Gebhardt, and Harvey L Levine. 1980. Decreased computerized tomography numbers in patients with presenile dementia: Detection in patients with otherwise normal scans. *Archives of Neurology*, 37(7):401–409.

National Institute on Aging. 2020. What Is Dementia? `https://www.nia.nih.gov/health/what-dementia`.

Sylvester Olubolu Orimaye, Jojo Sze-Meng Wong, and Karen J. Golden. 2014. Learning Predictive Linguistic Features for Alzheimer's Disease and related Dementias using Verbal Utterances. In *CLPsych@ACL*.

J. Ramírez, J.M. Górriz, D. Salas-Gonzalez, A. Romero, M. López, I. Álvarez, and M. Gómez-Río. 2013. Computer-aided diagnosis of Alzheimer's type dementia combining support vector machines and discriminant set of features. *Information Sciences*, 237:59–72. Prediction, Control and Diagnosis using Advanced Neural Computations.

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara Parser: A Fast and Accurate Dependency Parser. *Computing Research Repository*, arXiv:1503.06733. Version 2.

Marco Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "“Why Should I Trust You?”: Explaining the Predictions of Any Classifier". In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 97–101, San Diego, California. Association for Computational Linguistics.

Alireza Roshanzamir, Hamid Aghajan, and Mahdieh Soleymani Baghshah. 2021. Transformer-based deep neural network language models for Alzheimer's disease risk assessment from targeted speech. *BMC Medical Informatics and Decision Making*, 21:1–14.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Aharon Satt, Ron Hoory, Alexandra König, Pauline Aalten, and Philippe H Robert. 2014. Speech-based automatic and robust detection of very early dementia. In *Fifteenth Annual Conference of the International Speech Communication Association*.

Aharon Satt, Alexander Sorin, Orith Toledo-Ronen, Oren Barkan, Ioannis Kompatsiaris, Athina Kokonozi, and Magda Tsolaki. 2013. Evaluation of speech-based protocol for detection of early-stage dementia. In *Interspeech*, pages 1692–1696.

Rachel A Sluis, Daniel Angus, Janet Wiles, Andrew Back, Tingting Gibson, Jacki Liddle, Peter Worthy, David Copland, and Anthony J Angwin. 2020. An automated approach to examining pausing in the speech of people with dementia. *American Journal of Alzheimer's Disease & Other Dementias®*, 35:1533317520939773.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie 2.0: A continual pre-training framework for language understanding. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8968–8975.

Akshay Valsaraj, Ithihas Madala, Nikhil Garg, and Veeky Baths. 2021. Alzheimer's Dementia Detection Using Acoustic & Linguistic Features and Pretrained BERT. *2021 8th International Conference on Soft Computing & Machine Intelligence (ISCMI)*, pages 171–175.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Alfred Wahlforss and Alexander Aslaksen Jonasson. 2020. Early dementia diagnosis from spoken language using a transformer approach. *Alzheimer's & Dementia*, 16.

Jochen Weiner, Mathis Engelbart, and Tanja Schultz. 2017. Manual and Automatic Transcriptions in Dementia Detection from Speech. In *Interspeech*, pages 3117–3121.

World Health Organization. 2021. Dementia. https://www.who.int/news-room/fact-sheets/detail/dementia. Accessed: February 27, 2023.

Vithya Yogarajan, Jacob Montiel, Tony Smith, and Bernhard Pfahringer. 2021. "Transformers for Multi-label Classification of Medical Text: An Empirical Comparison". In *Artificial Intelligence in Medicine*, pages 114–123, Cham. Springer International Publishing.

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 *Interspeech*, volume 2020, pages 2162–6.

A Example of Saliency Highlighting Using LIME

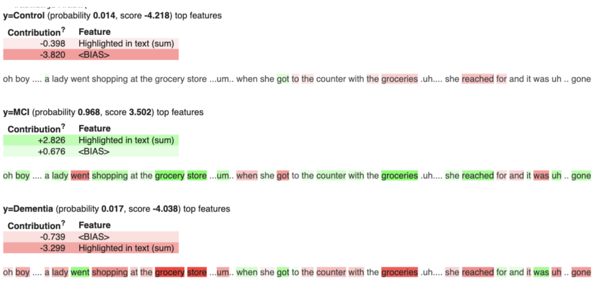

Figure 2: Example of an utterance from a patient exhibiting MCI. In conformance with data protection policies, a synthetic example is presented. The model's prediction may have been influenced by the presence of features such as "um" and "uh", which can indicate uncertainty on the part of the participant. This observation aligns with previous research that has identified the frequent use of filler words as an early indicator of dementia (Karlekar et al., 2018).

B Breakdown of the Datasets (original, oversampled and combined)

Dataset	Control	MCI	Dementia
Pitt	259	127	24
Oversampled Pitt	259	381	384
Combined P + H + K	259	127	34

Table 4: Breakdown of the Pitt dataset (original and oversampled) and the combined Pitt + Holland + Kempler (P + H + K) dataset.