Leveraging Natural Language Processing and Clinical Notes for Dementia Detection

Ming Liu[1,2,3], **Richard Beare**[1,2,3,4], **Taya Collyer**[1,2,3], **Nadine Andrew**[1,2,3], **Velandai Srikanth**[1,2,3,4]
[1] Peninsula Clinical School, Central Clinical School, Monash University, Melbourne, Australia
[2] National Centre for Healthy Ageing, Melbourne, Australia
[3] Peninsula Health, Melbourne, Australia
[4] Developmental Imaging, Murdoch Children's Research Institute, Melbourne, Australia
{grayming.liu, richard.beare}@monash.edu
{taya.collyer, nadine.andrew, velandai.srikanth}@monash.edu

Abstract

Early detection and automated classification of dementia has recently gained considerable attention using neuroimaging data and spontaneous speech. In this paper, we explore the problem of dementia detection with in-hospital clinical notes. We collected 954 patients' clinical notes from a local hospital in Melbourne and assign dementia/non-dementia labels to those patients based on clinical assessment and telephone interview. Given the labeled dementia data sets, we fine tune a ClinicalBioBERT using filtered clinical notes and conducted experiments on both binary and three class dementia classification. Our experiment results show that the fine tuned ClinicalBioBERT achieved satisfied performance on binary classification but performed poorly on three class dementia classification. We explore the difficulties we encountered applying ClinicalBioBERT to hospital text. Further analysis suggests that more human prior knowledge should be considered.

1 Introduction

Dementia describes a collection of symptoms that are caused by disorders affecting the brain. The global burden of dementia is large and expected to triple by 2050 in the absence of a treatment (Patterson, 2018). The application of deep learning to early detection and automated classification of dementia has recently gained considerable attention (Jo et al., 2019; Reuben et al., 2017), as rapid progress in neuroimaging techniques has generated large-scale multimodal neuroimaging data. The ADReSS challenge (Luz et al., 2020) released a benchmark dataset of spontaneous speech, which is acoustically pre-processed and balanced in terms of age and gender, defining a shared task through which different approaches to dementia recognition in spontaneous speech can be compared, several speech classification models were used for dementia detection, in which different types of linguistic features were extracted and fed into traditional statistical models. This study is an interesting proof of concept, with fewer than 100 patients. More recent studies (Calzà et al., 2021; Farzana et al., 2022) measured the impact of linguistic features (e.g. verbal disfluency tags) on dementia detection.

Dementia can be an underlying cause of hospital admissions, for example due to increased rates of falls in dementia sufferers. However the diagnosis associated with the admission will be a fracture, rather than dementia. In this paper, we test the possibility of early detection for dementia patients based on the in-hospital clinical notes. Specifically, we collected 954 patients' clinical notes from Melbourne Frankston hospital [1] and assign dementia/non-dementia/uncertainty labels. Given the labeled dementia data sets, we develop a deep learning model based on ClinicalBioBERT (Alsentzer et al., 2019). We experiment with both the binary (dementia/non-dementia) and the coarse (dementia/non-dementia/uncertainty) settings, and find that ClinicalBioBERT works well in the binary setting but performs poorly on the coarse setting, at the same time it still suffers from the low annotation problem, and the embedding representation is not effective as the structured representation (e.g. UMLS concept representation). "Poor" in this context is in comparison to traditional statistical and machine learning classifiers (not discussed here). Our main contributions are:

- We collected clinical text from a local hospital and provided a labeled data set for dementia detection.
- We developed a deep neural model based on ClinicalBioBERT and evaluated its performance on both binary and three-class coarse level dementia prediction, suggesting it works

[1] https://www.peninsulahealth.org.au/locations/frankston/

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 150–155

well on the binary setting but performs poorly on the coarse setting.

- We analyzed the representation power of the fine tuned ClinicalBioBERT, and find that the UMLS concept representation is stronger than the embedding one. As far as we are aware, our work is the first application to leverage deep models and clinical notes for dementia disease classification.

2 Dementia Dataset Construction

In this section, we describe how we collected the medical notes and acquired the gold-standard labels, we also show some of the basic data statistics.

2.1 Dataset collection and labeling

We recruited patients from two sources: i) a Cognitive Dementia and Memory Service (CDAMS) and ii) random selection based on attendance at the local health service. Patients attending CDAMS were split into two groups: those with a clinical diagnosis of dementia (*1a*) and those without (*1b*). Patients in group *1b* may have received a different diagnosis or not completed their assessment. Patients in group 2 were screened with the Telephone Interview for Cognitive Status (TICS-M, Australian version), with those scoring in the population average or better band after adjustment for age, sex and education (cohort *2a*) considered as free of dementia and those scoring below the average considered as uncertain (cohort *2b*). We collected documents from the in-patient electronic health record for a total of 954 patients. Table 1 shows the number of patients in each cohort. It can be seen there are much more patients in cohort *1b*, which is around half of all the patients. Also, we notice cohort 1 has more than two times patients than cohort 2, this imbalance may have some impact on later model development and cause low specificity issues.

2.2 Dataset statistics

Document Types There are various document types for the patients, including but not limited to patient demographics, medications, vital signs, past medical history description, radiology report and progress note. We noticed that the progress notes were the majority (24.89%) types for the patients, as shown in Table 1, patients in cohort 1 had more than 2% progress notes than that in cohort 2. The *1a* cohort has largest number of progress notes (26.61%), which is reasonable as those patients may have more times of visits than other groups.

Document Counts and Length We also calculated the statistics of document counts and length for each cohort. As shown in Table 2, document counts for patients from different cohorts vary significantly, while the document average lengths from the four cohorts is more or less similar. More specifically, patients in cohort 1 tend to have around 4 times as many documents (283) as those in cohort 2a (66). Patients in cohort 2a had fewer documents, because the randomly selected patients were usually less complex and had fewer admissions than cohort 1. However, we cannot use document count as an input feature for later statistical modelling as other complex disorders are likely to have similar document counts to dementia patients.

An example We show a progress note with demographic information removed for a patient from cohort *1a* as the following: *[Progress Note: Pt ambulated to toilet Independently with x1 assist While coming out from Toilet ,pt become agitated and aggressive towards author T/L involved and pt stating his Meds is not given Though Writer mentioned this matter to Treating Dr earlier, couldn't chart the meds as he hasn't had the list of meds Informed to Treating Dr and NIC Contacted wife over the phone and treating Dr spoke to her Pt become calm and has had Meds as per MAR.]* We notice three important characteristics for such clinical text: i) abbreviations, ii) spelling errors - clinical staff complete documents under time pressure and spelling errors are common. iii) Long distance context, as the history notes may also needed to give full interpretation for the current text.

3 Methodology

In this section, we describe the development of classification models using the data sets described above. There are two aspects to consider before the model development. First, what is the classifier's granularity? There could be three levels of input, i.e., sentence, document and patient (multi-document) level. Typically, it is more challenging to achieve high performance when the input text is longer. However, developing a sentence or single document level classification model requires fine grained annotation, which is often time consuming and expensive in the medical setting. Meanwhile,

Cohort	Description	Patient counts	Progress note Pct.
1a	*Diagnosed as dementia in CDAMS, Positive*	245	26.61%
1b	*No final diagnosis in CDAMS, Uncertain*	419	24.10%
2a	*Diagnosed as non-dementia via TICS, Negative*	196	22.94%
2b	*No diagnosis via TICS, Uncertain*	99	23.02%
Total	-	959	24.89%

Table 1: Patient statistics

Cohort	Max	Min	Mean	Std	Median
1a	2150 (4750)	1 (170)	283 (871)	356 (319)	125 (847)
1b	3770 (4678)	1 (239)	241 (898)	410 (308)	79 (887)
2a	737 (2586)	1 (393)	66 (813)	90 (269)	34 (775)
2b	347 (2829)	1 (400)	74 (893)	79 (340)	48 (822)
All	3770 (4750)	1 (170)	199 (873)	340 (309)	67 (844)

Table 2: The statistics for document counts (document length) in each cohort.

multi-instance learning may further improve the complexity in the prediction stage. Therefore, we aim to develop a patient level classification model directly. Second, what types of Machine Learning models can be used? We consider the recent deep neural models (e.g. BERT), but since BERT is pre-trained from generic text, we will fine tune a ClinicalBioBERT (Alsentzer et al., 2019) [2] due to its domain similarity.

Clinical Note Filtering and Compression For the BERT based classification model, we choose ClinicalBioBERT as the pre-trained LM, and fine tune it with the medical text from each patient. However, as most BERT based models can only take 512 tokens as the maximum input, it is necessary to compress each patient's notes within that length. We consider several strategies: The first one is to filter out the notes where there are structured notes, as these structure information are often progress notes and not disease specific. The second strategy is to annotate some key sentences and build a sentence level classifier, and use the classifier to filter and shorten the clinical notes. However, it is expensive and requires further human annotation. The third strategy is truncation based on the latest notes, as in table 2, we show the average clinical note length for all patients is 873, in our text pre-processing stage we notice there are at least 10 annotated UMLS concepts for a clinical note if the number of tokens in it is over 873. Therefore, we use a simple and realistic heuristic by keeping those medical notes in which there are at least 10 UMLS concepts, and aggregate the latest notes to represent the patient note summary.

Fine tune ClinicalBioBERT After getting the patient note summary, we pair those summaries with their cohort labels and fine tune ClinicalBioBERT. We add the [CLS] token at the beginning of the patient note summary and use it as the hidden representation. During fine tuning, we update all the transformer layers and use Adam(Kingma and Ba, 2014) as the optimizer.

4 Experiments

Experiment Setup We experimented with two classification schemes: binary and three-class. For the binary case, only the medical notes for patients from cohorts *1a* and *2a* were selected, which is an exact binary classification (1a v.s. 2a) setting. In contrast, in the three-class setting, we regard the *1b* and *2b* cohorts as the uncertainty group, which returns the three class (1a v.s. 2a v.s. uncertain) setting. For the ClinicalBioBERT model, we keep the default settings and trained 20 epochs until convergence. Like other biomedical settings, we use accuracy, precision, recall and Micro F1 as the evaluation metrics. We also add a keyword method as the naive baseline, where we use a pre-recognized 245 UMLS concept names as a keyword list, these concepts are recognized by human experts to correlated with dementia. If any of those concept names appear in the document, we give a prediction of

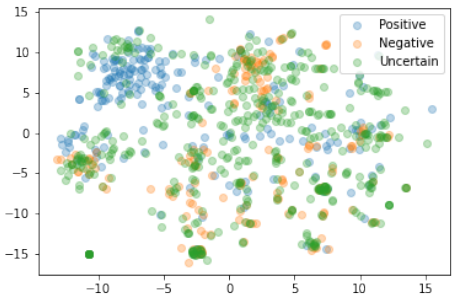
[2]The ClinicalBioBERT model was trained on all notes from MIMIC-III, a database containing electronic health records from ICU patients at the Beth Israel Hospital in Boston, MA. Model can be found from https://huggingface.co/emilyalsentzer/Bio_ClinicalBERT

Models	Accuracy	Precision	Recall	F1
Keyword-based (binary)	0.453 (0.021)	0.469 (0.023)	0.482 (0.025)	0.475 (0.021)
Keyword-based (coarse)	0.398 (0.051)	0.382 (0.052)	0.393 (0.043)	0.387 (0.052)
ClinicalBioBERT (binary)	**0.810 (0.041)**	**0.832 (0.051)**	**0.801 (0.052)**	**0.814 (0.050)**
ClinicalBioBERT (coarse)	0.458 (0.045)	0.449 (0.043)	0.406 (0.046)	0.381 (0.045)

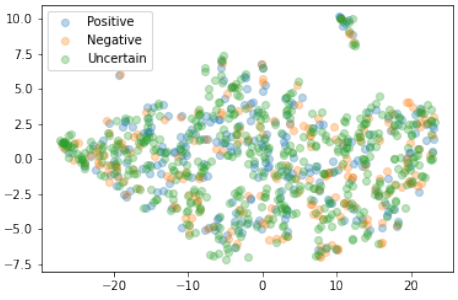
Table 3: Binary/coarse dementia classification results for the fine Tuned ClinicalBioBERT model, the binary classification are for (1a vs 2a), the coarse classification are for (1a vs 2a vs 1b, 2b) The numbers in the parenthesis show the standard deviation of the ten runs.

positive dementia, otherwise negative.

Results We perform 10-fold cross validation on the selected patients given the binary and three class setting. Table 3 shows the key results. In general, we find that the fine tuned ClinicalBioBERT performs well in the binary setting with a 0.81 accuracy, but it dropped significantly in the three class coarse setting. Meanwhile, it is shown that the performance of both models decreased around 20% from the binary to the three class setting.



(a) t-SNE for UMLS concept representation



(b) t-SNE for ClinicalBioBERT representation

Figure 1: We apply t-SNE for the 954 patients' feature representation with (a) 245 UMLS concepts and (b) the [CLS] embedding of the note summaries. In general, the UMLS concept representation distinguish positive and negative dementia patients better.

5 Analysis

Even though BERT based models show superior performance on most generic text classification tasks, the fine tuned ClinicalBioBERT does not exhibit satisfied results in the coarse setting in this study. We anticipate three reasons: First, the clinical notes are too long for ClinicalBioBERT to encode, since the standard BERT models can only take an input length of 512 tokens. Meanwhile, the dementia related text spans are quite sparse in the clinical notes, further text compression and selection heuristics are required. Furthermore, the BERT based modeling techniques cannot leverage expert prior knowledge, which in this study are the filtered UMLS concepts. To validate our hypothesis, We apply t-SNE for the 954 patients' feature representation with either the 245 UMLS concepts or the [CLS] embedding of the clinical note summary. As shown in figure 1 (a), the UMLS concept representation is more meaningful, as those positive dementia patients can be easily separated with those negative patients, while in figure 1 (b) there is no clear representation patterns for these three classes.

6 Related Work

Clinical text representation and classification When clinical text classification is used for disease detection tasks, it varies a lot from generic text classification: (i) Traditional text classification tasks take both precision and recall as the system measurement, while recall is considered to be top priority in most medical text classification tasks (Spasic et al., 2020) because doctors would never like to miss the information of any "likely" infected patients. That is, the system is being used to screen latent potential candidates. (ii) Annotation cost is higher in the medical domain (Wei et al., 2019) because professional skills from medical experts are needed. In common text annotation tasks, it is not necessary to hire highly skilled people and even

crowd sourcing can be used. (iii) The text in the medical domain contains a lot of abbreviations, jargon and acronyms for different medical concepts (Xu et al., 2007). (iv) There are patient records which are sequential and correlated within each other. A patient can have multiple reports, in which each report is the description of a specific time period. The classification for these multiple reports varies based on time, so there should be some level consistency. ClinicalXLNet (Huang et al., 2019) was recently developed to model such sequential clinical text. (v) Medical text for a patient can come from different sources (Yang and Wu, 2021) such as CT scans, blood scans and operation reports, etc.

Disease detection with NLP Before this study, we have previously explored automatic fungal disease detection with radiology reports (Liu et al., 2016, 2017; Baggio et al., 2019) and showed the effectiveness of various NLP models on clinical notes. Even though deep learning has revolutionized the ML applications, Sheikhalishahi et al. (2019) reviewed the ML models on chronic diseases with clinical notes and showed that more than 90% of the methods still relied on statistical models. Wang et al. (2020) conducted a systematic evaluation of NLP in medicine over the past 20 yeears, they showed that cancer (24.94%) was the most common subject area in NLP-assisted medical research on diseases, with breast cancers (23.30%, 24/103) and lung cancers (14.56%) accounting for the highest proportions of studies.

Dementia detection The application of deep learning to early detection and automated classification of dementia has recently gained considerable attention (Jo et al., 2019), as rapid progress in neuroimaging techniques has generated large-scale multimodal neuroimaging data. The ADReSS challenge (Luz et al., 2020) released a benchmark dataset of spontaneous speech, which is acoustically pre-processed and balanced in terms of age and gender, defining a shared task through which different approaches to dementia recognition in spontaneous speech can be compared. More recently, Farzana et al. (2022) measured the impact of verbal disfluency tags on denmentia detection.

Biomedical language models Most biomedical language models are pre-trained with BERT (Devlin et al., 2018) and related clinical text. For example, the ClinicalBioBERT (Alsentzer et al., 2019) model was trained on all notes from MIMIC-III (Johnson et al., 2016), a database containing electronic health records from ICU patients at the Beth Israel Hospital in Boston, MA. MedBERT (Rasmy et al., 2021) was pretrained on a structured EHR dataset of 28,490,650 patients.

7 Conclusion

In this work, we collected clinical text from a local hospital and leveraged deep neural models for dementia detection. We fine tuned a ClinicalBioBERT and evaluated its their performance on dementia classification, experiment results showed that the fine tuned model works well on binary dementia classification but fails on three class dementia classification. As for the future work, we will leverage more human prior knowledge and experiment with both statistical and deep neural models. Also, more structured patient representation using knowledge graphs will be considered.

8 Limitation

There are a few limitations of this study: First, the patient sample size for the validation cohorts was limited to 954 patients from a local hospital. As annotation in the medical setting is expensive and time consuming, we only get patient level labels and cannot pay the effort for document level annotations. The size and diversity of the data sample could be improved by collecting clinical notes for patients from other hospitals in different age groups and of similar clinical complexity. We did not perform cross label check for the sampled patients, as there is a large number of uncertain patients, among those patients there are still ones who suffer from dementia but not diagnosed. Second, more statistical models can bee developed. At the moment we only tried a keyword based model and a deep neural models. Traditional statistical models like Logistic Regression with biomedical concept features can also be considered. Furthermore, our study would have benefited from more model interpretability and human error analysis on the classifier predictions. We have plans to extend our current work with the above mentioned directions.

Acknowledgements

This work was supported by the Australian Medical Research Future Fund (RRDHI000088).

References

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*.

Diva Baggio, Trisha Peel, Anton Y Peleg, Sharon Avery, Madhurima Prayaga, Michelle Foo, Gholamreza Haffari, Ming Liu, Christoph Bergmeir, and Michelle Ananda-Rajah. 2019. Closing the gap in surveillance and audit of invasive mold diseases for antifungal stewardship using machine learning. *Journal of Clinical Medicine*, 8(9):1390.

Laura Calzà, Gloria Gagliardi, Rema Rossini Favretti, and Fabio Tamburini. 2021. Linguistic features and automatic classifiers for identifying mild cognitive impairment and dementia. *Computer Speech & Language*, 65:101113.

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*.

Shahla Farzana, Ashwin Deshpande, and Natalie Parde. 2022. How you say it matters: Measuring the impact of verbal disfluency tags on automated dementia detection. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 37–48.

Kexin Huang, Abhishek Singh, Sitong Chen, Edward T Moseley, Chih-ying Deng, Naomi George, and Charlotta Lindvall. 2019. Clinical xlnet: Modeling sequential clinical notes and predicting prolonged mechanical ventilation. *arXiv preprint arXiv:1912.11975*.

Taeho Jo, Kwangsik Nho, and Andrew J Saykin. 2019. Deep learning in alzheimer's disease: diagnostic classification and prognostic prediction using neuroimaging data. *Frontiers in aging neuroscience*, 11:220.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

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

Ming Liu, Gholamreza Haffari, and Wray Buntine. 2016. Learning cascaded latent variable models for biomedical text classification. In *Proceedings of the Australasian Language Technology Association Workshop 2016*, pages 128–132.

Ming Liu, Gholamreza Haffari, Wray Buntine, and Michelle Ananda-Rajah. 2017. Leveraging linguistic resources for improving neural text classification. In *Proceedings of the australasian language technology association workshop 2017*, pages 34–42.

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*.

Christina Patterson. 2018. World alzheimer report 2018.

Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. 2021. Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*, 4(1):1–13.

David B Reuben, Andrew S Hackbarth, Neil S Wenger, Zaldy S Tan, and Lee A Jennings. 2017. An automated approach to identifying patients with dementia using electronic medical records. *Journal of the American Geriatrics Society*, 65(3):658–659.

Seyedmostafa Sheikhalishahi, Riccardo Miotto, Joel T Dudley, Alberto Lavelli, Fabio Rinaldi, Venet Osmani, et al. 2019. Natural language processing of clinical notes on chronic diseases: systematic review. *JMIR medical informatics*, 7(2):e12239.

Irena Spasic, Goran Nenadic, et al. 2020. Clinical text data in machine learning: systematic review. *JMIR medical informatics*, 8(3):e17984.

Jing Wang, Huan Deng, Bangtao Liu, Anbin Hu, Jun Liang, Lingye Fan, Xu Zheng, Tong Wang, Jianbo Lei, et al. 2020. Systematic evaluation of research progress on natural language processing in medicine over the past 20 years: bibliometric study on pubmed. *Journal of medical internet research*, 22(1):e16816.

Qiang Wei, Yukun Chen, Mandana Salimi, Joshua C Denny, Qiaozhu Mei, Thomas A Lasko, Qingxia Chen, Stephen Wu, Amy Franklin, Trevor Cohen, et al. 2019. Cost-aware active learning for named entity recognition in clinical text. *Journal of the American Medical Informatics Association*, 26(11):1314–1322.

Hua Xu, Peter D Stetson, and Carol Friedman. 2007. A study of abbreviations in clinical notes. In *AMIA annual symposium proceedings*, volume 2007, page 821. American Medical Informatics Association.

Bo Yang and Lijun Wu. 2021. How to leverage multimodal ehr data for better medical predictions? *Conference on Empirical Methods in Natural Language Processing*.