Using Deep Learning to Identify Patients with Cognitive Impairment in Electronic Health Records

Abstract

Dementia is a neurodegenerative disorder that causes cognitive decline and affects more than 50 million people worldwide. Dementia is under-diagnosed by healthcare professionals — only one in four people who suffer from dementia are diagnosed. Even when a diagnosis is made, it may not be entered as a structured International Classification of Diseases (ICD) diagnosis code in a patient's charts. Indeed, information relevant to cognitive impairment (CI) is often found within electronic health records (EHR) but manual review of clinician notes by experts is both time consuming and often prone to errors. Automated mining of these notes presents an opportunity to label patients with cognitive impairment in EHR data. We developed natural language processing (NLP) tools to identify patients with cognitive impairment and demonstrate that linguistic context enhances performance for the classification task. We fine-tuned our attention based deep learning model, which can learn from complex language structures, and substantially improved accuracy (0.93) relative to a baseline NLP model (0.84). Further, we show that deep learning NLP can successfully identify dementia patients without dementia-related ICD codes or medications.

Keywords: EHR, NLP, Dementia

1. Introduction

Dementia is the most common neurodegenerative disease affecting older adults, progressing from mild cognitive impairment (MCI) to mild, moderate, and severe dementia. Dementia is under-diagnosed: dementia is not formally diagnosed or coded in claims data for over 50% of older adults living with probable dementia. Often, a diagnosis is given once patient has reached moderate dementia, and irreversible damage has already been done to the brain. The early detection of the first signs of cognitive impairment, however, is important for improving clinical outcomes and patient management. Tools that can efficiently and effectively analyze medical records for warning signs of dementia and recommend patients for follow up with a specialist can be critical to obtaining an early diagnosis for dementia. Such a tool could also be useful in recruiting into clinical trials as well as a variety of research studies ranging from in-silico trials for drug repurposing to evaluating how policies and programs meet the needs of patients and caregivers. We aim to use NLP to detect signs of cognitive impairment from unstructured clinician notes by using deep learning techniques. We apply our deep learning algorithm to patients in Mass General Brigham (MGB) Healthcare who have genotype data available from the MGB BioBank. An overview of our project can be found in Appendix A.

2. Related Works

Prior works have used NLP techniques to detect various diseases from EHR. (Rajkomar et al., 2018) used recurrent neural networks (long short-term memory (LSTM)) among others to predict inpatient mortality using EHR data from the University of California, San Francisco (UCSF) from 2012 to 2016, and the University of Chicago Medicine (UCM) from 2009 to 2016. (Glicksberg et al., 2018) performed phenotyping for diseases such as Attention Deficit Hyperactivity Disorder (ADHD) by clustering on word2vec embeddings from EHR of the Mount Sinai Hospital (MSH) in New York City. These studies have shown that the application of NLP techniques to EHR have improved disease detection, and that NLP techniques can be applied to dementia detecton to achieve similar results. Our work uses deep learning NLP techniques, which has achieved impressive results when applied to general text due to the use of word embeddings and attention-based models (Vaswani et al., 2017; Mikolov et al., 2013; Pennington et al., 2014; Peters et al., 2018; Devlin et al., 2018), but have had limited applications in healthcare, particularly in dementia research.

Table 1: Demographics of Data

Characteristic	**(N = 16428)**
Age (years) mean (SD)	73.01 (7.96)
Gender Male, *n* (%)	8740 (53.2)
Race, *n* (%)	
White	14896 (90.7)
Other/Not Recorded	608 (3.7)
Black	570 (3.5)
Hispanic	170 (1.0)
Asian	168 (1.0)
Indigenous	16 (0.01)
APOE Genotype, *n* (%)	
APOE $\varepsilon 2$	2028 (12.3)
APOE $\varepsilon 3$	10177 (62.0)
APOE $\varepsilon 4$	4223 (25.7)
Average Speciality Visits (SD)	1.67 (4.6)
Average PCP Encounters (SD)	5.25 (5.63)

3. Dataset, Preprocessing, and Annotations

Dataset Our dataset consisted of a cohort (N = 16,428) of patients from the Mass General Brigham (MGB) HealthCare (formerly Partner's Healthcare, comprising two major academic hospitals, community hospitals, and community health centers in the Boston area) system who were older than 60 years (as of July 13, 2021), had APOE genotype (Mahley and Rall Jr, 2000) (biggest genetic risk factor for dementia) data available from the MGB BioBank, and at least one clinician note with a dementia-related keyword. Table 1 shows demographics of the cohort of patients.

Preprocessing For each patient in our dataset, we extracted unstructured clinician notes, identified matches to 18 dementia-related keywords (Appendix D, including those related to memory, cognition, neuropsychological tests, and dementia diagnoses. We constructed sequences from the note text spanning each of these matches (of length 800 characters). Our cohort of 16,428 patients had 279,224 sequences with dementia-related keywords in total.

Annotations A subset of sequences was annotated for indication of cognitive impairments. We defined cognitive impairment as evidence of MCI, where one cognitive domain is involved, or dementia, where more than one cognitive domain is involved and activities of daily living are affected. Concern from the family of the patient or the patient was not considered as cognitive impairment. Experts annotated sequences using a web-based annotation tool (Appendix B) as 1) Yes, i.e., patient has CI; 2) No i.e., Patient does not have CI; and 3) Neither i.e., sequence has no information on patient's cognition. Appendix C shows examples of example sequences for all 3 classes.

We assigned 5,000 diverse sequences containing at least one match to every keyword from 5,000 unique patients for labeling. In order to expedite annotations, we utilized an "always pattern" scheme. An always pattern is defined as a phrase or regex expression that in any context indicates the phrase will be labeled with a particular class (i.e. yes, no, or neither). Once an always pattern is defined, all other sequences that match the pattern are automatically labeled with that always pattern's class. For examples of always patterns for all three classes (Yes, No, Neither), see Appendix C.

The final dataset of 8,656 annotated sequences from N = 2,487 unique patients was split between train (90%) and holdout test (10%) sets, stratified across label and proportion of sequences annotated manually and through always patterns. Validation datasets were split from the train set using techniques described in the Methodology section.

Table 2: Model Performance

Model	AUC	Accuracy	Sensitivity	Specificity	Micro F1	Macro F1	Weighted F1
TF-IDF	0.95	0.84	0.83	0.92	0.84	0.81	0.84
ClincialBERT	0.98	0.93	0.91	0.96	0.93	0.92	0.93

4. Methodology

We developed two NLP models for the classification task and compared them to each other.

(1) Logistic Regression with TF-IDF Vectors We performed TF-IDF (term frequency-inverse document frequency) vectorization on the annotated sequences and selected features based on a term's Pearson correlation coefficient with the outcome. L1 Regularized logistic regression (Tibshirani, 1996) was applied with the annotated cognitive impairment labels. We used 10-fold cross validation to determine the optimal lambda value and correlation coefficient threshold to select features.

(2) Transformer Based Sequence Classification Language Model We utilized a pre-trained language model called ClinicalBERT (Alsentzer et al., 2019), which was trained on the MIMIC II (Saeed et al., 2011) database containing EHR records from ICU patients. We used the implementation in the Huggingface Transformers (Wolf et al., 2019) and Simpletransformers (Rajapakse, 2020) packages. After text preprocessing, input texts were tokenized with the default tokenizer and converted to embeddings. The model was initialized with pre-trained parameters and later fine-tuned on our labeled training set. Optuna (Akiba et al., 2019) was used to perform a 20-trial study and tune the learning rate, adam epilson, and the number of train epochs on the held-out validation set to maximize AUC. An early stopping rule was used to prevent overfitting by ensuring that training stopped if the loss did not change substantially over 3 epochs.

5. Results

We evaluated each model based on sequence level class assignments. Model performance for each model on the held-out test set are shown in Table 2. To compute each metric, we used the threshold that maximized accuracy. The TF-IDF model achieved an AUC of 0.95 and accuracy of 0.84. The 20 words with the highest correlation coefficients using TF-IDF word vectorization are shown in Appendix E. While TF-IDF was able to identify the presence of a keyword or always pattern in a sequence, it was unable to the leverage the context around each keyword match. The context of the keywords and the agents within the sentence often contained useful information regarding a patient's cognitive status. For example, the sentence "Patient is caregiver for wife who has dementia" has the keyword dementia, but does not pertain to the patient's cognitive diagnosis. This led the baseline TF-IDF model to incorrectly predict sequences as evidence of cognitive impairment, resulting in a large count of false positives.

ClinicalBERT, with its more complex architecture, was able to leverage the context of the keyword matches within the sequences and overcome these issues. The fine-tuned ClinicalBERT model achieved an AUC of 0.98 and substantially improved accuracy to 0.93 (specificity of 0.96, sensitivity of 0.91, micro F1 of 0.93, macro F1 of 0.92, and weighted F1 of 0.93).

In order to generate patient level class assignments, we applied ClinicalBERT to all 186,730 sequences from the N = 13,941 unique patients that were not patient of our training/validation/set sets. With these sequence level predictions, we generated patient level class assignments by assigning patients a cognitive impairment label if their number of sequences predicted positive was greater than an empirically tuned threshold. We identified the most optimal threshold (from a range of 1 - 10) by comparing the percentage of patients being predicted as having cognitive impairment stratified by APOE allele to the percentages of patients with cognitive impairment related Meds/ICD codes stratified by APOE allele (MED/ICD code column in Table 3). Table 3 shows the comparison of Med/ICD codes to ClinicalBERT patient level class assignments with a sequence threshold of 2. As shown, ClinicalBERT was able to identify a significant proportion of patients that went undetected by current clinical methods, highlighting the utility of such a tool in a clinical setting.

Table 3: Comparison between Other Indicators of Cognitive Impairment and ClinicalBERT

	Count	Yes (%)	No/Ntr (%)	Med/ICD Code (%)
APOE ε2	1754	0.17	0.83	0.11
APOE ε3	8751	0.17	0.83	0.11
APOE ε4	3436	0.21	0.79	0.17

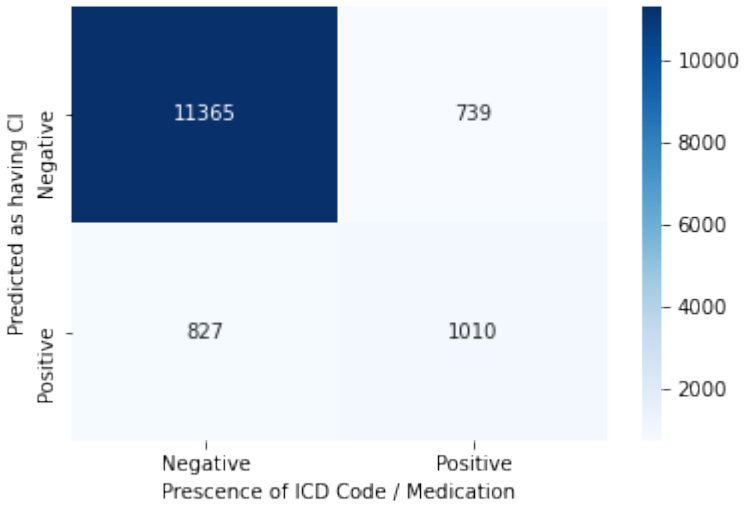

Figure 1: Confusion Matrix Patient Level Prediction Counts for ClinicalBERT

6. Conclusion and Future Work

We applied NLP algorithms to identify patients with cognitive impairment in EHR and compared a baseline TF-IDF model with an attention based deep learning model on performance of sequence level class assignment predictions. Our work can help combat the under-diagnosis of dementia and alert caregivers to do a formal cognitive evaluation or refer to specialists. Such a tool can be used to generate cohorts for dementia research studies to identify risk and protective factors of dementia as well as recruit patients into observational studies or clinical trials.

The deep learning model's performance was significantly better than the TF-IDF model as it was able to fully leverage the context of sequences. Our work illustrates the need of more complex, expressive language models for the nuanced task of detecting dementia in electronic health records.

We used the sequence level class assignments of the deep learning model to generate patient level classes. We show that our model can successfully identify patients with cognitive impairment who lack dementia-related ICD codes or medications in their records. However, a lack of patient level annotations prevents us from measuring the true accuracy of our results. In order to address this issue, we plan to generate 1000 patient level class assignments using our annotation tool. We also plan to further improve the generalizability of our models by labeling more sequences that do not match an always pattern. Further, we plan to use an active learning loop to pick particular sequences by using entropy and diversity measures. We will label uncertain sequences and use UMAP clustering (McInnes et al., 2018) on embeddings of the sequences to pick from each distinct cluster. The new gold-standard dataset will serve as the basis for the next iteration of the active learning loop to further improve model performance and develop a more generalizable model that can detect patients with cognitive impairment in electronic health records.

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2623–2631, 2019.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Benjamin S Glicksberg, Riccardo Miotto, Kipp W Johnson, Khader Shameer, Li Li, Rong Chen, and Joel T Dudley. Automated disease cohort selection using word embeddings from electronic health records. In *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2018: Proceedings of the Pacific Symposium*, pages 145–156. World Scientific, 2018.

Robert W Mahley and Stanley C Rall Jr. Apolipoprotein e: far more than a lipid transport protein. *Annual review of genomics and human genetics*, 1(1): 507–537, 2000.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.

Thilina Rajapakse. Simple transformers, 2020. URL https://simpletransformers.ai/.

Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaj, Michaela Hardt, Peter J Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, et al. Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine*, 1(1):1–10, 2018.

Mohammed Saeed, Mauricio Villarroel, Andrew T Reisner, Gari Clifford, Li-Wei Lehman, George Moody, Thomas Heldt, Tin H Kyaw, Benjamin Moody, and Roger G Mark. Multiparameter intelligent monitoring in intensive care ii (mimic-ii): a public-access intensive care unit database. *Critical care medicine*, 39(5):952, 2011.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771, 2019.

Appendix A. Overview

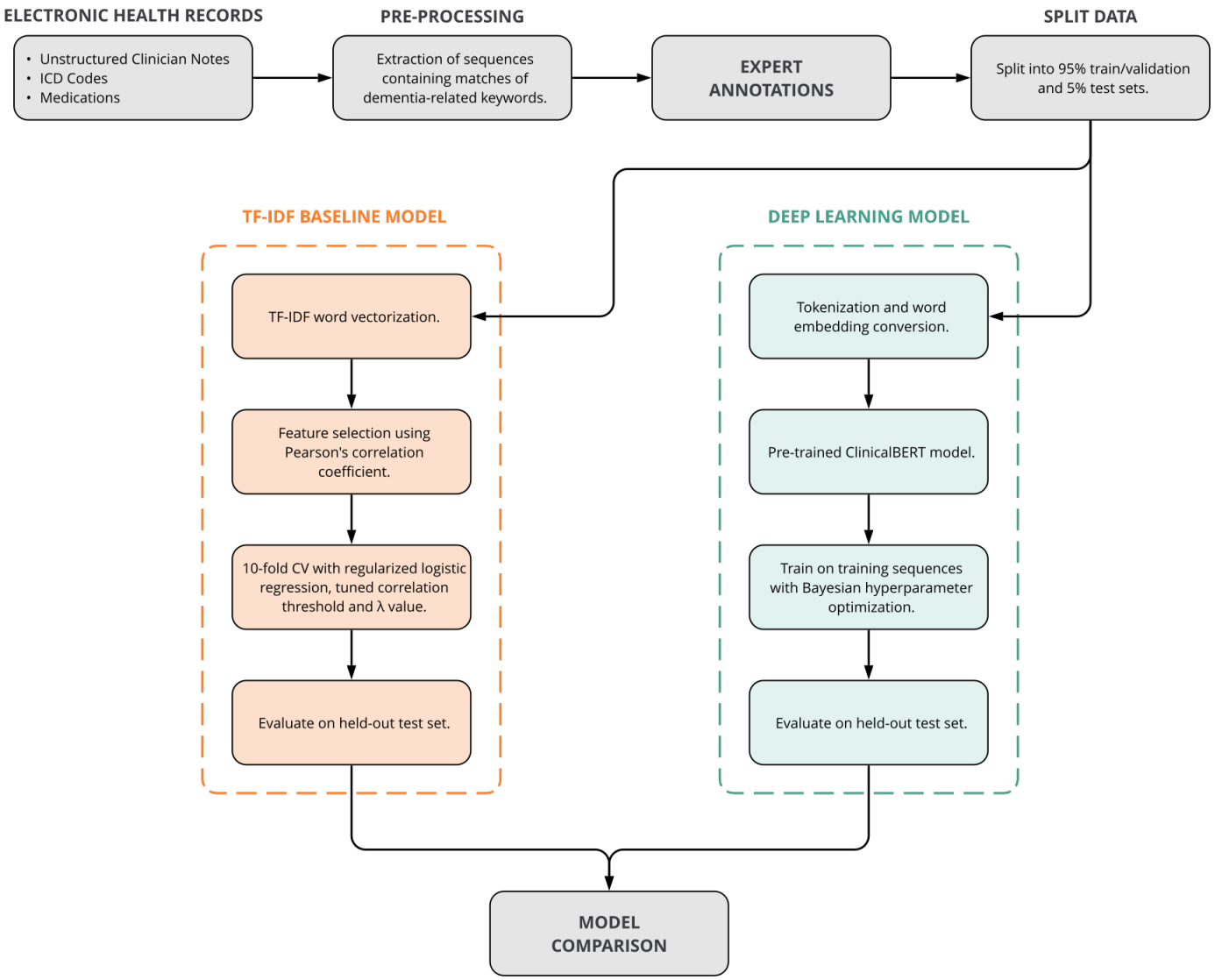

Figure 2: Overview

Appendix B. Pictures of UI Interface for Annotations

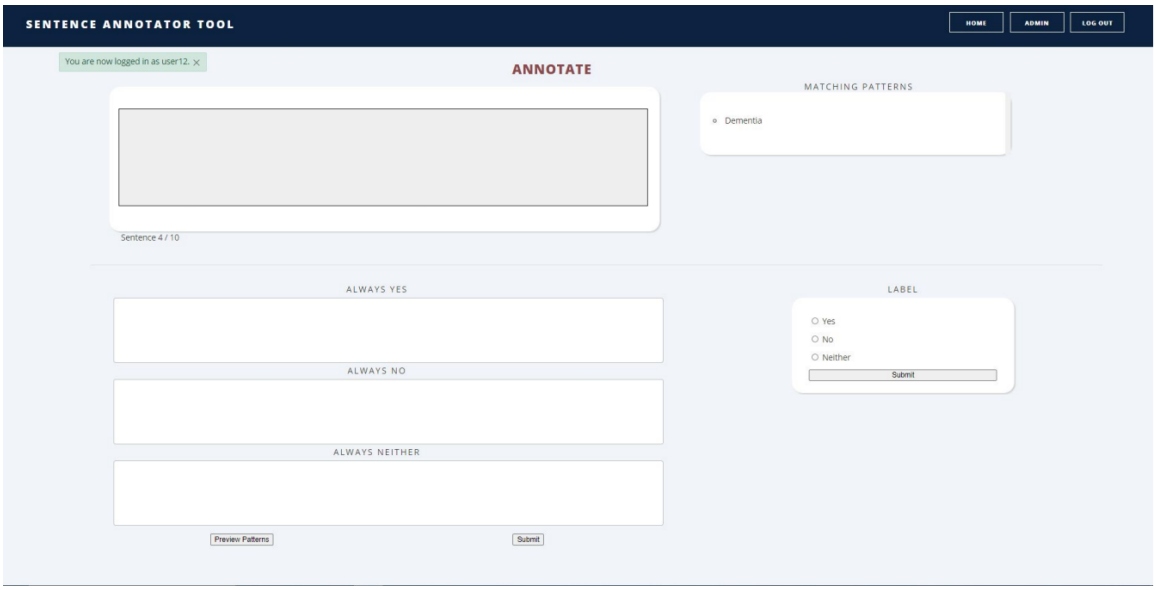

Figure 3: Annotation UI

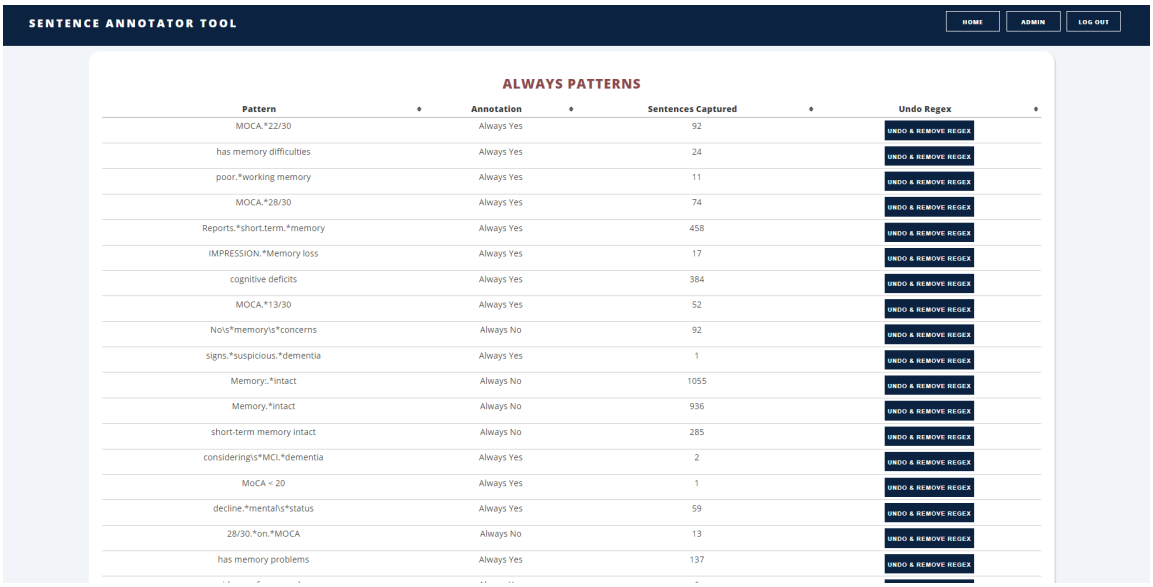

Figure 4: Always Pattern List Generated by Annotations

Appendix C. Example Sequences

Positive Sequences	Positive Always Patterns
1. Patient MOCA is 22/30.	1. (?i)\bMOCA\s*([0-9][12][0-5])\s*/\s*30
2. Patient with past medical history of dementia.	2. (?i)\bpast\s*medical\s*history\s*[^.]*(dementia)

Negative Sequences	Negative Always Patterns
1. Patient memory is intact.	1. (?i)Memory.*intact
2. No memory concerns.	2. (?i)No\s*memory\s*concerns

Neither Sequences	Neither Always Patterns
1. History: Father has Alzheimer's Disease	1. (?i)Father.*Alzheimer's\s*disease
2. Patient attends anticoagulation therapy daily.	2. (?i)anticoagulation

Figure 5: Example Sequences

Appendix D. Keywords

Keyword	Match Count
Memory	109218
Cognition	87655
Dementia	51034
Cerebral	45886
Cerebrovascular	36370
Cerebellar	26863
Cognitive Impairment	20267
Alzheimer	20581
MOCA	9767
Neurocognitive	7711
MCI	3889
Amnesia	3695
AD	2673
Lewy	2561
MMSE	2134
LBD	224
Corticobasal	147
Pick's	41

Table 4: Keywords indicative of Cognitive Impairment

Appendix E. Top TF-IDF Word Features

Word	Corr	Word	Corr
Intact	0.56	**Experiences**	0.36
Oriented	0.43	**Associations**	0.36
Concentration	0.42	**Homicidal**	0.36
Orientation	0.41	**Observation**	0.36
Sensorium	0.40	**Knowledge**	0.36
Perceptions	0.40	**Abstract**	0.36
Judgement	0.39	**Suicidal**	0.35
Fund	0.38	**Attention**	0.35
Insight	0.36	**Content**	0.34
Ideation	0.36	**Thought**	0.34

Table 5: Top 20 TF-IDF Word Features and their Correlation Coefficient