Natural Language Processing Methods to Extract Lifestyle Exposures for Alzheimer's Disease from Clinical Notes

Yoonkwon Yi*†, Zitao Shen*†, Anusha Bompelli‡, Fang Yu§, Yanshan Wang††, Rui Zhang‡**
†*College of Science & Engineering*, ‡*College of Pharmacy*, §*School of Nursing*, **Institute for Health Informatics*,
University of Minnesota, Minneapolis, ††*Mayo Clinic, Rochester, MN, USA*

***Abstract*—Due to the absence of medications on Alzheimer's disease (AD), lifestyle exposures that could improve cognitive functionality have become extremely important. Thus, the objective of the study was to show the feasibility of using natural language processing (NLP) methods to extract lifestyle exposures from clinical texts. The proposed named-entity recognition (NER) task's results indicate that NLP models can detect lifestyle information (i.e., excessive diet, physical activity, sleep deprivation and substance abuse) from clinical notes, which has the potential for improving efficiency in information acquisition and accrual for AD clinical trials.**

***Index Terms*—Alzheimer's disease, Lifestyle exposure, Electronic health records, Information extraction, Natural language processing, Machine learning, Deep learning**

I. Introduction

In the United States, Alzheimer's disease (AD) is the 6th leading cause of death [1]. Unfortunately, no treatments can yet prevent or cure AD. However, lifestyle factors have been associated with substantially reduced risk for AD or delay its onset, but the strength of evidence for each exposure varies. Individually, physical and cognitive activities show the strongest association with AD risk reduction, ranging from 11% to 44% [2]. Collectively, multiple lifestyle exposures seem to have an additive or synergistic effect on AD. A study showed a 60% lower risk of AD when four or all the five specified lifestyle behaviors (physical activity, not smoking, light-to-moderate alcohol consumption, high-quality diet, and cognitive activities) were followed [3].

Electronic health records (EHR) data is unstructured, thus making it difficult to extract desired information. Natural Language Processing (NLP) techniques have shown promising results in unlocking unstructured data to support clinical research. For example, Zhu and Razavian studied the use of deep learning models to predict AD using EHRs, showing the viability of utilizing EHRs to answer a variety of research questions in AD [4]. In our previous study, we found that lifestyle exposures such as substance use, physical activity, and diet had been recorded in clinical notes [5]. No studies have investigated the extraction of lifestyle exposures from EHRs using conventional learning models.

The objective of the study was to show the feasibility of using conventional machine learning and deep learning models to extract Alzheimer's Disease related lifestyle exposures from clinical texts. To the best of our knowledge, this is the first study to explore deep learning and conventional machine learning models for extracting lifestyle information. Our contributions include: 1) collecting and annotating lifestyle exposures to create a clinical corpus and 2) development and evaluation of conventional machine learning and deep learning models on extracting lifestyle knowledge.

† These two authors contributed equally to the work.
** Corresponding author

II. Methodology

The proposed named-entity recognition (NER) task was performed following these steps: 1) obtaining the AD patient's clinical texts with mentions of selected lifestyle exposures; 2) annotating selected sentences to develop the gold standard corpus (GSC); 3) dividing the GSC into 4 sub-datasets based on the lifestyle exposures' related categories; 4) developing the NER algorithms 5) evaluating the model performance.

A. Data Collection

Data was collected from the clinical data repository (CDR) from the University of Minnesota (UMN). The IRB (Institutional Review Board) approval was obtained for accessing the EHRs for patients with AD. We manually collected all CUIs associated with lifestyle exposures by using online UMLS Terminology searching browser based on our previous work [6]. We have identified 25,601 unique patients with AD in the CDR. Besides, there were 69,877 mentions on the lifestyle exposures related concept unique identifier (CUIs) among all EHRs. In total, 570 sentences with mentions of the lifestyle related keywords were randomly selected to form a sentence-level corpus for the following NLP model development.

B. Annotation

Annotation guideline was built based on a randomly generated small subset of 570 sentences, and the annotation task was performed using INCEpTION. Three annotators independently annotated 50 sentences. The team compared the disagreement and resolved it through discussion until a consensus was reached. The inter-rater agreement was calculated over 50 sentences using Fleiss' kappa score revealing kappa of 0.72 (considered as a substantial agreement). Then, the remaining 520 sentences were annotated with the "Beginning-Inside-Outside" (BIO) annotation schema. The annotated entities

include excess diet, physical activity, substance abuse, and sleep deprivation. For each entity, we further defined several related entity types.

After annotation, the selected sentences were reorganized into four subsets based on their associated CUIs, which indicated different types of lifestyle exposures. The number of sentences for each category of excess diet, physical activity, sleep deprivation, and substance abuse was 80 (14%), 200 (35%), 97 (17%), and 193 (34%), respectively. In addition, the total number of annotated tags within the notes for each entity related to the above-mentioned four lifestyle exposures was 82 (19%), 100 (24%), 25 (6%), and 215 (51%), respectively.

C. Models and Feature Selections

We trained and evaluated six conventional machine learning algorithms, including support vector machine (SVM), conditional random field (CRF), logistic regression, random forest, bagged decision trees, and K-nearest Neighbors (KNN), and one deep learning model: bi-directional long short-term memory (Bi-LSTM) in Python. The CRF model was used as a baseline for the comparison purpose.

Pre-processing: Before training, we cleaned the sentences by removing punctuation and special characters such as '*'.

Feature sets: In total three kinds of feature sets were explored in 4 sub-datasets individually:

- ***n*-grams**: The feature set applied the bag-of-words representation method. Unigrams, bigrams, trigrams, and their combinations, such as unigrams+bigrams, bigrams+trigrams, and unigrams+bigrams+trigrams. In addition, term frequency-inverse document frequency (TF-IDF) was applied for adding weights on *n*-grams related features. Weights on TF-IDF was changed based on *n*-grams choices. Then, we selected the top 300 features based on their term frequency in the corpus.
- **context**: Tokens within the neighborhood window of the target token were considered as features. The sizes of the window were chosen as 3, 5, and 7.
- **combined**: we also combined the previous two types of features sets. For example: [window size 3+bigram]

Model details: For the NER task, we formulated the classification problem as a multi-class classification problem. Also, for tuning, each sub-data set's model (7 total) was tuned with GridSearch inside each training fold. In term of specific machine learning algorithms, we experimented with a range of hyperparameters, including but not limited to kernel-related parameters for SVM (i.e. gamma), penalty related parameters for logistic regression (i.e. penalty type), the number of neighbors for KNN, and tree-related parameters for the ensemble method (i.e. number of trees).

Besides the conventional machine learning models, we also applied Bi-LSTM to capture the contextual information among sentences. This model consisted of three types of layers: the embedding layer, the Bi-LSTM layer, and the time-distributed layer. The word embedding was obtained through training a Word2Vec model on the 570 sentences.

Evaluation: While evaluating the models, for each sub-dataset, a 5-fold CV was performed. The F1, precision, and recall scores were used as the evaluation metrics. All seven models with the feature sets followed the same procedure. Lastly, a model with the highest F1 score for the specific lifestyle category was determined.

III. Results

An individual sub-dataset for each lifestyle exposure was used for corresponding lifestyle exposure models.

For excessive diet, the feature set of window size of 7 was selected. The bagging model performed the best in F1, precision, and recall scores of 0.88, 0.83, and 0.94, respectively.

For physical activity, the final chosen feature set was window size 3. The random forest model outperformed the other six models in terms of F1 and recall scores. In regards to precision, the CRF model performed the best with 0.79.

Regarding sleep deprivation, the chosen feature set was unigrams+bigrams+trigrams. The KNN model performed the best in terms of F1 score. For precision, the Bi-LSTM delivered the best result. In regards to the highest recall score, logistic regression outperformed other models, achiving a score of 0.98.

For substance abuse, we selected tokens within a window size of 3 and bigrams. The random forest model achieved the highest F1 and recall scores of 0.77 and 0.80, respectively. For precision, the SVM model performed the best with 0.80.

IV. Conclusion

In this study, we developed and compared several NLP methods to automatically extract lifestyle exposure in clinical notes. We manually curated annotations for excess diet, physical activity, sleep deprivation, and substance abuse. The best methods were bagging, random forest, KNN, and random forest, respectively. Thus, this paper has demonstrated the feasibility of using NLP methods to automatically extract lifestyle exposures from EHRs. Overall, this paper will become a cornerstone of future exploration of understanding lifestyle exposures for AD with massive data and more advanced learning algorithms.

References

[1] NIH, "Alzheimer's disease fact sheet." [Online]. Available: https://www.nia.nih.gov/health/alzheimers-disease-fact-sheet

[2] K. S. Frederiksen, L. Gjerum, G. Waldemar, and S. G. Hasselbalch, "Physical activity as a moderator of alzheimer pathology: A systematic review of observational studies," *Current Alzheimer Research*, vol. 16, no. 4, p. 362–378, 2019.

[3] K. Dhana, D. A. Evans, K. B. Rajan, D. A. Bennett, and M. C. Morris, "Healthy lifestyle and the risk of alzheimer dementia: Findings from 2 longitudinal studies," *Neurology*, vol. 95, no. 4, pp. e374–e383, 2020.

[4] W. Zhu and N. Razavian, "Graph neural network on electronic health records for predicting alzheimer's disease," 2019.

[5] X. Zhou, H. Liu, and Y. Wang, "A comparison of lifestyle interventions for alzheimer's disease extracted from clinical notes and literature," *2018 IEEE International Conference on Healthcare Informatics Workshop (ICHI-W)*, 2018.

[6] X. Zhou, Y. Wang, S. Sohn, T. M. Therneau, H. Liu, and D. S. Knopman, "Automatic extraction and assessment of lifestyle exposures for alzheimer's disease using natural language processing," *International Journal of Medical Informatics*, vol. 130, p. 103943, 2019.