MEDINFO 2023 — The Future Is Accessible
J. Bichel-Findlay et al. (Eds.)

doi:10.3233/SHTI231038

Using Natural Language Processing to Extract and Classify Symptoms Among Patients with Thyroid Dysfunction

Sy HWANG[a], Sujatha REDDY[a], Katherine WAINWRIGHT[a], Emily SCHRIVER[a], Anne CAPPOLA[a] and Danielle MOWERY[a,1]
[a]*University of Pennsylvania, Philadelphia, PA, USA*
ORCiD ID: Sy Hwang https://orcid.org/0000-0003-3851-9521, Sujatha Reddy https://orcid.org/0000-0002-9474-0469, Katherine Wainwright https://orcid.org/0000-0003-1727-0341, Emily Schriver https://orcid.org/0000-0003-4522-1029, Anne Cappola https://orcid.org/0000-0002-0883-9754, Danielle Mowery https://orcid.org/0000-0003-3802-4457

Abstract. In the United States, more than 12% of the population will experience thyroid dysfunction. Patient symptoms often reported with thyroid dysfunction include fatigue and weight change. However, little is understood about the relationship between these symptoms documented in the outpatient setting and ordering patterns for thyroid testing among various patient groups by age and sex. We developed a natural language processing and deep learning pipeline to identify patient-reported outcomes of weight change and fatigue among patients with a thyroid stimulating hormone test. We built upon prior works by comparing 5 open-source, Bidirectional Encoder Representations from Transformers (BERT) to determine which models could accurately identify these symptoms from clinical texts. For both fatigue (f) and weight change (wc), Bio_ClinicalBERT achieved the highest F1-score (f: 0.900; wc: 0.906) compared BERT (f: 0.899; wc: 0.890), DistilBERT (f: 0.852; wc: 0.912), Biomedical RoBERTa (f: 0.864; wc: 0.904), and PubMedBERT (f: 0.882; wc: 0.892).

Keywords. Natural language processing, machine learning, electronic health records

1. Introduction

In the United States, an estimated 20 million Americans have thyroid dysfunction [1]. The symptoms of hypothyroidism and hyperthyroidism are myriad and nonspecific, particularly in older individuals and thus testing is required to distinguish thyroid disease from other etiologies [2]. Prominent symptoms reported by patients include fatigue and weight change, with weight loss from thyroid overactivity and weight gain from underactivity [3,4]. In order to understand the relationship between symptoms of fatigue and weight change in the outpatient setting and ordering patterns for thyroid testing in academic medical centers, we must be able to reliably identify documented symptoms in the electronic health record (EHR). However, in the EHR, symptoms are often under-coded using billing codes and documented in the clinical free-texts. Natural language processing

[1] Corresponding author: Danielle Mowery, dlmowery@pennmedicine.upenn.edu

and deep learning approaches can extract symptoms from a large corpus of clinical texts. Several open-source Bidirectional Encoder Representations from Transformers (BERT) frameworks have been trained on large corpora. However, BERT is largely understudied as a viable approach to symptom detection. Our study aims to determine how well each BERT framework identifies fatigue and weight change symptoms from clinical texts and how one can intuitively interpret the prediction by a BERT model.

2. Methods

This pilot study was approved by the University of Pennsylvania Institute Review Board. Clinical notes reviewed by the annotators were de-identified using De-ID [5] and PHIlter [6]. Co-authors SH and ES created a dataset for patient-reported outcomes extraction. We identified patients greater than 18 years of age who had at least one TSH test result and thyroid medication order between January 1, 2015 through December 31, 2019. We excluded any patients who had at least one or more of the following exclusionaries including thyroid medications, combination thyroid hormone therapy, T3 therapy, antithyroid medications, generic amiodarone, brand amiodarone, checkpoint inhibitors as well as diagnoses in their encounter of hypo function and other disorders of the pituitary gland, multiple myeloma, hypothyroidism, hyperthyroidism, thyroid cancer, or pregnancy. We randomly sampled 200 patient encounters and their associated outpatient clinical notes from family medicine and endocrinology generated in the visit preceding their first TSH test and thyroid medication order.

2.1. Annotating Patient-Reported Outcomes

Author AC led creation of an annotation scheme for encoding fatigue and weight change described in clinical notes. For fatigue, we created 2 classes to encode sentences from the text: **fatigue affirmed** reported fatigue experienced recently by the patient, and **not fatigue affirmed** reported fatigue as denied or absent, in the past, or reason associated with reported fatigue, e.g., insomnia. For weight change, we created 2 classes to encode sentences from the text: **weight change** perceived or actual increase or decrease in weight and **no weight change** reported failed attempt at increase or decrease in weight, unchanged weight, plans or desires for weight change, or reason associated with reported weight change, e.g., edema. Co-authors KW and SJ iteratively annotated 5 batches of clinical notes with consensus review to formulate the annotation schema. We report their final IAA using three metrics: Cohen's kappa, Krippendorf's Alpha, and F1-score.

2.2. Classifying Patient-Reported Outcomes

To determine how well open-source BERT frameworks identify thyroid dysfunction-related symptoms from clinical texts, we fine-tuned 5 different BERT-based language models (LMs) to predict these symptom classes.

- **BERT**[1] is a Transformer architecture-based model pretrained on the entire English Wikipedia dataset using two distinct self-supervised tasks, masked language modeling (MLM) and next sentence prediction (NSP).

- **DistilBERT**[2] uses an approximation technique called knowledge distillation in the pre-training phase to retain 97% of BERT's original capabilities while being less than half the size of the model [7].
- **Biomedical RoBERTa**[3] is an off-shoot of the RoBERTa model, which diverges from the original BERT model by foregoing the NSP task, using dynamically changing masking patterns in the pretraining and training on a larger dataset that includes the Books Corpus and Common Crawl Dataset. This specific model deploys a second phase of domain-adaptive pretraining continued from RoBERTa on 2.68 million scientific papers from Semantic Scholar [8].
- **PubMedBERT**[4] is a domain-specific BERT language model trained on PubMed ab- stracts from scientific articles, explicitly for the purpose of comparing them against continual pretraining of general-domain language models on a comprehensive set of biomedical NLP benchmarks.
- **Bio_Clinical BERT**[5] is a pre-trained BERT language model initialized from PubMed article abstracts and PubMed Central article full texts then tuned using a clinical corpus of notes from the Medical Information Mart for Intensive Care (MIMIC version III) dataset.

Author SH trained all classifiers on a 5-partition cross-validated dataset with stratified randomized folds. Hyperparameter optimization for the fine-tuning process was conducted with a Tree-structured Partizen Estimator sampling function and the F1-score as a primary evaluation metric using Optuna.

3. Results

We conducted an annotation and automation study of symptoms described by patients and documented by clinicians in clinical notes. We also applied a novel visualization approach for interpreting deep learning algorithms based on textual clues within the sentence.

3.1. Annotating Patient-Reported Outcomes

We assessed inter-annotator agreement. A1 review of A2 annotations resulted in IAA: Cohen's kappa=0.944, Krippendorf's Alpha=0.983, and F1 = 0.981. Conversely, A2 review of A1 annotations resulted in IAA: Cohen's kappa=0.953, Krippendorf's Alpha=0.992, and F1 = 0.964. Across both datasets, fatigue affirmed and no weight change were the most frequent observed classes.

3.2. Classifying Patient-Reported Outcomes

In **Table 1**, we observe the model performance on the test set for predicting a sentence as fatigue affirmed or weight change. Overall, we observed high performance across all BERT frameworks. For fatigue affirmed, the highest F1-score (as well as log loss, accuracy, and precision) was achieved by Bio_ClinicalBERT (F1: 0.900) followed by BERT (F1: 0.899). In terms of recall (R), BERT (R: 0.900) and Biomedical RoBERTa (R: 0.900) outperformed Bio_ClinicalBERT (R: 0.894). For weight change, the highest F1-score and all other measures was also achieved by Bio_ClinicalBERT (F1: 0.906) followed by Biomedical RoBERTa (F1: 0.904). Biomedical RoBERTa achieved equal recall to Bio_ClinicalBERT.

Table 1. Model performance by language model for fatigue affirmed and weight change.

Language Models	Log Loss	Accuracy	Precision	Recall	F1	AUROC
Fatigue Affirmed						
BERT	0.139	0.954	0.898	**0.900**	0.899	**0.935**
DistilBERT	0.180	0.916	0.894	0.736	0.727	0.852
Biomedical RoBERTa	0.172	0.935	0.832	**0.900**	0.864	0.923
PubMedBERT	0.131	0.947	0.890	0.879	0.882	0.923
Bio_ClinicalBERT	**0.130**	**0.956**	**0.908**	0.894	**0.900**	0.934
Weight Change						
BERT	0.215	0.918	0.863	0.919	0.890	0.919
DistilBERT	0.232	0.912	0.855	0.911	0.881	0.912
Biomedical RoBERTa	0.197	0.928	0.872	**0.941**	0.904	0.931
PubMedBERT	0.212	0.921	0.869	0.916	0.892	0.920
Bio_ClinicalBERT	**0.194**	**0.930**	**0.874**	**0.941**	**0.906**	**0.932**

In **Figure 1**, we visualize the word feature contribution using SHapley Additive exPlanations (SHAP) values as an indicator of force to either increase (pink) or decrease (blue) the prediction toward the position class (e.g., fatigue affirmed and weight change). For the subjective symptom expression "Pt complaining of feeling sleepy all day" indicating of **fatigue affirmed**, the pink terms "feeling" and "sleepy" provide a greater force for predicting fatigue affirmed compared to the other terms in the sentence. Conversely for the sentence "...Negative for chills, fever, malaise, fatigue and weight loss", terms for "weight" and "loss" as well as the negations term "negative" are strongly colored blue. In aggregate with other terms, this forces the prediction to **no weight change**.

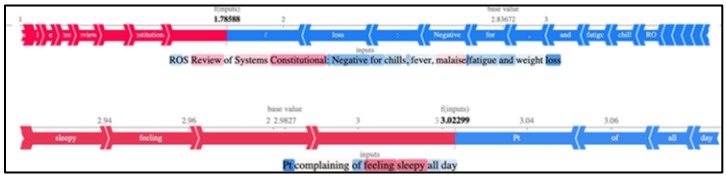

Figure 1. Sentence visualized using SHapley Additive exPlanations (SHAP) values for two example sentences.

4. Discussion

Unsurprisingly, Bio_ClinicalBERT outperformed the other BERT-based architectures for classifying statements of fatigue affirmed (F1: 0.900) and weight change (F1: 0.906). This demonstrates the added value of using domain-specific contextual embeddings that can accurately capture the clinical sublanguage used by clinical staff documenting information in the EHR. However, the performance by the remaining models was still comparable. We suspect that statements of fatigue affirmed and weight change are likely stated similarly in clinical texts, scientific literature, and non-clinical texts. Also, the number of statements using medical jargon were few. Compared to Bio_ClinicalBERT, the DistilBERT language model achieved reasonable performance for fatigue affirmed (F1: 0.727) and weight change (F1: 0.881) with less computational time, representing a reasonable baseline. Although, we focused on automation of only two thyroid dysfunction symptoms from our schema, this methodology could be applied broadly to assess symptoms of fatigue

and weight change in numerous conditions, including in cancer and specifically to assess patient response to thyroid dysfunction therapy.

5. Conclusions

We learned that patient-reported symptoms associated with thyroid dysfunction could be reliably annotated from clinical notes. State-of-the-art BERT models applied using a transfer learning approach can support classification of thyroid dysfunction symptoms from clinical text while maintaining model explainability.

Acknowledgement

This research was supported by the National Institute of Aging of the National Institutes of Health under Award Number K24-AG042765.

References

[1] American Thyroid Association. General Information/Press Room; 2022. Available from: https://www.thyroid.org/media-main/press-room.
[2] Trivalle C, Doucet J, Chassagne P, Landrin I, Kadri N, Menard JF, Bercoff E. Differences in the signs and symptoms of hyperthyroidism in older and younger patients. J Am Geriatr Soc. 1996 Jan;44(1):50-3, doi: 10.1111/j.1532-5415.
[3] Carlé A, Pedersen IB, Knudsen N, Perrild H, Ovesen L, Laurberg P. Hypothyroid symptoms and the likelihood of overt thyroid failure: a population-based case-control study. Eur J Endocrinol. 2014 Nov;171(5):593-602, doi: 10.1530/EJE-14-0481.
[4] Boelaert K, Torlinska B, Holder RL, Franklyn JA. Older subjects with hyperthyroidism present with a paucity of symptoms and signs: a large cross-sectional study. J Clin Endocrinol Metab. 2010 Jun;95(6):2715-26, doi: 10.1210/jc.2009-2495.
[5] Gupta D, Saul M, Gilbertson J. Evaluation of a deidentification (De-Id) software engine to share pathology reports and clinical documents for research. Am J Clin Pathol. 2004 Feb;121(2):176-86, doi: 10.1309/E6K3-3GBP-E5C2-7FYU.
[6] Norgeot B, Muenzen K, Peterson TA, Fan X, Glicksberg BS, Schenk G, Rutenberg E, Oskotsky B, Sirota M, Yazdany J, Schmajuk G, Ludwig D, Goldstein T, Butte AJ. Protected Health Information filter (Philter): accurately and securely de-identifying free-text clinical notes. NPJ Digit Med. 2020 Apr 14;3:57. doi: 10.1038/s41746-020-0258-y.
[7] Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108. 2019 Oct 2, doi: 10.48550/arXiv.1910.01108.
[8] Gururangan S, Marasović A, Swayamdipta S, Lo K, Beltagy I, Downey D, Smith NA. Don't stop pretraining: Adapt language models to domains and tasks. arXiv preprint arXiv:2004.10964. 2020 Apr 23, doi: 10.48550/arXiv.2004.10964.

[1] https://huggingface.co/bert-base-cased
[2] https://huggingface.co/distilbert-base-uncased
[3] https://huggingface.co/allenai/biomed_roberta_base
[4] https://huggingface.co/microsoft/PubMedBERT-base-uncased-abstract
[5] https://huggingface.co/emilyalsentzer/Bio_ClinicalBERT