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

doi:10.3233/SHTI231047

Development of a Natural Language Processing System to Identify Clinical Documentation of Electronic Cigarette Use

Patrick R. ALBA[a,b,1], Qiwei GAN[a,b], Mengke HU[a,b], Shu-Hong ZHU[c], Scott E. SHERMAN[d], Scott L. DUVALL[a],
Mike CONWAY[d]

[a] VA Informatics and Computing Infrastructure, VA Salt Lake City Health Care System, UT, USA
[b]Department of Internal Medicine Division of Epidemiology, University of Utah School of Me[1]dicine, Salt Lake City, UT, USA
[c]The Herbert Wertheim School of Public Health and Human Longevity Science
[d]Department of Population Health, New York University School of Medicine, NY, USA
[e]School of Computing & Information Systems, University of Melbourne, Parkville, VIC, Australia
ORCiD ID: Patrick R. Alba https://orcid.org/0000-0002-5176-5447

Abstract. Electronic Nicotine Delivery Systems (ENDS) use has increased substantially in the United States since 2010. To date, there is limited evidence regarding the nature and extent of ENDS documentation in the clinical note. In this work we investigate the effectiveness of different approaches to identify a patient's documented ENDS use. We report on the development and validation of a natural language processing system to identify patients with explicit documentation of ENDS using a large national cohort of patients at the United States Department of Veterans Affairs.

Keywords: Natural language processing, electronic cigarettes, public health, preventative medicine

1. Introduction

Electronic cigarettes – e-cigarettes, e-cigs, vapes, or Electronic Nicotine Delivery Systems (ENDS) – are now well-established products in most developed and developing countries [1]. In the United States, almost 15% of adults have used an ENDS device at least once [2]. The widespread use of ENDS devices continue to pose regulatory challenges to governments and health authorities due to uncertainties regarding both their long-term safety, and the role of ENDS in precipitating smoking initiation and nicotine addiction in children and young adults [3].

Given the potential public health and clinical significance of ENDS use, little is

[1] Corresponding Author: Patrick R. ALBA, email: patrick.alba@utah.edu

currently known regarding how clinicians document ENDS use in the Electronic Health Record (EHR), with existing evidence suggesting that ENDS use is substantially under-documented [1,4,5]. While there has been influential work conducted on the automated extraction of smoking status from EHRs using Natural Language Processing (NLP) algorithms work extending current smoking status detection methods to encompass ENDS use are less well developed [6].

In previous work we reported on the development of an ENDS use status annotation scheme and corpus derived from the Department of Veterans Affairs (VA) clinical notes [4]. In this paper, using this corpus and schema, we report on the development and evaluation of a high-performance, high-throughput NLP system capable of automating, at scale, the identification of ENDS use status at the VA.

2. Methods

Our initial cohort included 12 million VA patients who had utilized clinical services at least once between 2010 and June-2022, generating over 4 billion clinical documents. It is challenging to process all these documents with limited computing resources. To address this challenge, we implement a workflow shown in Figure 1 which tackles this problem in four main steps: 1) All clinical documents are stored in Microsoft SQL Server and are full text indexed. We first use full text index search with a manually developed list of key terms which largely covers smoking and ENDS. This reduces relevant cohort documents to about 45 million. 2) We then use a manually developed list of regular expressions of ENDS key terms to process all these 45 million documents, resulting in approximately 3.5 million documents containing ENDS key terms. 3) The system next extracts these ENDS key terms and their surrounding context. 4) Finally, multiple transformer models and classification algorithms were tested to classify the context into ENDS usage categories. All instances of an extracted term were classified into one of five categories described in previous work: Active-User, Usage-Unknown, Irrelevant, Former-User, and Non-User [4].

In the first and second step, all key terms associated with ENDS terminology act as the basis for the ENDS name entity recognition (NER) identified by the NLP system. The process of identifying relevant key terms can be found in previously published work and is also available in this project's git repository, this work also includes a detailed summary of the training and testing set selection process, which was largely influenced by the need to account for dramatic terminology shifts over time. Once the ENDS keywords are identified, the surrounding context is extracted for classification, including the previous sentence, the sentence in which the ENDS key word is identified, and the following sentence, if available.

In the classification step, we train and evaluate several machine learning models on training and validation dataset, respectively, including traditional machine learning-based algorithms (Logistics regression with TF-IDF vectors/ Word2Vector embeddings) and transformer-based NLP models (Clinical Bert). Although we did not perform extensive hyperparameter tuning on all models tested, we fine-tuned the transformer-based models with a learning rate of 10^{-4} and trained with 5 epochs. We selected the best model using F-score of the test dataset, and retrained the model on all training, validation, and test datasets for processing.

The end-to-end NLP pipeline (Figure 1) with the final ClinicalBert classification model was implemented with PySpark on a local machine with a CPU of 128 logical cores and 2 TeraByte memory which took approximately 4 days for extracting and inference.

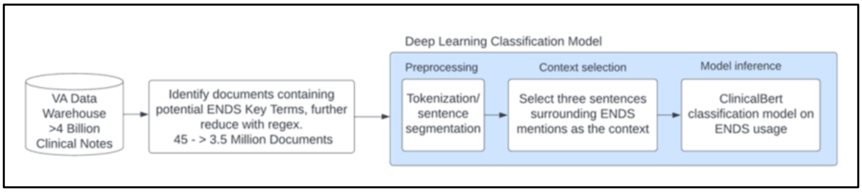
Deep Learning Classification Model
Preprocessing
Context selection
Model inference
VA Data Warehouse >4 Billion Clinical Notes
Identify documents containing potential ENDS Key Terms, further reduce with regex. 45 - > 3.5 Million Documents
Tokenization/ sentence segmentation
Select three sentences surrounding ENDS mentions as the context
ClinicalBert classification model on ENDS usage

Figure 1. End-to-end NLP System to Identify ENDS Usage from Clinical Documents.

3. Results

3.1. NLP Model Performance

In our experiments of classifying ENDS usage, the more traditional methods, such as logistic regression models provide a baseline performance of 0.82 (F-score). The transformer-based models outperformed the more traditional machine learning models, with the Bert for sequence classification model with embeddings pretrained on medical text (ClinicalBert) achieving the best overall performance of 0.89 on the validation dataset. We report performance on the final model, which was trained on the complete training and validation datasets and tested on test dataset. The performance of the final model is presented in Table 1 with the overall accuracy of the model as 0.90, weighted precision as 0.89, weighted recall as 0.90, weighted F1 score as 0.90, and an average AUC calculated at 0.962 (one vs. rest), and 0.936 (one vs. one).

Table 1. Performance of the final NLP pipeline for each category evaluated, and the total number of instances found in the testing set(support).

Categories	Recall	Precision	F1 Score	Support
Active User	0.94	0.91	0.93	408
Usage Unknown	0.94	0.93	0.93	327
Non-User	0.73	0.79	0.76	102
Irrelevant	0.75	0.86	0.80	32
Former User	0.50	0.70	0.58	32

4. Discussion

This work has resulted in the creation of a high-performance system capable of identifying ENDS use status of VA patients. Our system exhibited a particularly high performance in identifying individuals actively using ENDS products at the time of documentation (F-score 0.93), a key requirement for both future clinical decision support efforts and retrospective epidemiological analysis.

Although recent research has indicated that the application of transformer-based NLP methods to clinical text does not always yield better performance than traditional machine learning-based NLP methods [7], the limited comparisons done in this work found that transformer models outperformed traditional machine learning methods for this concept.

The research presented here does include several limitations. While the system showed excellent performance identifying Active ENDS users from all others, our random sampling method resulted in little training and testing data for the former and non-smoking classifications made by this system. Meaning the system can be used to accurately identify patients with affirmed ENDS documentation, but further work is needed to better identify past and non-users. Second, it is important to note that the NLP system depends on clinical documentation of a patient's ENDS use, little is still known about the extent to which clinicians actively document a patient's ENDS use, and what data does exist suggests substantial heterogeneity in clinical practice [8]. Therefore, any subsequent work studying patients identified by this system will need to account for the possible underrepresentation of actual e-cigarette usage. Future work to study these possible issues with sensitivity could be done by comparing to cohorts with existing survey data. Future work may also improve upon the results achieved here by testing an even more specific transformer lexicon/model, as many of the current errors could be due to terminology missing from the model vocabulary.

With this study we sought to investigate the best methods, and ultimately implement a system to identify ENDS use in the EHR. We have since applied this algorithm to process all notes for our cohort of patients which will be described in more detail in future work.

5. Conclusions

This work demonstrates the feasibility of developing a high-performance, high-throughput NLP system capable of automating, at scale, the identification of ENDS use status at the VA.

Acknowledgements

Research reported in this publication was partially supported by the National Institute on Drug Abuse of the National Institutes of Health under award number R03DA047577 and received further support in terms of resources and facilities from the Department of Veterans Affairs (VA) Informatics and Computing Infrastructure under award number VA HSR RES 13-457. The views expressed in this article are those of the authors and do not necessarily reflect the position or policy of the National Institutes of Health, the Department of Veterans Affairs, or the United States government.

References

[1] Young-Wolff KC, Klebaner D, Folck B, Tan ASL, Fogelberg R, Sarovar V, Prochaska JJ. Documentation of e-cigarette use and associations with smoking from 2012 to 2015 in an integrated healthcare delivery system. Prev Med. 2018 Apr;109:113-8, doi: 10.1016/j.ypmed.2018.01.012.

[2] Villarroel MA, Cha AE, Vahratian A. Electronic cigarette use among U.S. adults, 2018. NCHS Data Brief. 2020 Apr;(365):1-8.

[3] Balfour DJK, Benowitz NL, Colby SM, Hatsukami DK, Lando HA, Leischow SJ, Lerman C, Mermelstein RJ, Niaura R, Perkins KA, Pomerleau OF, Rigotti NA, Swan GE, Warner KE, West R. Balancing consideration of the risks and benefits of e-cigarettes. Am J Public Health. 2021 Sep;111(9):1661-72, doi: 10.2105/AJPH.2021.306416.

[4] Conway M, Alba PR, Zhu SH, Patterson OV. Vaping at the VA: developing an annotated corpus of electronic cigarette mentions in clinical notes at the department of veterans affairs. AMIA Annu Symp Proc. 2022 Feb;2021:343-51.

[5] Winden TJ, Chen ES, Wang Y, Sarkar IN, Carter EW, Melton GB. Towards the standardized documentation of e-cigarette use in the electronic health record for population health surveillance and research. AMIA Jt Summits Transl Sci Proc. 2015 Mar;2015:199-203.

[6] Young-Wolff KC, Klebaner D, Folck B, Carter-Harris L, Salloum RG, Prochaska JJ, Fogelberg R, Tan ASL. Do you vape? leveraging electronic health records to assess clinician documentation of electronic nicotine delivery system use among adolescents and adults. Prev Med. 2017 Dec;105:32-6, doi: 10.1016/j.ypmed.2017.08.009.

[7] Gao S, Alawad M, Young MT, Gounley J, Schaefferkoetter N, Yoon HJ, Wu XC, Durbin EB, Doherty J, Stroup A, Coyle L, Tourassi G. Limitations of transformers on clinical text classification. IEEE J Biomed Health Inform. 2021 Sep;25(9):3596-607, doi: 10.1109/JBHI.2021.3062322.

[8] Hurst S, Conway M. Exploring physician attitudes regarding electronic documentation of e-cigarette use: a qualitative study. Tob Use Insights. 2018 Jul;11:1179173X18782879, doi: 10.1177/1179173X18782879.