Resource Summary Report

Generated by NIF on May 17, 2025

Smoking NLP Challenge Data

RRID:SCR_008644
Type: Tool

Proper Citation

Smoking NLP Challenge Data (RRID:SCR_008644)

Resource Information

URL: https://www.i2b2.org/NLP/DataSets/Main.php

Proper Citation: Smoking NLP Challenge Data (RRID:SCR_008644)

Description: The data for the smoking challenge consisted exclusively of discharge summaries from Partners HealthCare which were preprocessed and converted into XML format, and separated into training and test sets. I2B2 is a data warehouse containing clinical data on over 150k patients, including outpatient DX, lab results, medications, and inpatient procedures. ETL processes authored to pull data from EMR and finance systems Institutional review boards of Partners HealthCare approved the challenge and the data preparation process. The data were annotated by pulmonologists and classified patients into Past Smokers, Current Smokers, Smokers, Non-smokers, and unknown. Second-hand smokers were considered non-smokers. Other institutions involved include Massachusetts Institute of Technology, and the State University of New York at Albany. i2b2 is a passionate advocate for the potential of existing clinical information to yield insights that can directly impact healthcare improvement. In our many use cases (Driving Biology Projects) it has become increasingly obvious that the value locked in unstructured text is essential to the success of our mission. In order to enhance the ability of natural language processing (NLP) tools to prise increasingly fine grained information from clinical records, i2b2 has previously provided sets of fully deidentified notes from the Research Patient Data Repository at Partners HealthCare for a series of NLP Challenges organized by Dr. Ozlem Uzuner. We are pleased to now make those notes available to the community for general research purposes. At this time we are releasing the notes (~1,000) from the first i2b2 Challenge as i2b2 NLP Research Data Set #1. A similar set of notes from the Second i2b2 Challenge will be released on the one year anniversary of that Challenge (November, 2010).

Synonyms: NLP Data Set #1C

Resource Type: data or information resource, database

Keywords: nlp datasets

Funding:

Resource Name: Smoking NLP Challenge Data

Resource ID: SCR_008644

Alternate IDs: nif-0000-32739

Record Creation Time: 20220129T080248+0000

Record Last Update: 20250507T060630+0000

Ratings and Alerts

No rating or validation information has been found for Smoking NLP Challenge Data.

No alerts have been found for Smoking NLP Challenge Data.

Data and Source Information

Source: SciCrunch Registry

Usage and Citation Metrics

We found 14 mentions in open access literature.

Listed below are recent publications. The full list is available at NIF.

Chen F, et al. (2024) Examining the Generalizability of Pretrained De-identification Transformer Models on Narrative Nursing Notes. Applied clinical informatics, 15(2), 357.

Dada A, et al. (2024) Information extraction from weakly structured radiological reports with natural language queries. European radiology, 34(1), 330.

Skreta M, et al. (2021) Automatically disambiguating medical acronyms with ontology-aware deep learning. Nature communications, 12(1), 5319.

Chen Q, et al. (2021) Specialists, Scientists, and Sentiments: Word2Vec and Doc2Vec in Analysis of Scientific and Medical Texts. SN computer science, 2(5), 414.

Zhang Z, et al. (2021) Combining data augmentation and domain information with TENER model for Clinical Event Detection. BMC medical informatics and decision making, 21(Suppl

9), 261.

Grabar N, et al. (2020) CAS: corpus of clinical cases in French. Journal of biomedical semantics, 11(1), 7.

Shen Z, et al. (2019) A Lightweight API-Based Approach for Building Flexible Clinical NLP Systems. Journal of healthcare engineering, 2019, 3435609.

Balyan R, et al. (2019) Using natural language processing and machine learning to classify health literacy from secure messages: The ECLIPPSE study. PloS one, 14(2), e0212488.

Palmer EL, et al. (2019) Building a tobacco user registry by extracting multiple smoking behaviors from clinical notes. BMC medical informatics and decision making, 19(1), 141.

Li Z, et al. (2019) Integrating shortest dependency path and sentence sequence into a deep learning framework for relation extraction in clinical text. BMC medical informatics and decision making, 19(Suppl 1), 22.

Reátegui R, et al. (2018) Comparison of MetaMap and cTAKES for entity extraction in clinical notes. BMC medical informatics and decision making, 18(Suppl 3), 74.

Lee HJ, et al. (2018) Identifying direct temporal relations between time and events from clinical notes. BMC medical informatics and decision making, 18(Suppl 2), 49.

Liu Z, et al. (2017) Entity recognition from clinical texts via recurrent neural network. BMC medical informatics and decision making, 17(Suppl 2), 67.

Oellrich A, et al. (2015) Generation of silver standard concept annotations from biomedical texts with special relevance to phenotypes. PloS one, 10(1), e0116040.