Medical Subdomain Classification Using Unstructured Clinical Documents and Machine Learning-Based Natural Language Processing Approach

HARVARD MEDICAL SCHOOL
MASSACHUSETTS GENERAL HOSPITAL

Wei-Hung Weng, MD[1,2*], Kavishwar B. Wagholikar, MBBS, PhD[2,3], and Henry C. Chueh, MD, MS[2,3]

[1] Department of Biomedical Informatics, Harvard Medical School, Boston, MA
[2] Laboratory of Computer Science, [3]Department of Medicine, [4]Department of Neurology, Massachusetts General Hospital, Boston, MA

Abstract

The medical subdomain of a clinical document is useful metadata for developing machine learning models. To classify the medical subdomain of a document accurately, we have constructed an automated machine learning-based natural language processing (NLP) pipeline and developed accurate medical subdomain classifiers based on the content of the document. We constructed the pipeline using cTAKES, the UMLS Metathesaurus and Semantic Network, and learning algorithms to extract features from two datasets — documents from Integrating Data for Analysis, Anonymization, and Sharing (iDASH) data repository (n = 431) and clinical notes from Massachusetts General Hospital (MGH) (n = 91,237), and built medical subdomain classifiers with different combinations of clinical feature representations and learning algorithms. The performance of classifiers and model portability across two datasets were evaluated. The linear support vector machine algorithm-trained medical subdomain classifier using hybrid bag-of-words and clinically relevant UMLS concepts as the feature representation, with term frequency-inverse document frequency-weighting, outperformed other classifiers on iDASH and MGH datasets with F1 scores of 0.932 and 0.934, and areas under ROC curve (AUC) of 0.957 and 0.964, respectively. Classifiers for half of medical subdomains were found to be portable across two datasets at the threshold of F1 score of 0.7. Our study shows that with the machine learning-based NLP approach, it is possible to develop medical subdomain classifiers having sufficient performance for clinical applications. Portable medical subdomain classifiers may also be used across datasets from different institutes.

Background

Motivation

- The medical subdomain, such as cardiology, gastroenterology and neurology, is metadata of a document that is necessary to enhance the effectiveness of clinical machine learning-based NLP models by considering specialty-associated conditions
- Unstructured clinical documents have been regarded as a powerful resource to solve different clinical questions by providing detailed patient conditions, the thinking process of clinical reasoning, and clinical inference, which usually cannot be obtained from the other components of the electronic health record (EHR) system
- Automated clinical document classification is an essential component of clinical predictive analytics that can extract knowledge and categorize documents into predefined document-level thematic labels

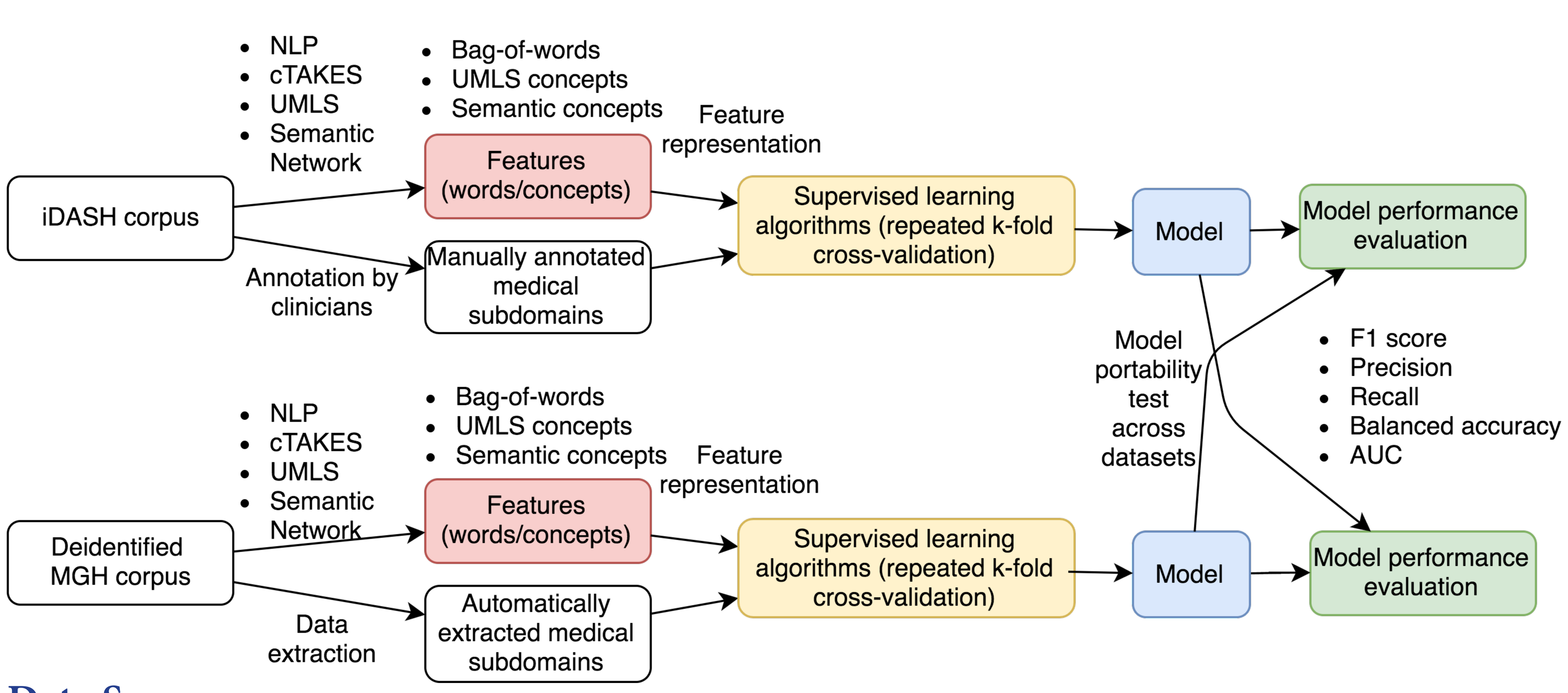
Current / Proposed Approach

- Automated document classification task was usually performed via rule-based knowledge engineering, by manually implementing a set of expert intelligence rules
- Machine learning-based NLP approach -- NLP-derived feature representation and supervised learning algorithm using pre-classified documents
- MEDLINE documents
  - Hybrid word and phrase representation with a support vector machine (SVM) algorithm
  - The Medical Subject Headings (MeSH) ontology as a feature representation with a maximum entropy algorithm
- Chest radiograph reports
  - Medical Language Extraction and Encoding System (MedLEE) with UMLS Metathesaurus to identify medical concepts and classify into six clinical conditions

Research Objectives

- Developed an automated machine learning-based NLP pipeline to build subdomain classifiers
- Examined the performance between the classifiers using different clinical feature representation methods, weighting strategies, and supervised learning algorithms
- Investigated the portability of classifiers across two real-world clinical datasets

Materials and Methods



Data Source

- 431 clinical documents from the iDASH repository + 91237 MGH clinical notes from Partners HealthCare RPDR (IRB:2016P000011)

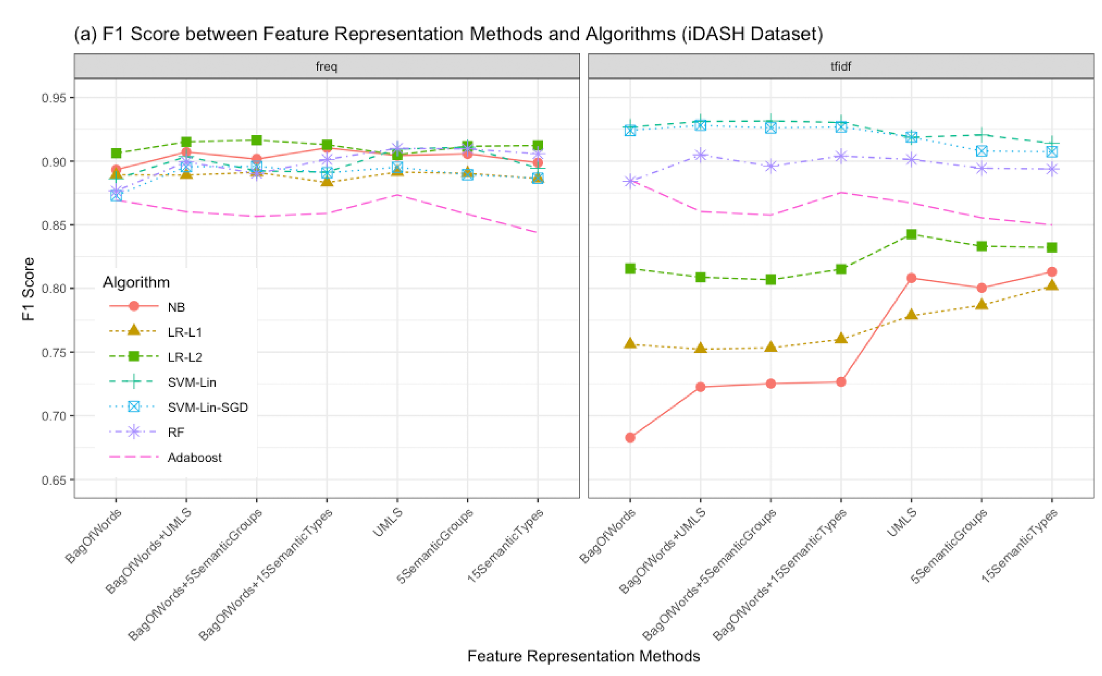
NLP and Machine Learning

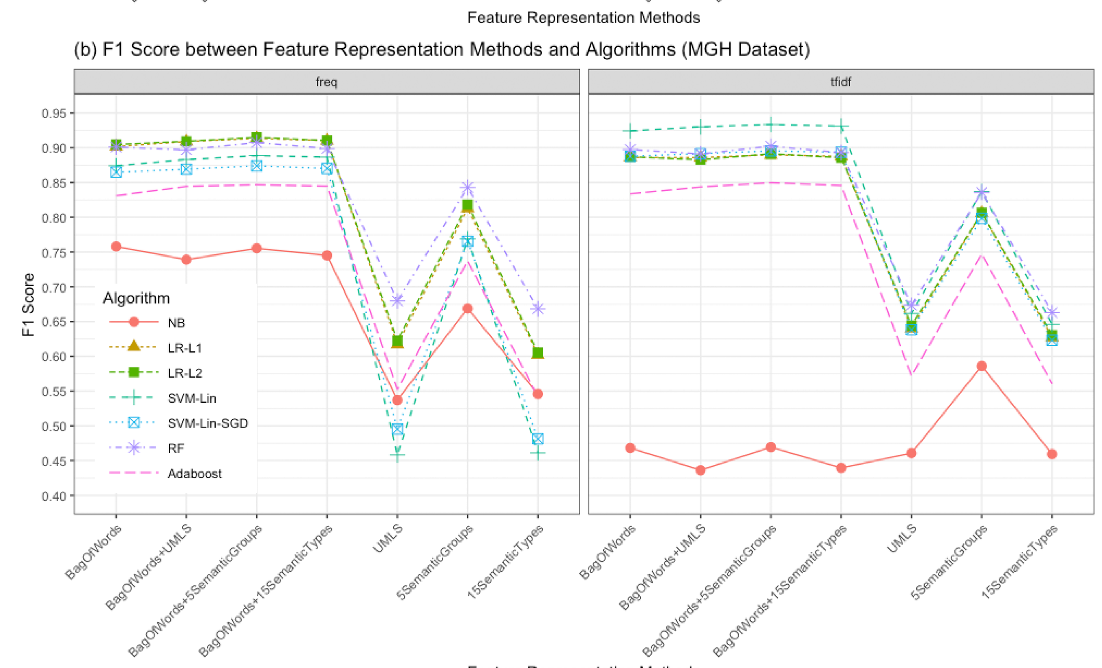
- Using cTAKES, UMLS Metathesaurus and Semantic Network. [1-3]
- 98 classifiers for each dataset
- Seven feature representation methods
  - Bag-of-Words / SNOMED concepts / UMLS concepts / Concepts restricted by semantic groups or types / words-concepts combination
- Two vector representation methods
  - Frequency count / tf-idf
- Seven supervised learning algorithms
  - Naive Bayes, multinomial logistic regression (L1-, L2 penalization), linear SVM (with/without SGD), random forest, adaptive boosting
- Accuracy, precision, recall, F1, AUC

TUI	Semantic group	Semantic type description
T017	Anatomy	Anatomical Structure
T022	Anatomy	Body System
T023	Anatomy	Body Part, Organ, or Organ Component
T033	Disorders	Finding
T034	Phenomena	Laboratory or Test Result
T047	Disorders	Disease or Syndrome
T048	Disorders	Mental or Behavioral Dysfunction
T049	Disorders	Cell or Molecular Dysfunction
T059	Procedures	Laboratory Procedure
T060	Procedures	Diagnostic Procedure
T061	Procedures	Therapeutic or Preventive Procedure
T121	Chemicals & Drugs	Pharmacologic Substance
T122	Chemicals & Drugs	Biomedical or Dental Material
T123	Chemicals & Drugs	Biologically Active Substance
T184	Disorders	Sign or Symptom

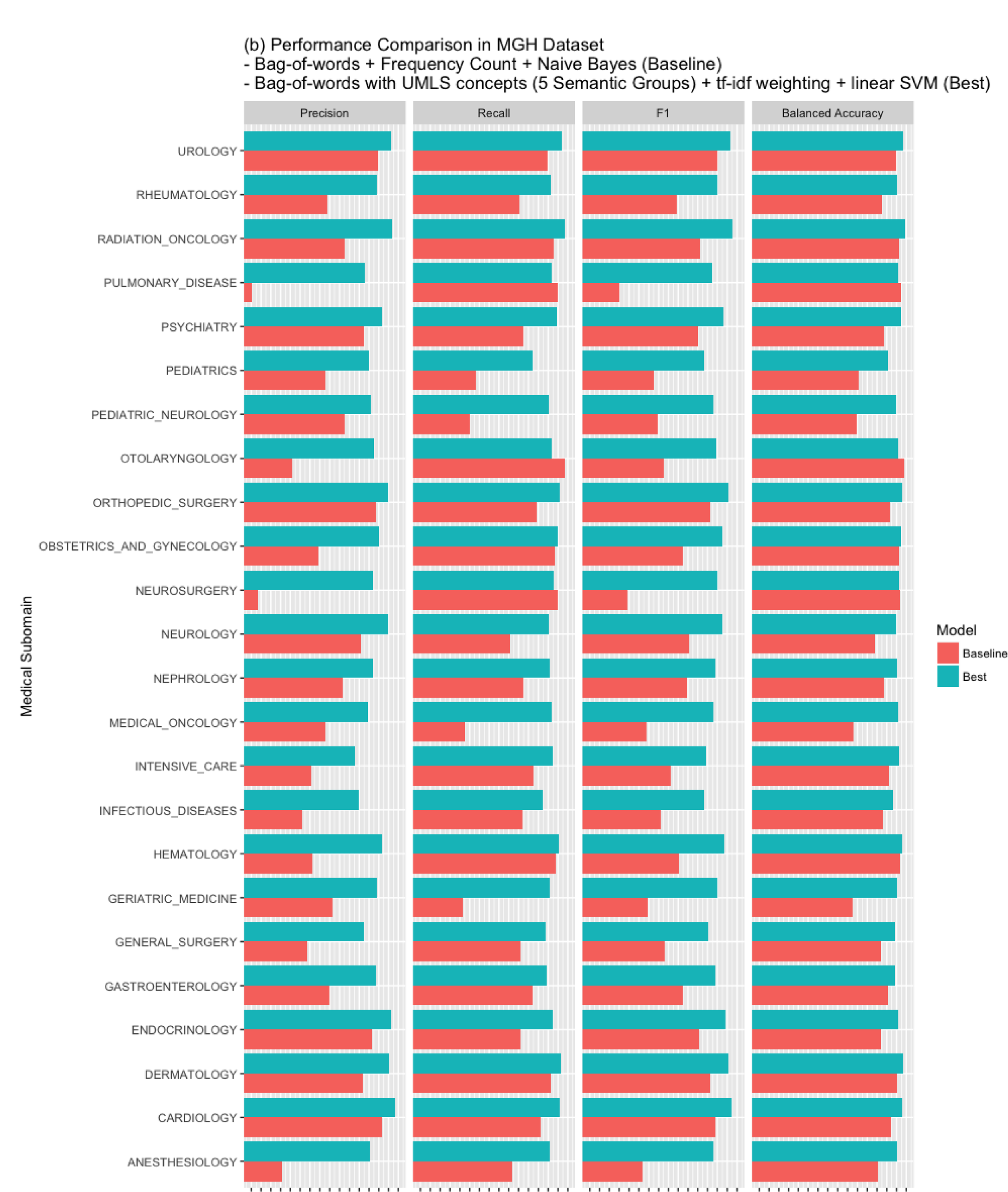
Result

Dimension of feature space	iDASH	MGH
Bag-of-words (Vocabulary size)	10150	160097
UMLS concepts	4750	25456
UMLS concepts restricted to five semantic groups	4531	24457
UMLS concepts restricted to 15 semantic types	3634	18520
Bag-of-words + UMLS concepts	14900	185553
Bag-of-words + UMLS concepts restricted to five semantic groups	14681	184554
Bag-of-words + UMLS concepts restricted to 15 semantic types	13784	161949





GASTROENTEROLOGY	
c0014876 (Esophagus) c1278919 (Entire esophagus) murmur vomit mucosa c0009378 gallbladd pancrea distal moder sever transfer c1278925 (Entire cecum) c0038351 (Stomach) c0007531 (Cecum) c0009368 (Colon structure (body structure)) portion duodenum rectum duct	ibd le precancer coliti c0009378 (colonoscopy) abduct ppi relax rheum methocarbamol c1457887 (Symptoms) c0231377 (At risk for impaired home maintenance management) hypercholesterolemia hcc sptrg thiim esophagu c0021853 (Intestines) formalin c0392916 (Intracellular ferritin) cmd
PSYCHIATRY	
time orient c0438696 (Suicidal) development hospit contact c0033975 (Psychotic Disorders) c0018524 (Hallucinations) need feel affect c0344315 (Depressed mood) famili data thought seroquel laboratori patient bipolar deferred.axi live father	psychopharmacolog citalopram c0004457 (Axis vertebra) retrain director unabl lorazepam licsw registr haldol suicid lexapro report memori wish c0033573 (Psychotic Disorders) nasosept card psychiatr c0442967 (Salvage procedure) psych



From iDASH to MGH					From MGH to iDASH				
Subdomain	AUC	Precision	Recall	F1	Subdomain	AUC	Precision	Recall	F1
Cardiology	0.828	0.923	0.715	0.806	Cardiology	0.731	0.829	0.500	0.624
Gastroenterology	0.802	0.396	0.691	0.503	Gastroenterology	0.832	1.000	0.664	0.798
Neurology	0.877	0.745	0.859	0.798	Neurology	0.775	0.902	0.567	0.696
Psychiatry	0.803	0.907	0.613	0.732	Psychiatry	0.941	0.794	0.900	0.844
Pulmonary	0.820	0.197	0.692	0.307	Pulmonary	0.545	1.000	0.089	0.164
Nephrology	0.770	0.573	0.561	0.567	Nephrology	0.634	0.750	0.273	0.400

Model Performance

- Algorithm selection plays important role
  - SVM with linear kernel and regularized logistic regression performed better
- Clinical feature selection method has a great impact
  - Words-concepts combination yielded the best performance
  - Semantic groups-derived concepts augmented the performance of bag-of-words model
  - Bag-of-words features were required for better modeling
- Tf-idf weighting adjustment improves the performance with some supervised learning algorithms.
- The best-performing models improves the performance of almost all subdomain classifiers

Top Features

- Words and UMLS concepts
- Different in two datasets
- Some meaningless features may due to noise fitting
- Extracting the top features improves the interpretability of machine learning classifiers

Model Portability

- The iDASH model has better portability comparing to the MGH model
- Document heterogeniety > document size

Next Steps

- More features (n-grams) and combination
- Preserving sequential information (e.g. recurrent neural network)
- Transfer learning for model portability

Conclusion

The purpose of the study is to classify the medical subdomain of an unstructured or semi-structured clinical document accurately, and we proved that the machine learning-based NLP approach could be an optimal solution for building portable medical subdomain classifiers for clinical documents. Using two sets of clinical notes, we found that the selection of the classifier-building combination of the clinical feature representation and supervised learning algorithm is important to yield a better-performed and portable medical subdomain classifier. We plan to integrate the information of both medical subdomain and clinical expert to build hierarchical models for consolidating our methods, and may adopt techniques of domain adaptation and transfer learning to improve the performance of model portability and construct a well-generalizable solution.

Reference

[1] Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. Journal of the American Medical Informatics Association. 2010 Sep;17(5):507–13.
[2] Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Research. 2004 Jan 1;32(90001):267D–270.
[3] McCray AT. An Upper-Level Ontology for the Biomedical Domain. Comparative and Functional Genomics. 2003;4(1):80–4.