Cohort selection in clinical trials using enriched word-vector features with SVM and CNN models

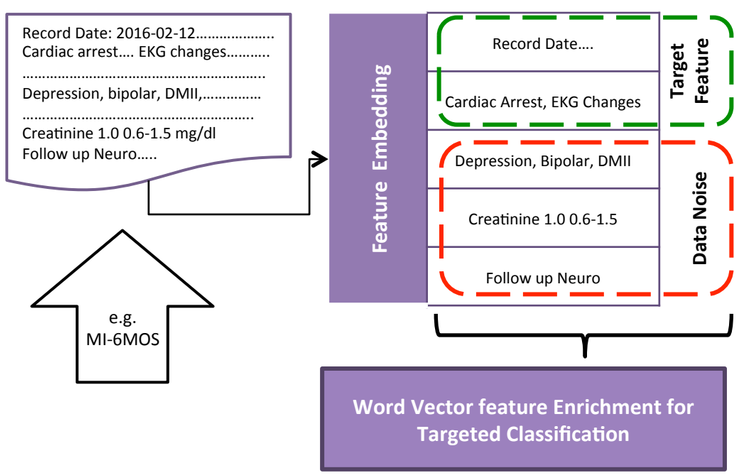
Chi-Jen Chen[1], Neha Warikoo[2,3], Yung-Chun Chang[1], Hsyien-Chia Wen[4], and Jin-Hua Chen[1]

[1]Graduate Institute of Data Science, Taipei Medical University, Taipei 106, Taiwan,
[2]Institute of Biomedical Informatics, National Yang-Ming University, Taipei 115, Taiwan
[3]Bioinformatics Program, Taiwan International Graduate Program, Institute of Information Science, Academia Sinica, Taipei 115, Taiwan
[4]School of Heal Care Administration, Taipei Medical University, Taipei, Taiwan.

I. Abstract

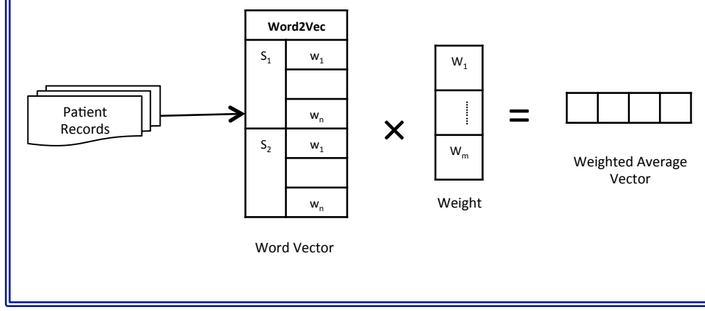
- Motivation: Admittance in clinical trial requires meeting specific medical criteria on part of patients, easily deducible via mining patient EHRs.
- Objective: Determining the status of a checklist of 13 criteria from a set of clinical out patient records. Cohort selection is targeted towards identifying cardiovascular and diabetes risk factors.
- Method: Our learning model focuses on feature enrichment via selective representation. Ranking metrics viz. log likelihood and trigger word similarity are employed to isolate noisy data.

II. Challenges with Noisy Representation

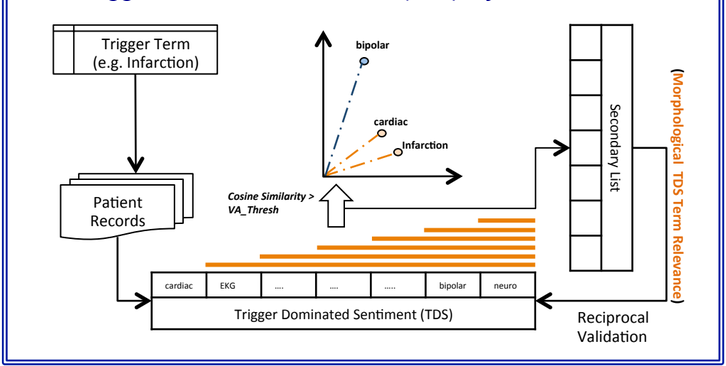


III. Feature Enrichment

Log Likelihood Ratio (LLR) weighted Embedding

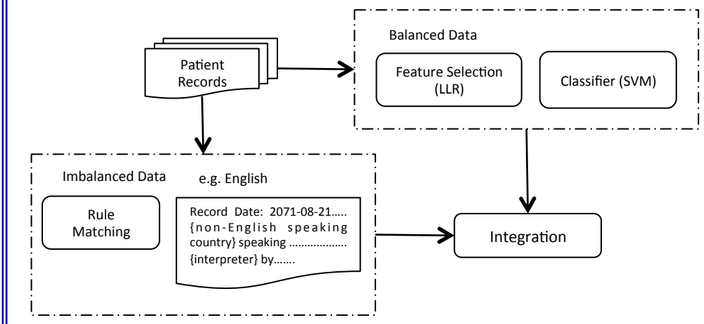


Trigger Dominated Sentiment (TDS) refined Features

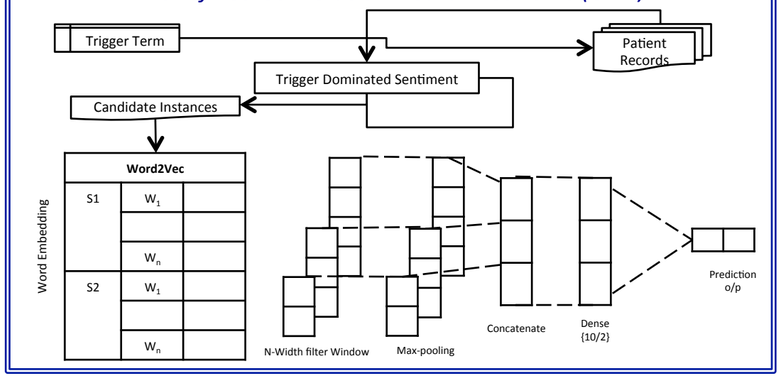


IV. System Workflow

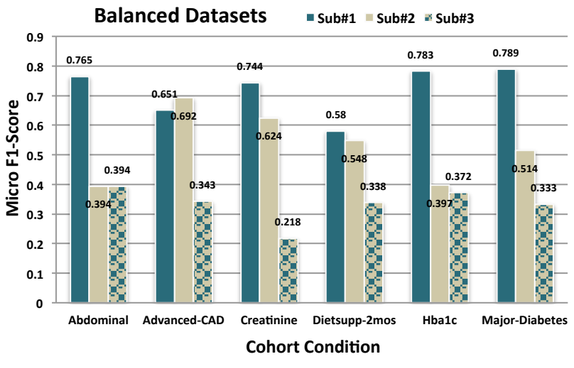
LLR weighted Support Vector Machines (SVM)

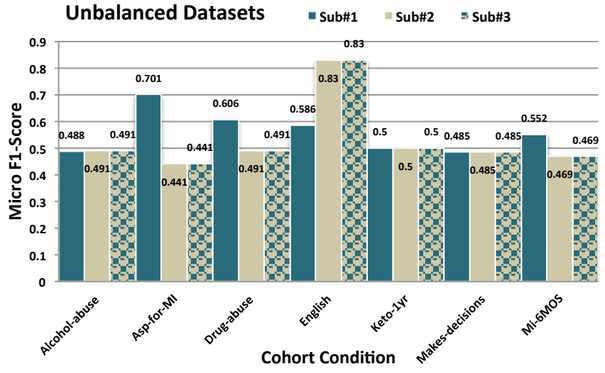


TDS refined Convolution Neural Networks (CNN)



V. Experiments & Results





Overall System performance

Metrics / Run#	*Micro*	*Macro*
Sub # 1 (TDS)	0.8192	0.6334
Sub # 2 (LLR)	0.7850	0.5293
Sub # 3 (LLR + Rules)	0.7326	0.4391

VI. Discussion

- Reciprocal validation subjected Trigger terms are consistent in retaining criteria specific information across all 13 categories.
- Heuristic patterns define better representations for certain criteria like 'English' and 'Drug-Abuse'. E.g. If word 'speaking' is present in sentence, then screen for country origin information.
- Co-occurrence of cross category terms like 'drugs, alcohol' impacts true symptom identification in selective classes.

VII. Acknowledgment

Ministry of Science and Technology of Taiwan: MOST 107-2634-F-001-005, MOST 107-2319-B-400-001, MOST 107-2410-H-038-017-MY3, MOST 106-2218-E-038-004-MY2

Presented at 2018 National NLP Clinical Challenges (n2c2) Workshop @ American Medical Informatics Association Fall Symposium, San Francisco, USA