



HAL
open science

Contrastive Clustering of medical reports to characterise Adverse Drug Reactions

Xuchun Zhang, Michel Riveill, Milou-Daniel Drici

► To cite this version:

Xuchun Zhang, Michel Riveill, Milou-Daniel Drici. Contrastive Clustering of medical reports to characterise Adverse Drug Reactions. Statlearn 2023 workshop : Challenging problems in Statistical Learning, Apr 2023, Montpellier, France. . hal-04282263

HAL Id: hal-04282263

<https://hal.science/hal-04282263v1>

Submitted on 13 Nov 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License

Xuchun ZHANG¹, Michel RIVEILL,¹ Milou-Daniel. DRICI,²

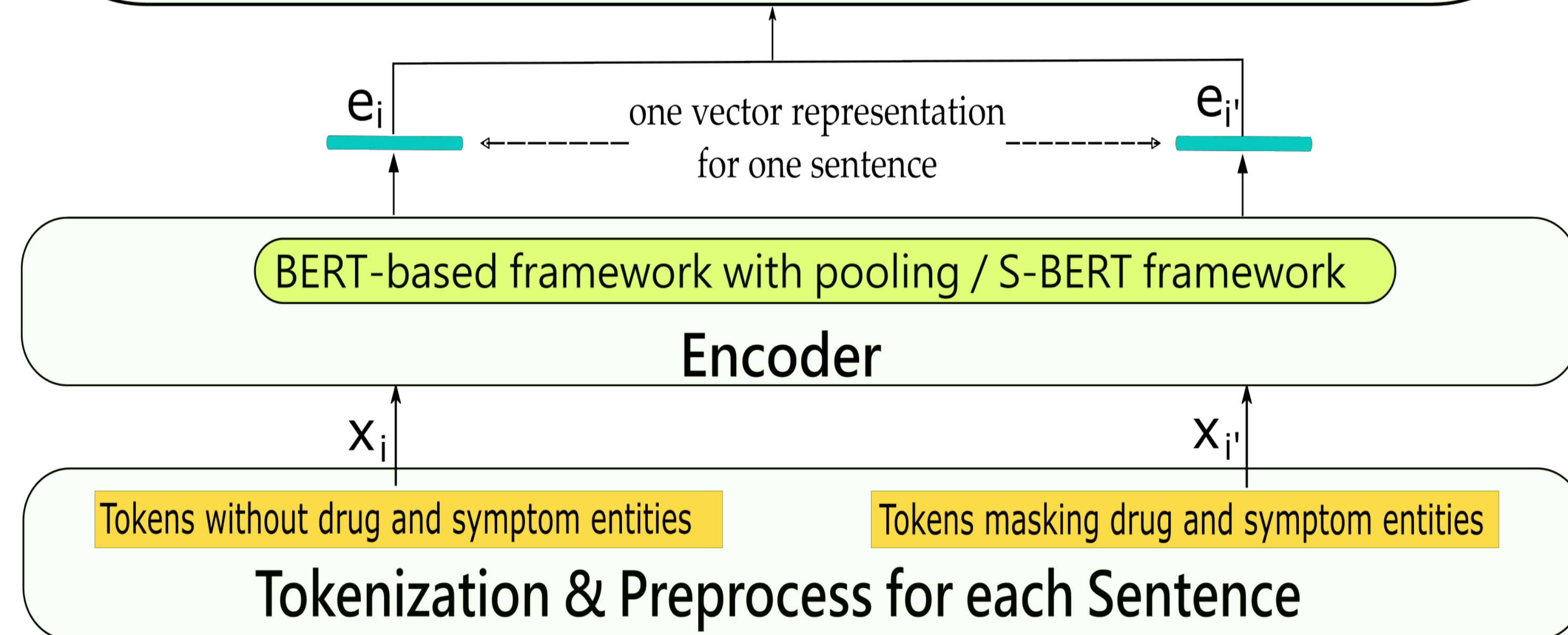
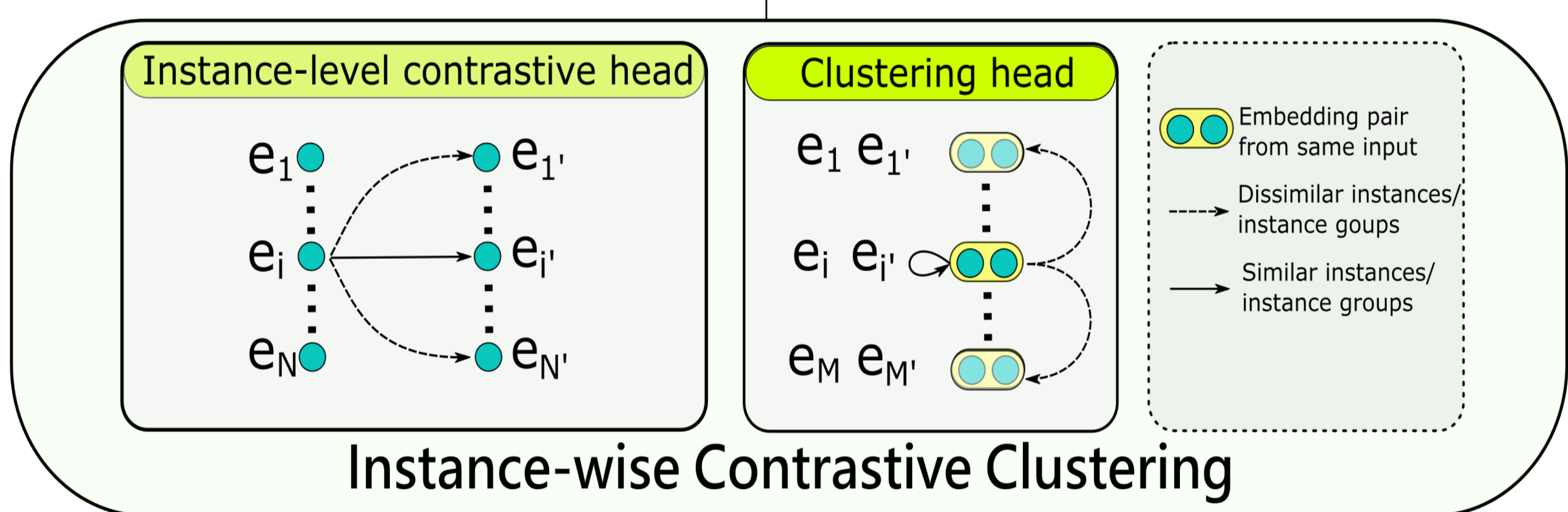
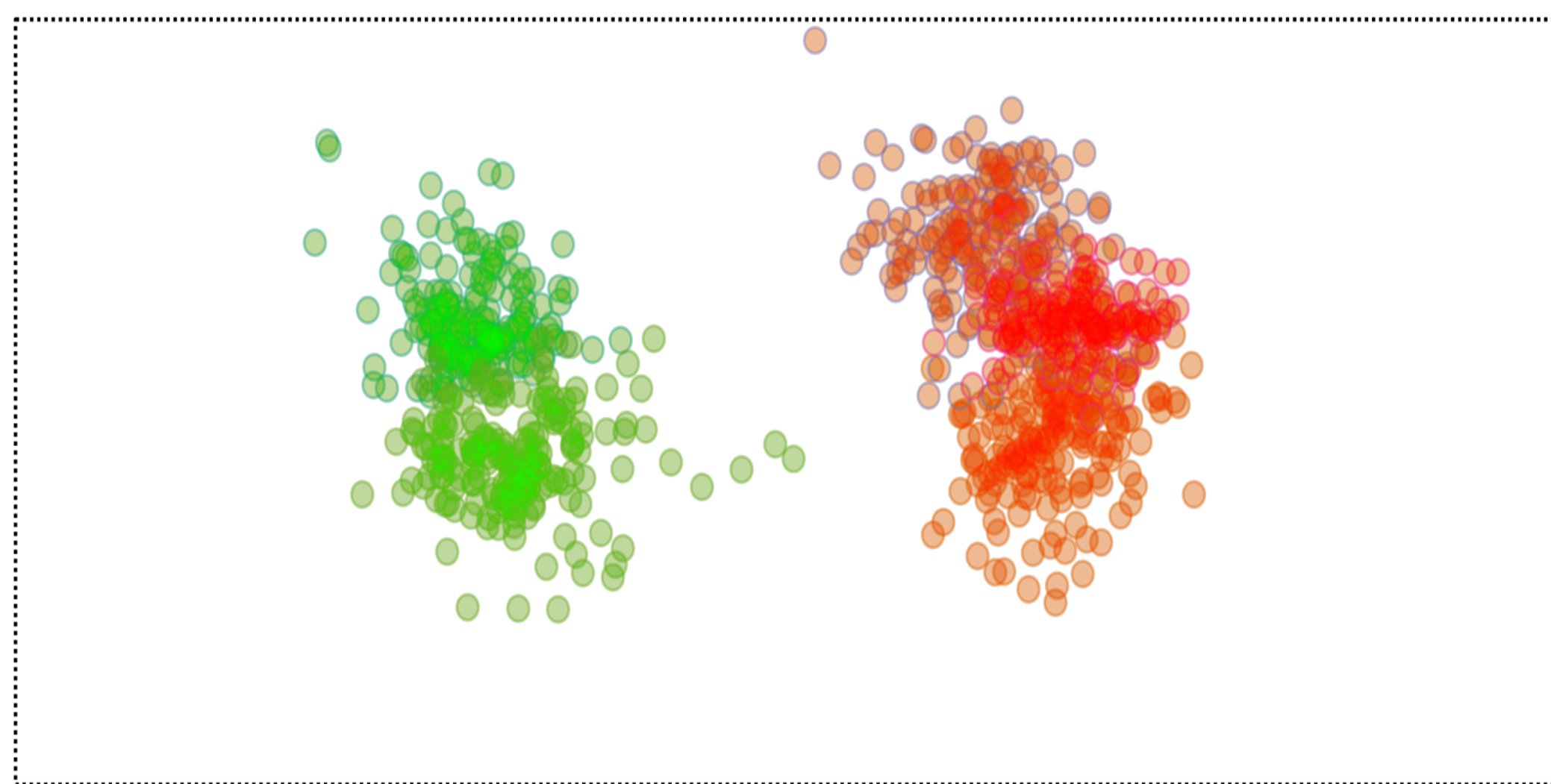
1: Université Côte d'Azur, CNRS, INRIA, I3S Laboratory, France {xuchun.zhang, michel.riveill}@inria.fr
2: Université Côte d'Azur, CNRS, CHU Nice, France drici.md@chu-nice.fr

1. Motivation

- Adverse Drug Reactions (ADRs) being mandatory for healthcare professionals, pharmacovigilance still suffers from a significant under-reporting, accounting for only 5-10% of all ADRs.
- Identification of ADRs relies on the well-trained health professionals, while there are still an enormous amount of documents waiting to be reviewed.
- Annotating such electronic health records (EHRs) is very expensive, yet supervised approaches need a large amount of annotated data.

2. Structure of Model Pipeline

- Assumption: "the ADR relation lies in the contexts of the drug-symptom target entities". More specifically, we assume there exists a short text around the entities pair which is sufficient for the representation of ADR.
- Our goal is to separate the ADR-related text blocks (noted as positive blocks β^+) from non-related ones (noted as negative blocks β^-).



He was better controlled on **<Drug>Velcade</Drug>**, but developed significant **<Symptom>peripheral neuropathy</Symptom>**. ADR

... he attributes to the use of **<Drug>vasoconstrictors</Drug>** while he was in the Intensive Care Unit. non-ADR

- BioBERT** model is pre-trained and fine-tuned on biomedical corpora instead of employing general domain text corpora in BERT. Both BERT and BioBERT are token-wise language model.
- Sentence-BERT** applies a siamese fine-tuning structure on basic BERT model to capture features in semantic space, which transform a whole sentence directly into vector representation.

4. Contrastive Learning Enhanced Clustering (CLEC)

- For the Instance-level Contrastive head, the model is trained to distinguish between similar and dissimilar data by evaluating the loss¹

$$\ell_i^C = -\log \frac{\exp(\text{sim}(x_i, x_{i'})/\tau)}{\sum_{j=1}^M \mathbb{1}_{j \neq i'} \exp(\text{sim}(x_i, x_j)/\tau)}$$

The Instance Contrastive Loss is $\mathcal{L}_{IC} = \sum_{i=1}^M \ell_i^C / M$

- For the clustering head, we use the Student's t -distribution² to measure the similarity between mapped point z_i and centroid μ_k :

$$q_{ik} = \frac{(1 + \|z_i - \mu_k\|^2 / \alpha)^{-\frac{\alpha+1}{2}}}{\sum_{k'} (1 + \|z_i - \mu_{k'}\|^2 / \alpha)^{-\frac{\alpha+1}{2}}}$$

and refine the clusters with the help of an auxiliary target distribution³ $p_{ik} = \frac{q_{ik}^2 / f_k}{\sum_{k'} q_{ik}^2 / f_{k'}}$ with $f_k = \sum_i q_{ik}$ as soft cluster frequencies,

The objective is a KL divergence loss between the soft assignments q_i and the auxiliary distribution for both instance pairs:

$$\mathcal{L}_C = \sum_{(i,i')} \ell_i^C / M + \ell_{i'}^C / M, \text{ where } \ell_i^C = \sum_j \sum_k p_{ik} \log \frac{p_{ik}}{q_{ik}}$$

- Total objective of CLEC: $\mathcal{L}_{CLEC} = \mathcal{L}_{IC} + \eta \mathcal{L}_C$.

5. Results and Discussion

Experimental Data

- Data source from MADE challenge 2018, whose corpora are only electronic health records.
- The information about medications and symptoms are given.
- MADE 1d1s** where we extract those who has exactly one drug and one symptom as the "perfect situation", a nearly balanced dataset with short textual blocks.
- MADE multi-d-s** has longer blocks of the EHR corpus with an almost balanced distribution between negative and positive blocks. Each block contains at least one drug and one symptom.
- In BioBERT based model, we use average pooling for short text input in **MADE 1d1s** and took $[CLS]$ representation for long inputs in **MADE multi-d-s**

We have chosen a fully supervised approach (Bag of Words + Logistic Regression Classifier) as the upper bound baseline, and a BoW with random guessing as the lower bound.

The results for unsupervised approaches (*) are always followed by a KMedoids clustering

Category	Exp	MADE 1d1s		
		Prec	Recall	F1
Supervised	BOW+LR	0.702	0.797	0.746
Unsupervised	BioBERT* ⁴	0.651	0.663	0.650
	SBERT* ⁴	0.733	0.615	0.665
Unsupervised	BioBERT+CLEL	0.768	0.669	0.715
	SBERT+CLEL	0.646	0.711	0.677
Supervised	BOW+Random	0.514	0.529	0.522

Category	Exp	MADE multi-d-s		
		Prec	Recall	F1
Supervised	BOW+LR	0.809	0.847	0.827
Unsupervised	BioBERT* ⁴	0.604	0.634	0.619
	SBERT* ⁴	0.653	0.673	0.663
Unsupervised	BioBERT+CLEL	0.582	0.827	0.673
	SBERT+CLEL	0.660	0.773	0.702
Supervised	BOW+Random	0.509	0.440	0.472

- Both BioBERT-based model and SBert-based Model benefit the contrastive clustering comparing to our previous work.

- For short EHR data, BioBERT based model benefits more from its biomedical domain specified dictionary comparing to

the SBERT-based model,

- SBert model has more advantage for long textual EHR data,

References

- He, Kaiming et al. "Momentum Contrast for Unsupervised Visual Representation Learning." 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019): 9726-9735.
- Maaten, Laurens van der and Geoffrey E. Hinton. "Visualizing Data using t-SNE." Journal of Machine Learning Research 9 (2008): 2579-2605.
- Xie, Junyuan et al. "Unsupervised Deep Embedding for Clustering Analysis." ArXiv abs/1511.06335 (2015): n. pag.
- X. Zhang, M.-D. Drici, M. Riveill. "Unsupervised Text Clustering to characterize Adverse Drug Reactions from hospitalization reports". ASPAI 2022