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

doi:10.3233/SHTI231232

Automated Feature Selection from Medical Literature

Alberto PURPURA[a,1], Tobia BOSCHI[a], Francesca BONIN[a], Rodrigo ORDONEZ-HURTADO[a], Natasha MULLIGAN[a], Joao H. BETTENCOURT-SILVA[a] and Alessandra PASCALE[a]
[a] *IBM Research Europe – Dublin*

Abstract. We propose an automated approach to rank the most salient variables related to a certain clinical phenomenon from scientific literature. Our solution is an automated approach to improve the efficiency of the collection of different health-related measures from a population, and to accelerate the discovery of novel associations and dependencies between health-related concepts.

Keywords. Digital health, natural language processing, feature selection

1. Introduction

We describe an automated solution, grounded on statistical methods and Natural Language Processing (NLP) to discover the most significant variables related to a certain phenomenon, and suggest them as new targets of a clinical study or survey. We focus on Quality of Life (QoL) as the target phenomenon for the presentation of our approach, which can be also used to discover any other type of associations.

2. Methods

We consider over 17M PubMed abstracts published between 2000 and 2022, annotate them with Unified Medical Language System (UMLS) entities using ScispaCy [1] and store them in an Elasticsearch index – the annotation process took about 10 days. Next, we compute different association scores between UMLS entities representing QoL concepts that we manually extracted from both the UMLS metathesaurus[2] and each of the entities recognized by ScispaCy. The measures we compute to represent the Strength of Association (SoA) between a given entity *e* and QoL concepts are: (i) the Association Rule Confidence Score (ARCS) i.e., the ratio between the number of times an entity *e* co-occurs with any QoL-related entity and the total number of times QoL entities found in our index; (ii) the Unnormalized Co-occurrence Frequency (UCF) of an entity *e* with any of the UMLS entities defining a QoL concept; (iii) the Semantic Association Score

[1] Corresponding Author: Alberto Purpura, IBM Research Europe – Dublin, email: alp@ibm.com

[2] The manually selected UMLS identifiers of QoL entities are C0518214, C0034380, C2015891, C0281588, C0451401, C4049260, C5548092, C3476057, C0451498, C1963786, C3686815, C4534509, C5387806, C4279947, C3874813, C5545434, C2058096, C3836984, C3837215, C3837014, C5539750, C3482969, C3171917, C0518906 and C4331194.

(SAS), obtained with the rule-based approach SemRep [2]; (iv) the Logarithm of the Normalized Frequency (LNF) of entity *e* over the total number of sentences in our database; and (v) a Binary Association Score (BAS) (0 or 1) indicating whether entity *e* is semantically related to any of the selected QoL-related entities in the UMLS semantic network. These metrics are then aggregated to form a single score using a Linear Regression model. To train our regression model, we employ the SHARE dataset [3], and we focus our analysis on 1518 patients. For each patient, we retrieve 14 variables and use the CASP12 index to measure their QoL. To gauge the association between QoL and the other variables in the SHARE data, we regress CASP12 indexes against each of them and compute the coefficient of determination of marginal regressions. We then use these coefficients to train our regression model.

3. Results

The Mean Absolute Error (MAE) of our regression model that we obtain in a cross-validation setting (leave-one-out) is 0.04. Among the top 10 UMLS entities most related to QoL discovered by our system we observed UMLS concepts related to depression, sleep disturbances, physical activity, and gastrointestinal problems (e.g., C0086132, C0011570, C0872084, C0028734). These features – many of which were absent in the SHARE data – give some insights on what aspects of an individual's life are more important to track in a study focused on QoL. This information, combined with the domain knowledge of medical professionals on the context of a certain medical study, can help the design of better targeted experiments and optimize the data collection process by narrowing down the choices of items to track.

4. Conclusions

We described a solution to discover the most statistically relevant factors affecting the QoL of a population from medical literature. Our approach produced novel relevant insights – absent from its training data – to study the wellbeing of a person.

Acknowledgements

This work has been funded through the EU project SEURO grant no. 945449.

References

[1] Neumann M, King D, Beltagy I, Ammar W, editors. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. Proceedings of the 18th BioNLP Workshop and Shared Task; 2019 Aug; Florence, Italy; p.319-27, doi: 10.18653/v1/W19-5034.

[2] Kilicoglu H, Rosemblat G, Fiszman M, Shin D. Broad-coverage biomedical relation extraction with SemRep. BMC Bioinformatics. 2020; 21:1-28, doi:10.1186/s12859-020-3517-7.

[3] Börsch-Supan A, Brandt M, Hunkler C, Kneip T, Korbmacher J, Malter F, Schaan B, Stuck S, Zuber S. Data resource profile: the Survey of Health, Ageing and Retirement in Europe (SHARE). Int J Epidemiol. 2013 Aug; 42(4):992-1001, doi:10.1093/ije/dyt088.