INTRODUCTION

Despite the recent digitalization of patient information by the use of millions of electronic health records (EHR), some of the most important information remains hidden to machines in unstructured clinical notes and reports. Natural language processing (NLP) can extract unstructured information to enable precision medicine and predict epidemic trends through big data approaches. However, the rise of new conditions, such as COVID-19, frequently leads to the rise of unstandardized medical terms that are challenging to extract for the current NLP pipelines. Our NLP/Machine Learning (ML) enabled system improves the identification and extraction of medical concepts by integrating human-provided feedback into a confidence score. This allows users to give directed feedback on the data correctness.

Apache cTAKES

cTAKES (Clinical Text Analysis & Knowledge Extraction System) analyzes unstructured electronic health records and extracts medical concepts and health information.

NLP Analysis

NLP analytics are used to evaluate precision of matching between concepts in the clinical notes and in cTAKES medical database.

In text: Coronavirus

- Concept: SARS-CoV-2
- BLEU Score: 0
- Levenstein Similarity: 0.5
- Jaccard Similarity: 0.12
- Cosine Similarity: 0.76
- EIMo: 0.82

SARS-CoV-2

Range Text: Coronavirus

Concept Ontology: SARS-CoV-2

Location: null

Negation: False

History of: False

Is the concept correct?

- Yes
- No

Is the location correct?

- Yes
- No

User Feedback

Users provide feedback (‘yes’ or ‘no’) to extracted clinical concepts and the respective modifier words through UI. Feedback is used as a label to train an ML model along with NLP the metrics.