

# Human-Enabled Healthcare NLP

## Discovering Inaccurate Extracted Information from Unstructured Clinical Notes Using Natural Language Processing

Team Members: Chloe Kim, Layne Wei, Marcel Schaack  
Advised by: Dr. Gundolf Schenk, Dr. Gabriel Gomes,  
Dr. Angela Rizk-Jackson, Eugenia Rutenberg



## 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 un/structured 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

HEALTH • COVID-19

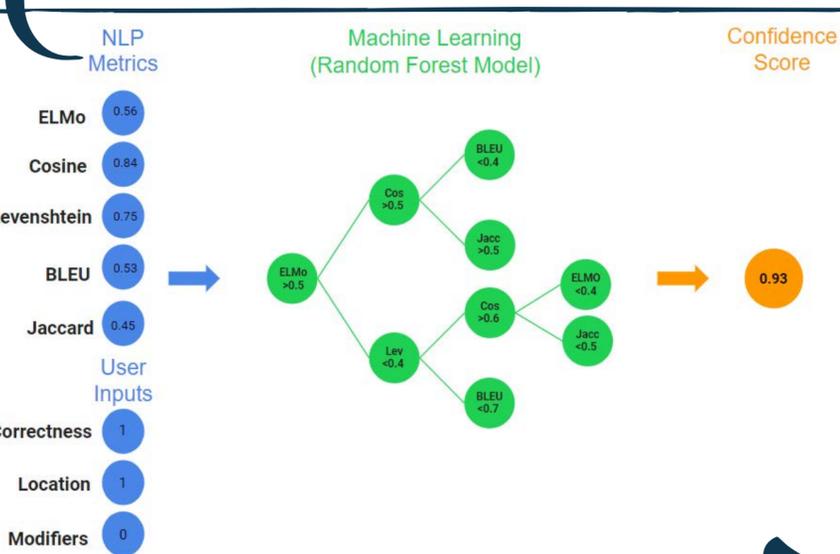
**Coronavirus** Researchers Are Using High-Tech Methods to Predict Where the Virus Might Go

Natural language processing (NLP) is one tool used by BlueDot to track diseases with the company being successful in detecting diseases around the globe. For instance, BlueDot analyzes human languages around the world and use the information to assist them forecast disease outbreaks.

**cTAKES**  
In Text: Coronavirus  
Concept Ontology: SARS-CoV-2

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

[Go back](#)

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



## RESULTS

We collected **over 1500** feedback datapoints from users, and used this, together with the computed NLP metrics to build a model that identifies inaccurate extracted information at a **96% accuracy** (97% F1). Our **human-in-the-loop** system can utilize user-provided feedback to self-improve through **active learning**. It, thus, represents a superior method to collect and use feedback data and can effectively be used to increase the reliability of medical NLP pipelines.

### Possible Text Sources:



Journals



Agencies



News



EHRs

### ML Prediction

Trained random forest model predicts the correctness of the extracted concepts and computes reliable confidence scores which are used to improve cTAKES identification

### 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