



Citation Recognition in Legal Documents

Named Entity Recognition with Transformer-Based Deep Learning Methods



Definition of Task

This bachelor thesis develops a system for recognizing citations in legal documents. It is carried out in collaboration with Swisslex, an online research platform for legal documents.

In order to link documents in their web application, citations and their specific parts need to be identified in the text.

The current citation recognition system is based on regular expression matching. This approach is not robust and is hard to maintain.

Approach

1 Recognizing Citations

Test F1 Score: 95.7%

Regex solution



~ 21 Mio samples
3 classes (LAW, CASELAW, LITERATURE)



google-bert/bert-base-multilingual-uncased

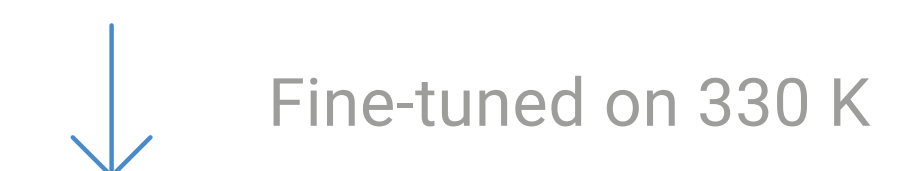
2 Recognizing Parts of a Citation

Test F1 Score: 98.9%

meta-llama/Llama-3.3-70B-Instruct



~ 380 K synthetic samples
21 classes (COURT, DATE, CONSIDERATION, PAGE, etc.)



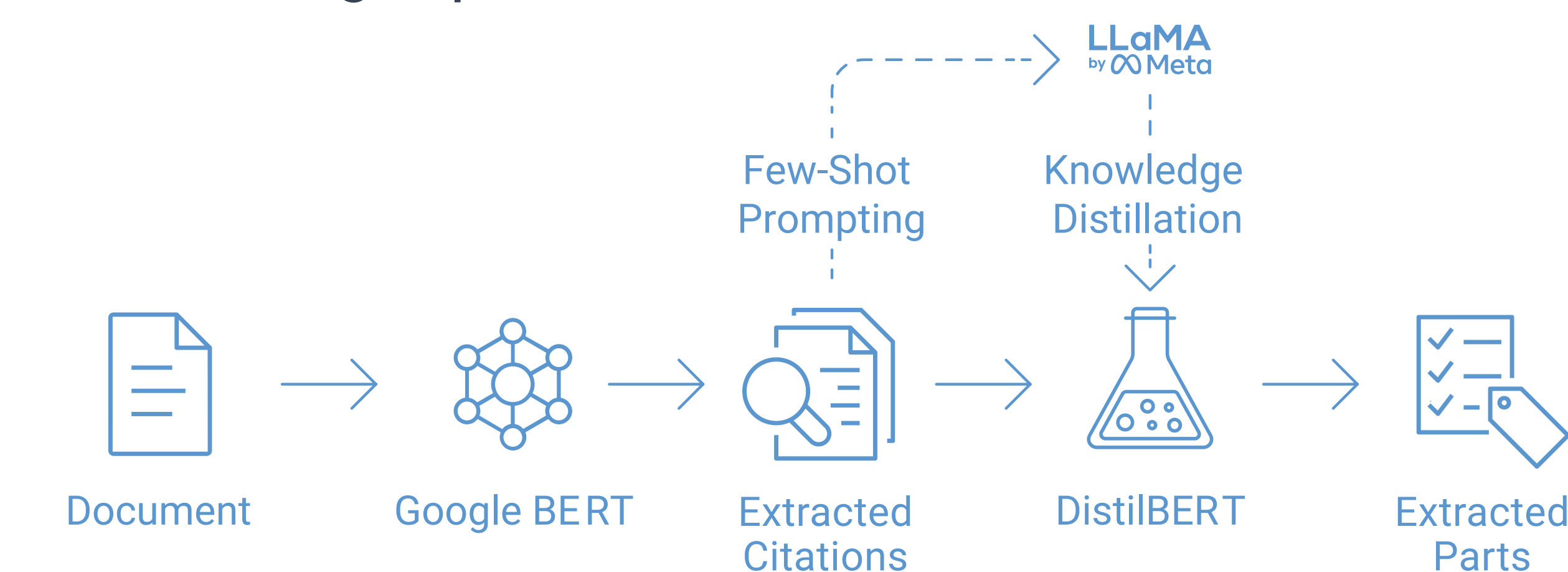
distilbert/distilbert-base-uncased

Knowledge distillation

Technology Stack



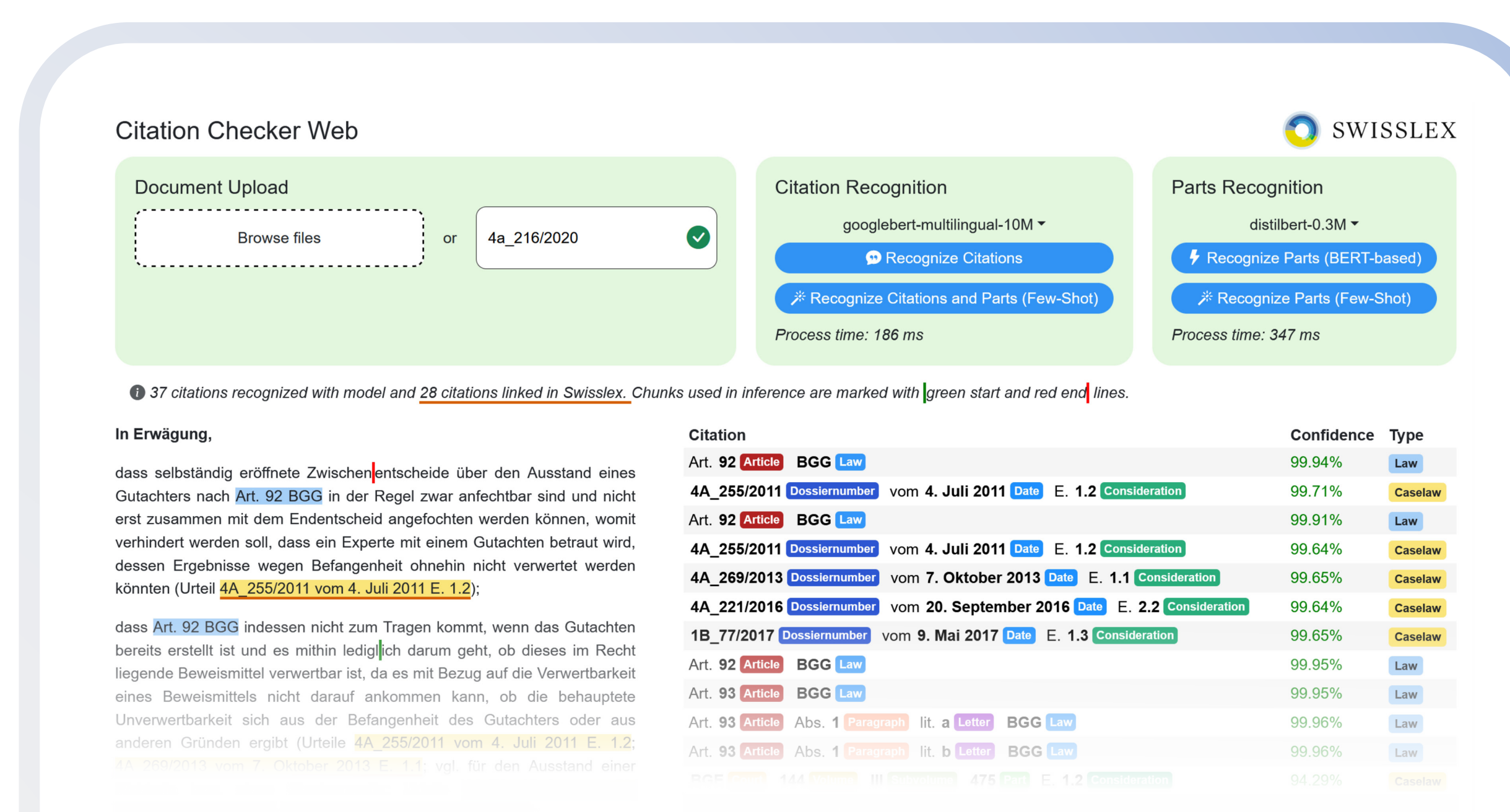
Processing Pipeline



Solution

- Web application for testing and exploration of citation recognition
- Console application for bulk processing documents
- Inference API to serve fine-tuned models

Web-Application to Test Citation Recognition

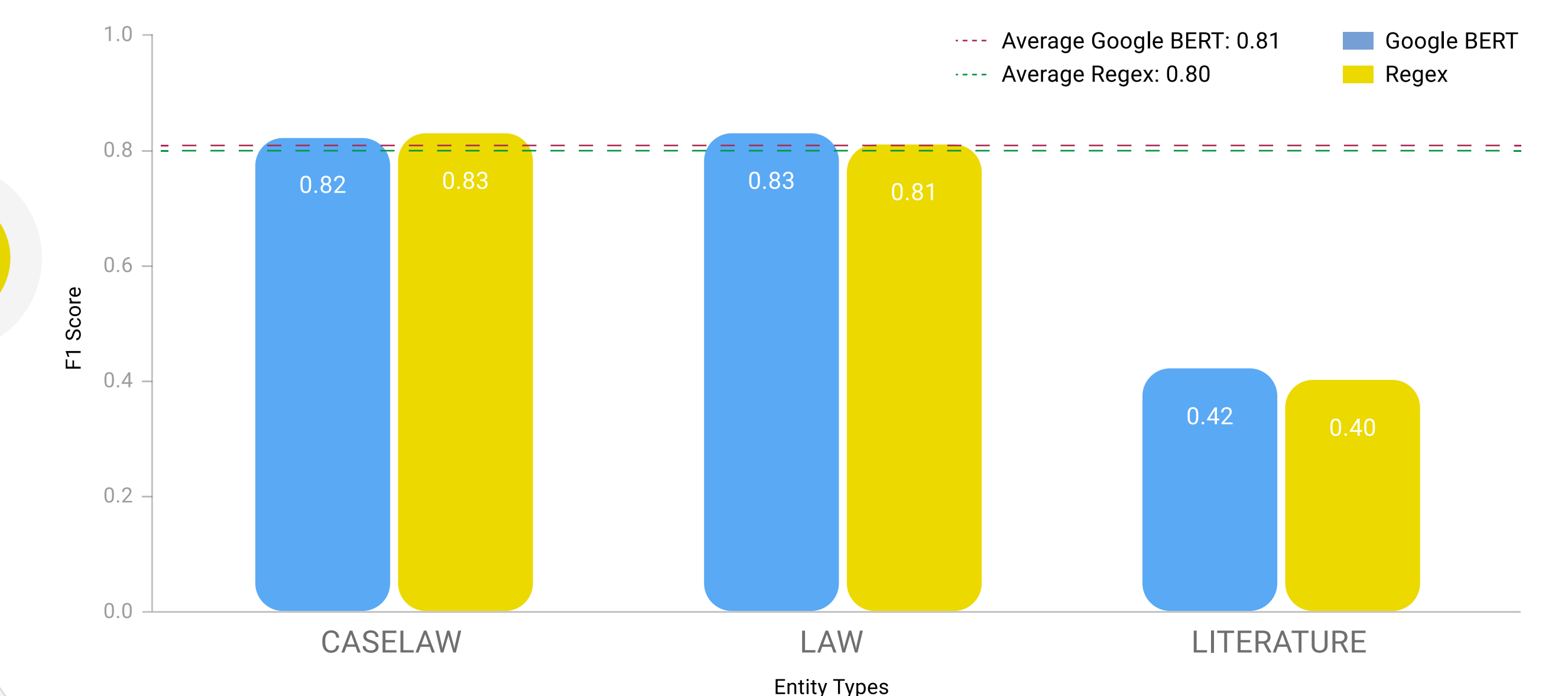


Results

- ✓ Performance matches regex-based system
- ✓ Can handle deviations such as typing errors and new citation formats better
- ✓ Improved performance in recognizing parts of a citation with more granular labels
- ✓ Processes 3000 documents in 32 minutes

Regex-Solution vs. Fine-Tuned Google BERT

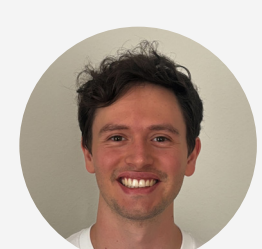
Comparison on manually annotated data



Jasmin Fitz



Dario Glasl



Mirio Eggmann