Named Entity Recognition

Group 6 - Team Octane

202318005 Aditi Singh
202318032 Kruti Patel
202318044 Hani Soni
202318061 Garvika Singh

Problem Statement

In the era of rapidly expanding domain-specific data, efficient extraction of meaningful entities from unstructured text is crucial for knowledge discovery and decision-making.

- Existing NER models are trained on general datasets and fail to capture domain-specific terminologies, resulting in poor performance in specialized fields.This project focuses on developing and training a Named Entity Recognition (NER) model for two distinct domains: Medical and Solar Panel Technology Patents.

- For the Medical domain, entities such as "Chemicals," "Phenotypes," and "Anatomical Structures" are critical for applications like biomedical research, drug discovery, and clinical data analysis.

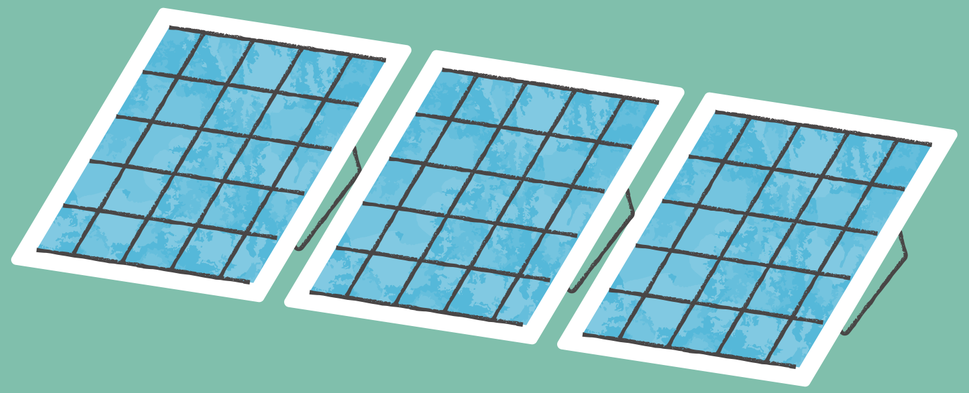
- For the Solar Panel Technology domain, entities like "Components," "Manufacturing Process," and "Performance Metrics" are essential for understanding technological innovations and patent trends.

Exploring Different Datasets

1. Medical Domain Data

https://huggingface.co/datasets/knowledgator/biomed_NER

2. Solar Patent Data



Ectracted from Google Patents

Medical Domain Data Text Snippet


```
"text": "The purpose of this study is to investigate the time course of
PF-00241939 concentration in the blood following dosing by oral inhalation
from dry powder inhalers. IgA nephropathy (IgAN) is the commonest primary
glomerulonephritis worldwide. In Hong Kong, IgAN accounts for
approximately 30% of all primary glomerular diseases, and a significant
proportion of young patients (< 50 years of age) on dialysis therapy are
sufferers of primary IgAN. To date, no specific therapeutic agent has been
consistently shown to halt the progression of IgAN to end-stage renal
failure, particularly in patients with persistent significant proteinuria
and the presence of chronic tubulointerstitial inflammation on kidney
biopsy. In recent years, angiotensin-converting enzyme inhibitors (ACEI)
have been found capable of significantly reducing proteinuria in some IgAN
patients, while others, particularly those with the ACE DD genotype,
showed either absent or unsatisfactory response to angiotensin blockade.
",
```

Solar Patent Data Text Snippet

"text": "the invention discloses a kind of integrated circuit (ic) apparatus and manufacture the method for semiconductor device.provide a kind of area effective high voltage unipolarity esd protection device (300), comprise p-type substrate (303); one p-trap (308-1), is formed in the substrate and size is confirmed as comprising the n+ contact area and p+ contact area (310,312) that are connected to cathode terminal; second independent p-trap (308-2), is formed in the substrate and size is confirmed as only comprising the p+ contact area (311) being connected to anode terminal; and electric floating n-shaped isolation structure (304,306,307-2), formed in the substrate with around and be separated the first semiconductor regions and the second semiconductor regions.when being applied above the positive voltage of trigger voltage level to cathode terminal and anode terminal, esd protection device makes intrinsic thyristor trigger and returns pattern for rapid, to provide by the low impedance path of described structure for discharging to esd electric current."

Entities Medical

```
NEW_VALID_ENTITIES = {
    "CHEMICALS",
    "ACTIVITY",
    "PHENOTYPE",
    "ANATOMICAL STRUCTURE",
    "GROUP",
    "GENE AND GENE PRODUCTS",
    "GEOGRAPHICAL AREAS",
    "CLINICAL DRUG",
    "SIGNALING MOLECULES",
    "PRODUCTS",
    "DISORDERS",
    "ORGANIZATIONS"}
```

Entities Solar

```
VALID_ENTITIES = {
    "Components",
    "Efficiency Improvement Techniques",
    "Energy Storage,
    "Innovation Objectives",
    "Manufacturing Process",
    "Materials-CM",
    "Module Structures",
    "Performance Metrics"
}
```

Weed seed inactivation in soil mesocosms via biosolarization with mature compost and tomato processing waste amendments Biosolarization is a fumigation alternative
•CHEMICALS •ACTIVITY •CHEMICALS •CHEMICALS •ACTIVITY •ACTIVITY

that combines passive solar heating with amendment-driven soil microbial activity to temporarily create antagonistic soil conditions, such as elevated temperature and
•CHEMICALS •CHEMICALS

acidity, that can inactivate weed seeds and other pest propagules. The aim of this study was to use a mesocosm -based field trial to assess soil heating, pH, volatile fatty
•CHEMICALS •CHEMICALS •CHEMICALS

acid accumulation and weed seed inactivation during biosolarization. Biosolarization for 8 days using 2% mature green waste compost and 2 or 5% tomato processing
•ACTIVITY •ACTIVITY •CHEMICALS

residues in the soil resulted in accumulation of volatile fatty acids in the soil, particularly acetic acid, and> 95% inactivation of Brassica nigra and Solanum nigrum seeds.
•CHEMICALS •CHEMICALS •CHEMICALS •CHEMICALS •CHEMICALS

Inactivation kinetics data showed that near complete weed seed inactivation in soil was achieved within the first 5 days of biosolarization. This was significantly greater
•CHEMICALS •ACTIVITY

than the inactivation achieved in control soils that were solar heated without amendment or were amended but not solar heated. The composition and concentration of
•CHEMICALS

organic matter amendments in soil significantly affected volatile fatty acid accumulation at various soil depths during biosolarization. Combining solar heating with organic
•CHEMICALS •CHEMICALS •CHEMICALS •CHEMICALS •ACTIVITY •CHEMICALS

matter amendment resulted in accelerated weed seed inactivation compared with either approach alone.
•ACTIVITY

the moisture and temperature stability in the doped graphene is improved by forming a hydrophobic organic layer on
•Materials-CM •Materials-CM

the surface of the doped graphene on which dopants are doped. the degradation of the doped graphene can be
•Materials-CM •Materials-CM •Efficiency I...
•Materials-CM

prevented by the improved stability; thus, a transparent electrode comprising the doped graphene including a
•Components-C... •Materials-CM

hydrophobic organic layer can be useful in a solar cell and a display device.
•Materials-CM

Entities Medical Summary

Avg text length: 42 tokens per text

Corpus Size: 4k texts - Fully Labelled

Total Entities: 12

Entities Solar Summary

Avg text length: 38 tokens per text

Corpus Size: 20k texts - 680 Labelled

Total Entities: 8

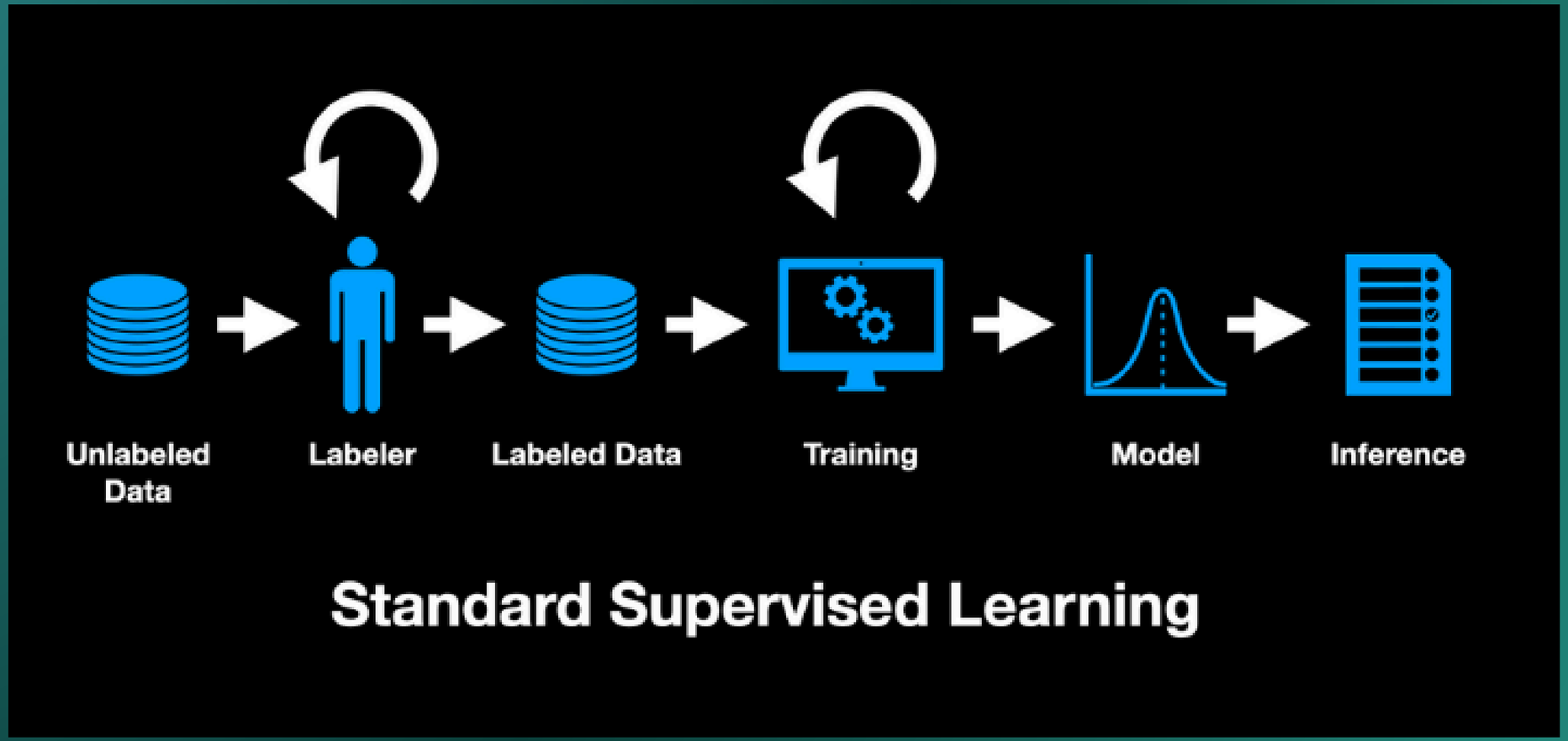
Data Format

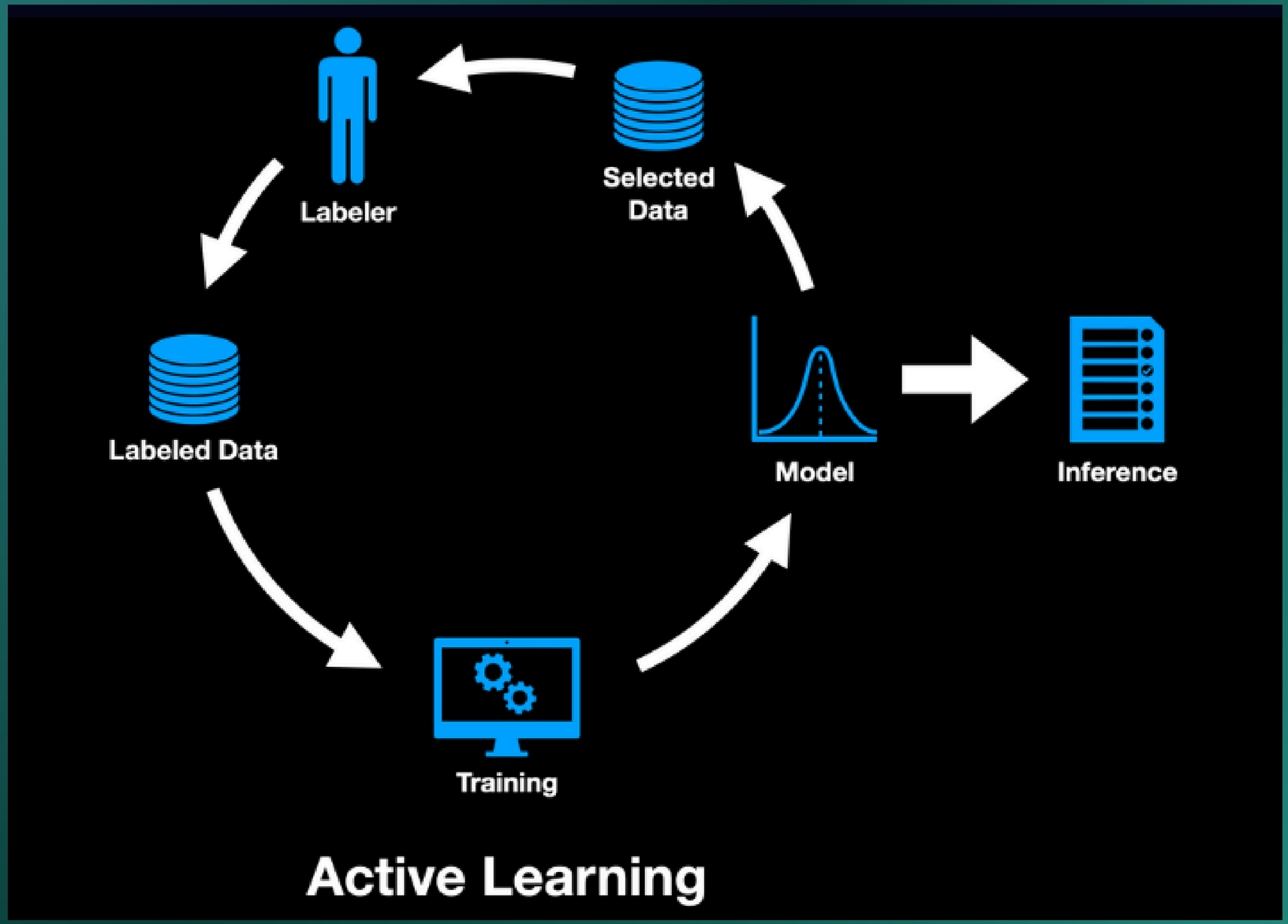

```
{"id": 4568, "text": "The purpose of this study", "aid":
"3badea39c1e1eddbc3d6b459ecf869b9", "entities": [{"id": 181925, "label":
"CHEMICALS", "start_offset": 77, "end_offset": 93}, {"id": 181927,
"label": "CLINICAL DRUG", "start_offset": 136, "end_offset": 150}, {"id":
181928, "label": "GENE AND GENE PRODUCTS", "start_offset": 158,
"end_offset": 164}]}
```

Data Annotations Using Active Learnig

When to use active learning approach

- Domain Expertise Required for Annotation:
- High Annotation Cost
- Limited Labeled Data Availability


Unlabeled Data
Labeler
Labeled Data
Training
Model
Inference
Standard Supervised Learning


Labeler
Selected Data
Labeled Data
Model
Inference
Training
Active Learning

GLiNER:

- GLiNER is designed for lightweight, efficient Named Entity Recognition (NER) tasks
- Resource-constrained environments.
- Zero Shot Learning on custom defined entities.
- Pre-Trained on synthetic data generated by chatgpt

```python
from gliner import GLiNER

model = GLiNER.from_pretrained("urchade/gliner_multi_pii-v1")

text = """
Harilala Rasoanaivo, un homme d'affaires local d'Antananarivo, a enre
"""

labels = ["work", "booking number", "personally identifiable informat
entities = model.predict_entities(text, labels)

for entity in entities:
    print(entity["text"], "=>", entity["label"])
```

Our Problem - 100% of solar patent data was unlabelled

Kick started inital 200 annotations using GLiNER zero shot

200 Mannualy Verified Samples

Model Training

Model Inference on Unlabelled Corpus

Least Confidence Sampling bottom 200 least confident sentences

Manual Validation of these sentences

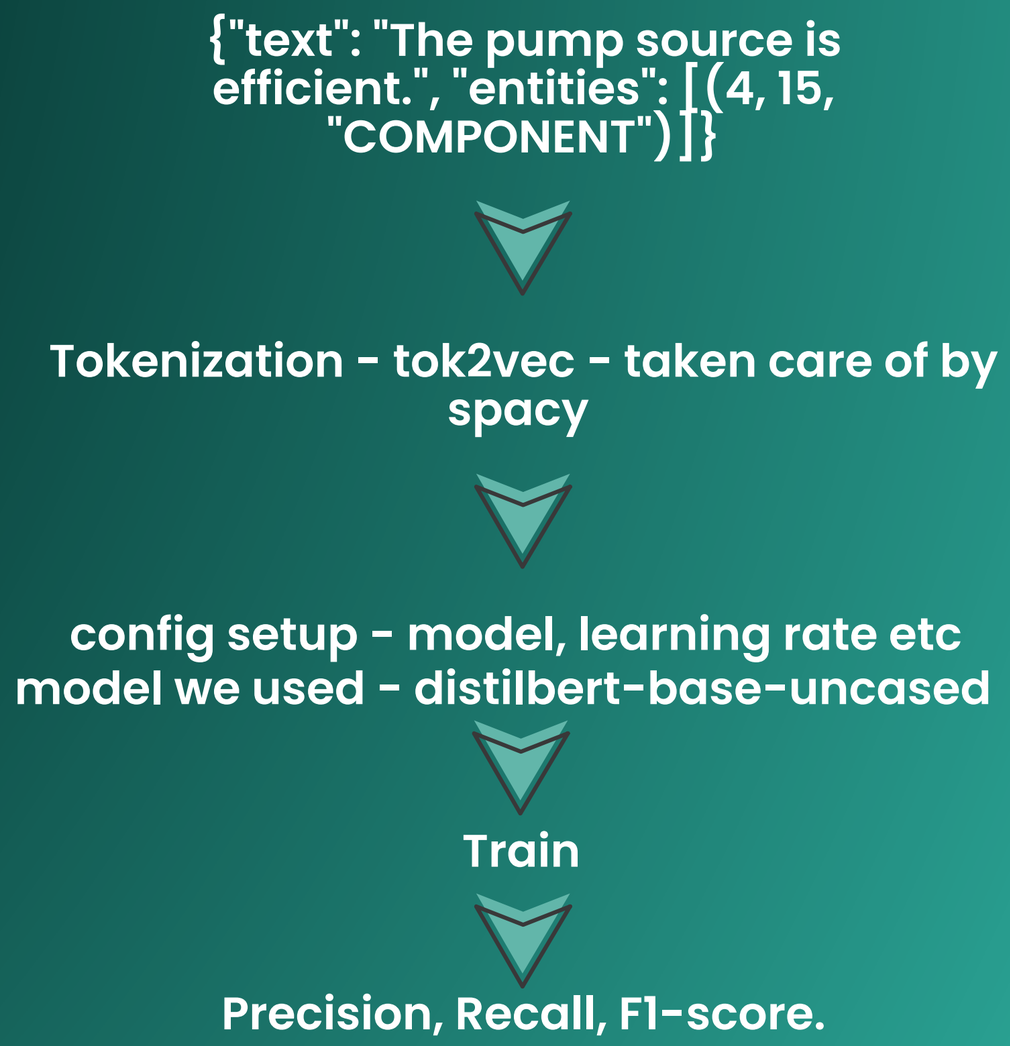
Model Training on (All Prev + Curr) batch sentences

NER trainig with Spacy

We used spacy’s automated pipeline with defaut config and custome label and model

Steps -

- Data Preparation
- Pipeline Configuration
- Model Initialization
- Fine-tuning
- Evaluation
- Inference



Least Confidence Sampling

$$\phi^{LC}(x) = 1 - P(y^*|x;\theta)$$

Where:

- $P(y^*|x;\theta)$ is the probability of the most likely class label y^* for a given input x (the sentence or word).
- $\phi^{LC}(x)$ is the confidence score of the input x.

Sentence: "The invention relates to nanostructured solar cells."

Step-by-step Explanation:

1. **Training the Classifier**:
   - Initially, a small set of sentences (1% of the training data) is manually labeled and used to train the classifier C.
2. **Prediction for each word**:
   - For each word in the sentence, the classifier predicts probabilities for each entity class (e.g., "Materials-CM", "Components-CP").
   - Suppose the classifier assigns the following probabilities to the word "solar" (hypothetical):
     - "Solar":
       - "Materials-CM": 0.20
       - "Components-CP": 0.80
3. **Calculate Least Confidence Score for each word**:
   - The Least Confidence score for "solar" is:
     $$\phi^{LC}(solar) = 1 - \mathrm{P}(\text{Components-CP} \mid \text{solar}) = 1 - 0.80 = 0.20$$
4. **Sentence Confidence Score**:
   - The sentence confidence score is calculated by averaging the LC scores of all words:
     - For simplicity, assume the word "nanostructured" has an LC score of 0.10, and "cells" has an LC score of 0.50.
     - The mean sentence confidence score is:
       $$\phi^{LC}(\text{sentence}) = \frac{0.20 + 0.10 + 0.50}{3} = 0.27$$
5. **Rank Sentences**:
   - After processing multiple sentences in the unlabeled set U, the sentences are ranked by their confidence scores. Sentences with the highest LC scores (i.e., those with the least confident predictions) are selected for annotation and added to the labeled set.

Evalation

Medical Data

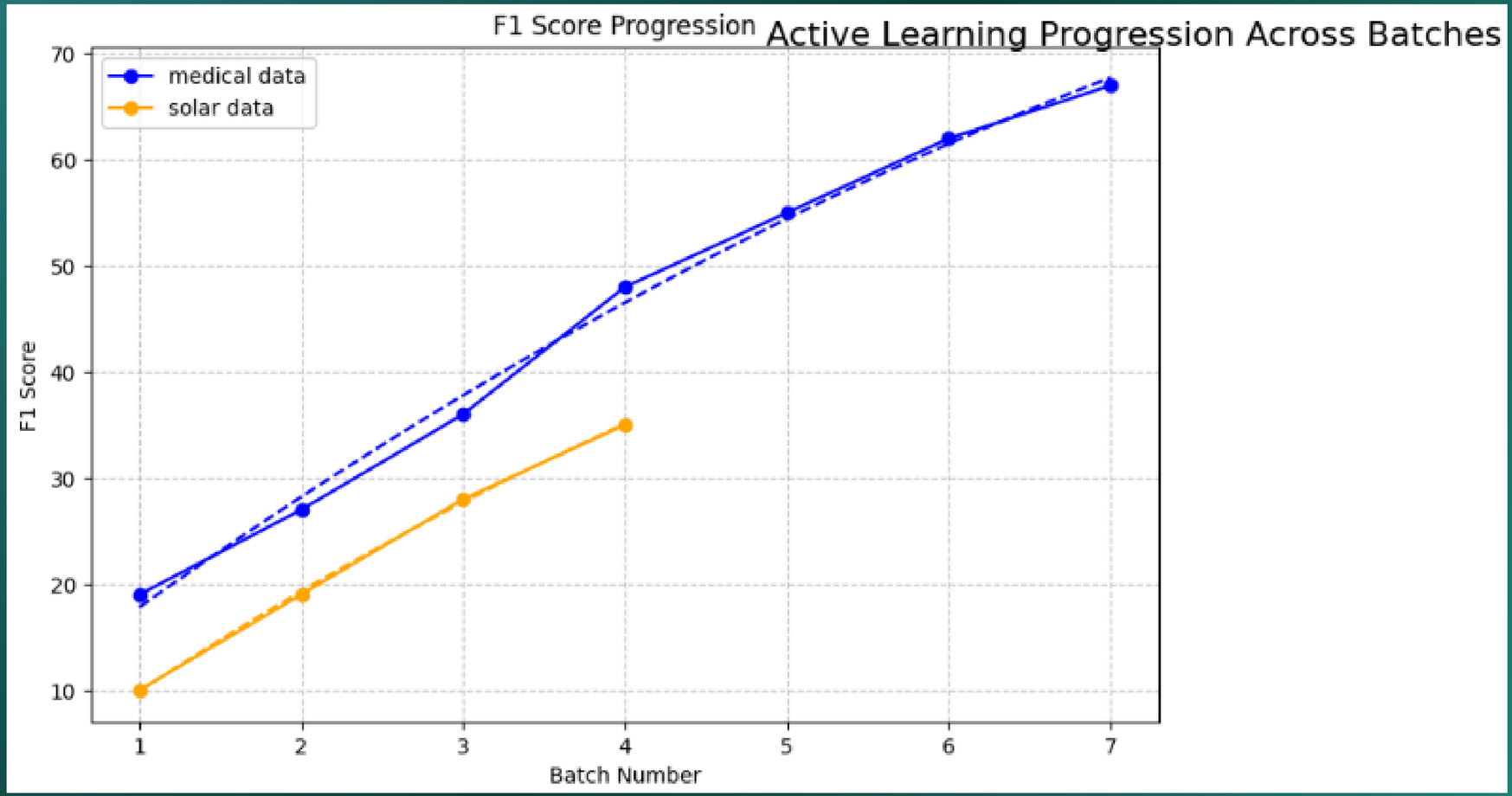
batch: 7


```
"overall": {
  "precision": 0.7537,
  "recall": 0.6044,
  "f1": 0.6727,
  "sup": 2117,
  "predicted": 1731
}
```


Solar Data

batch: 4


```
"overall": {
  "precision": 0.5137046861184792,
  "recall": 0.2744449629617385,
  "f1": 0.3577586206896552,
  "sup":1310
  "predicted": 1131
}
```



Link to inferece