EXTRACTING KEYPHRASES AND RELATIONS FROM SCIENTIFIC PUBLICATIONS

Team No. 55

Introduction to NLP

Project Presentation

Team Members

Aakash Singh 2021201087

Mayank Mukundam 2021201057

Sourabh Patidar 2021201089

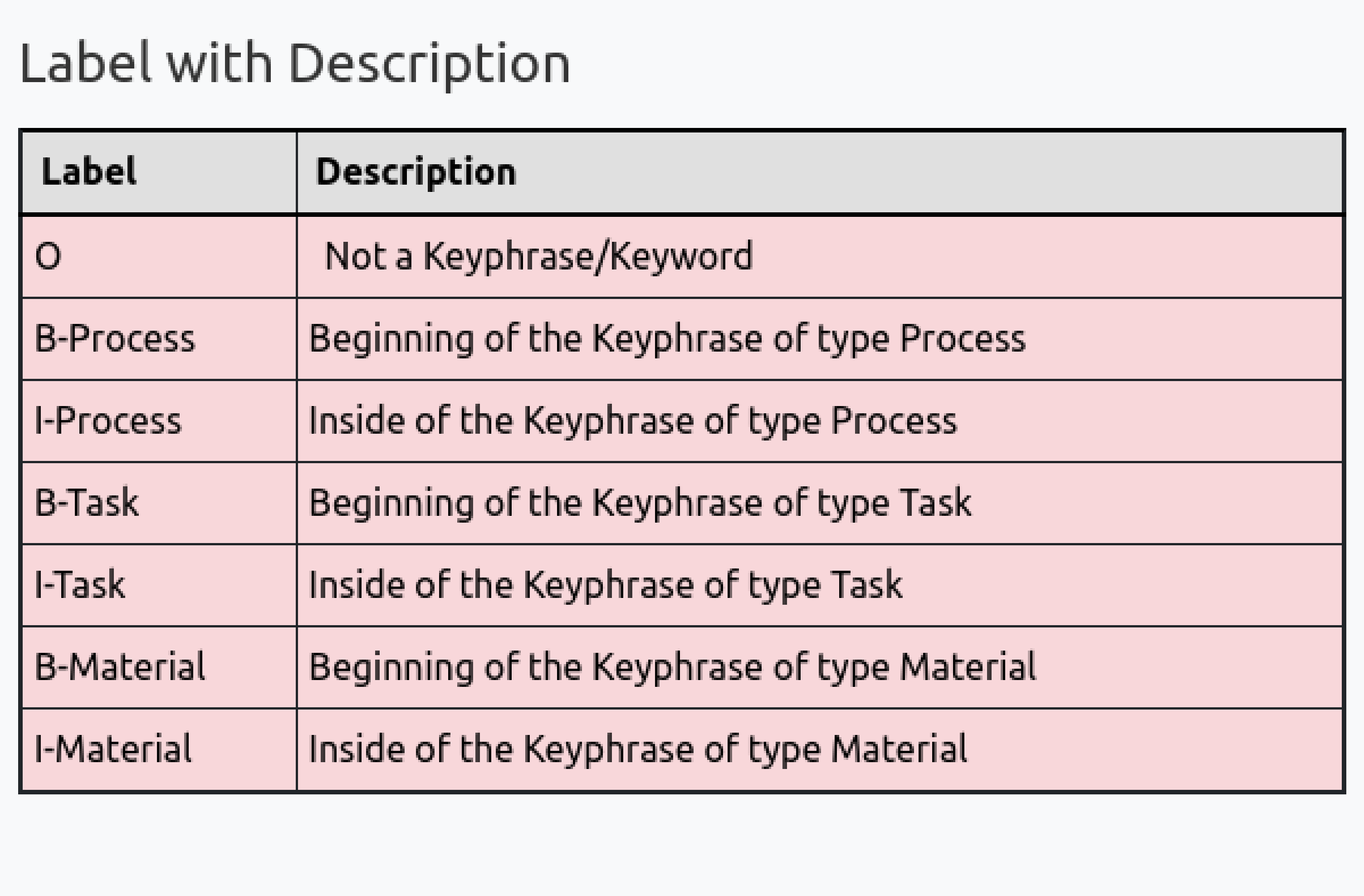
PROBLEM STATEMENT

- Our task is to present a method to extract keyword or keyphrases and relation between them from given piece of scientific text.
- This task can be classified further into 3 subtasks :
  - Extracting Keyphrases: Extract Key Scientific Phrases
  - Classify the Keyphrases: Classify the KeyPhrase in Process, Task & Material.
  - Identify the relationship: Identify the relations between Keyphrases

DATASET

- We have used the SemEval 2017 Task 10 dataset.
- The dataset is divided into train, validation and test set.
- Training set, Validation set and Test set contains of 350, 100 and 50 annotations and text file each.
- Text file contains the extract from some article, it’s a 200-300 words paragraph.
- Annotation file contains the keyphrase boundaries and type of keyphrase.

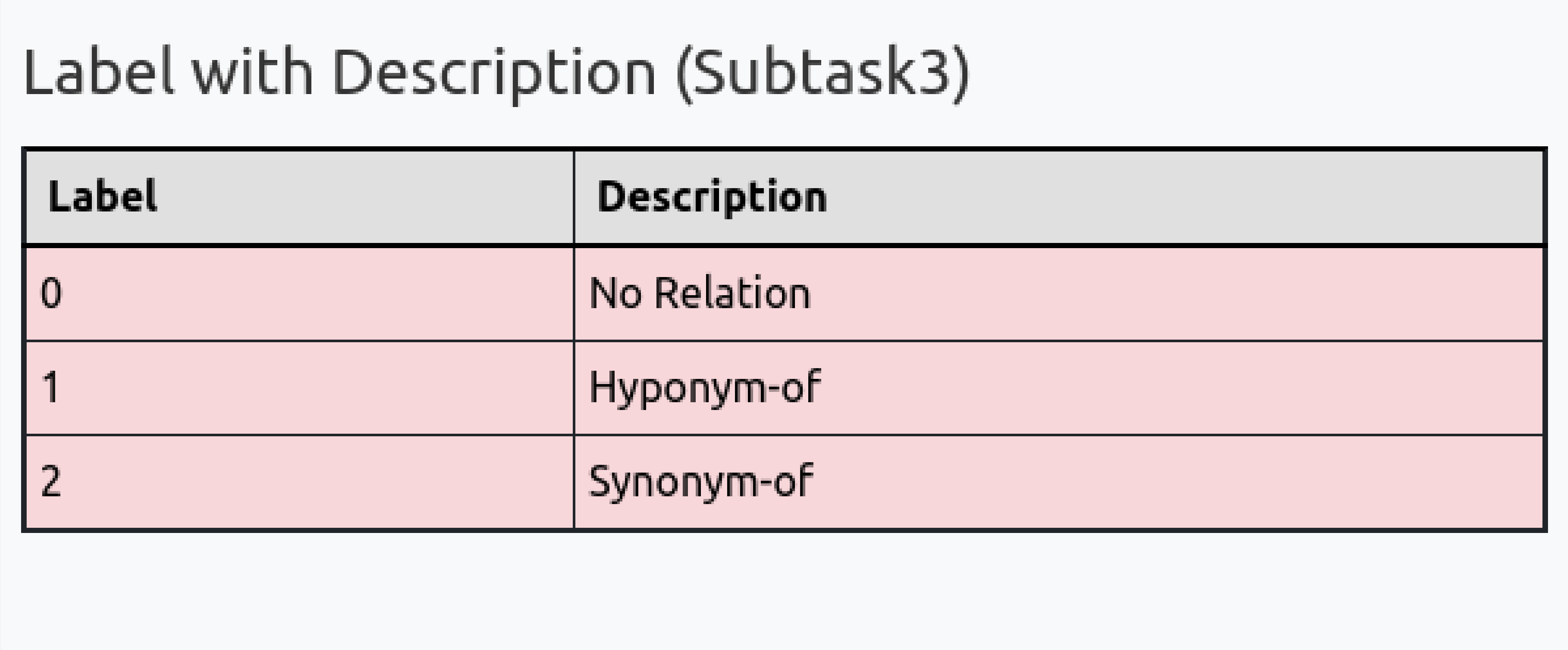
Data Preprocessing (for subtask 1 and 2)



Label with Description

Label	Description
O	Not a Keyphrase/Keyword
B-Process	Beginning of the Keyphrase of type Process
I-Process	Inside of the Keyphrase of type Process
B-Task	Beginning of the Keyphrase of type Task
I-Task	Inside of the Keyphrase of type Task
B-Material	Beginning of the Keyphrase of type Material
I-Material	Inside of the Keyphrase of type Material

DATA PREPROCESSING (FOR SUBTASK 3)



Label with Description (Subtask3)

Label	Description
0	No Relation
1	Hyponym-of
2	Synonym-of

METHODOLOGY (OVERVIEW)

- The task was divided into 3 subtask.
- We have combined the subtask1(i.e. Keyword extraction) and subtask2(i.e. Keyword Classification).
- And Subtask3(i.e. identifying relations) was performed and evaluated independent of the previous 2 subtasks.

METHODOLOGY(PART-1 AND PART-2)

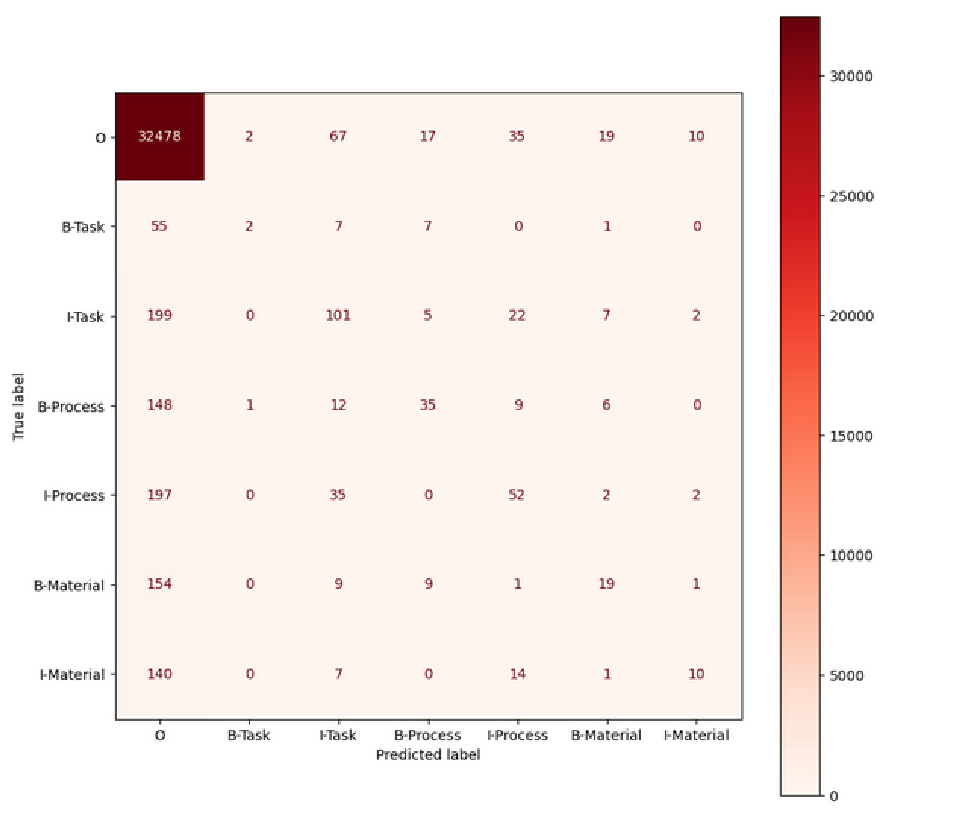
- We have used the SciBERT, the variation of BERT model which is pre-trained on the Scientific Data.
- The model is finetuned on the training set which has 55135 tokens.
- The model has been tested on the 18259 tokens.
- SciBERT -
  - Model architecture same as BERT-base model, just pre-trained on data of Scientific Domain
  - It has its own vocab, scivocab that is built to best match the scientific domain.

METHODOLOGY(PART-3)

- We had a pair of entities and relationship between them.
- Using sentence-transformer library, we have calculated the embeddings of each entities.
- Each embedding vector is of size 384.
- Concatenated both the entities, to get a vector of size 768.
- Trained a svm classifier on training data (around 1350 datapoints).

EVALUATION (SUBTASK 1 & 2)

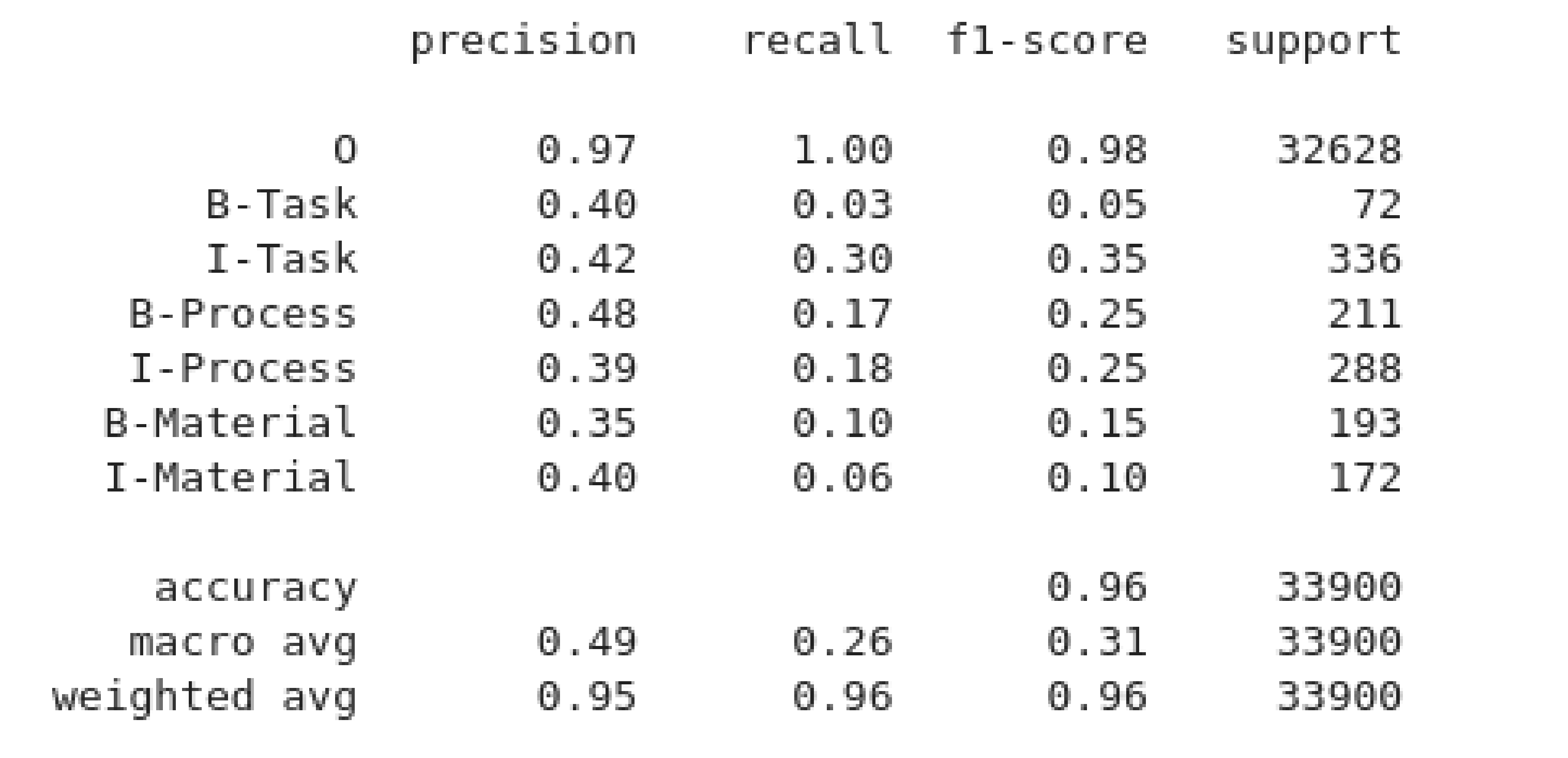
- Confusion Matrix



True label \ Predicted label	O	B-Task	I-Task	B-Process	I-Process	B-Material	I-Material
O	32478	2	67	17	35	19	10
B-Task	55	2	7	7	0	1	0
I-Task	199	0	101	5	22	7	2
B-Process	148	1	12	35	9	6	0
I-Process	197	0	35	0	52	2	2
B-Material	154	0	9	9	1	19	1
I-Material	140	0	7	0	14	1	10

EVALUATION (SUBTASK 1 & 2)

- Classification Report



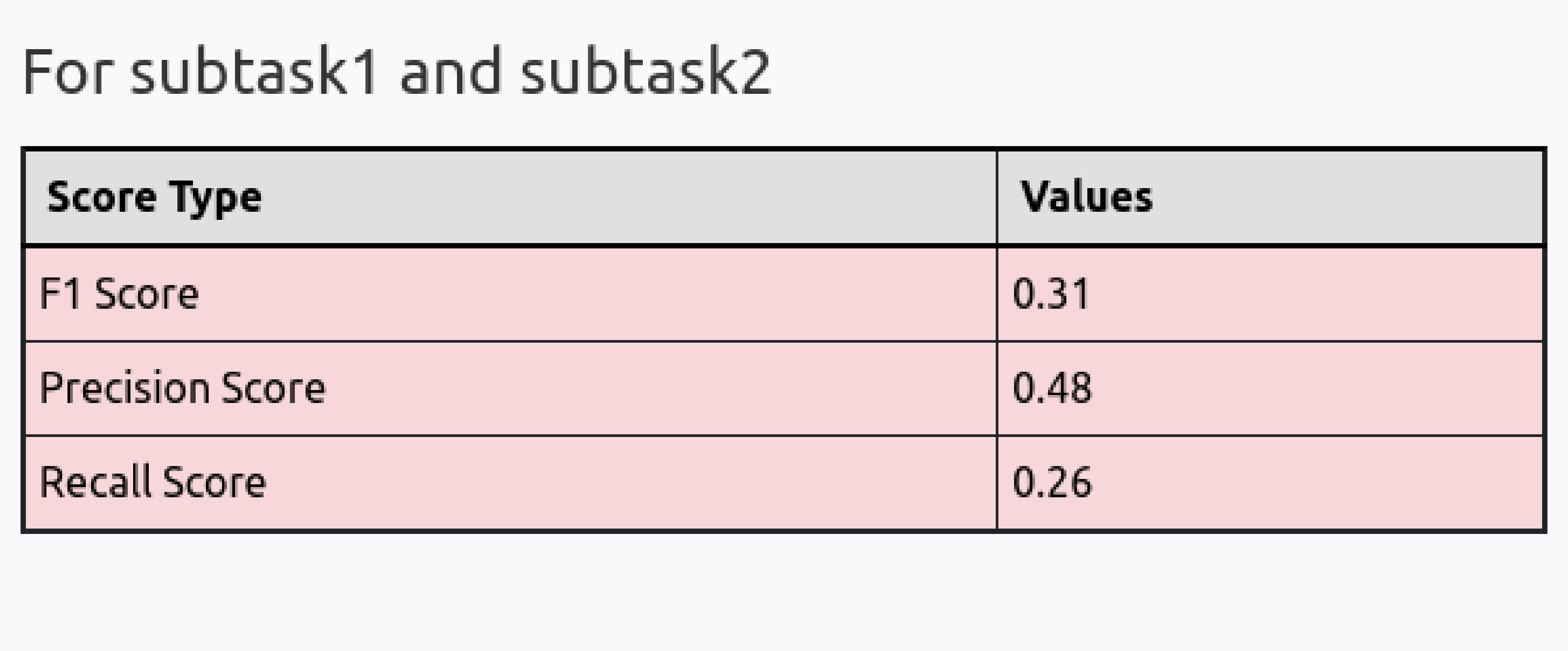
```
              precision    recall  f1-score   support

           O       0.97      1.00      0.98     32628
      B-Task       0.40      0.03      0.05        72
      I-Task       0.42      0.30      0.35       336
   B-Process       0.48      0.17      0.25       211
   I-Process       0.39      0.18      0.25       288
  B-Material       0.35      0.10      0.15       193
  I-Material       0.40      0.06      0.10       172

    accuracy                           0.96     33900
   macro avg       0.49      0.26      0.31     33900
weighted avg       0.95      0.96      0.96     33900
```

EVALUATION (SUBTASK 1 & 2)

- Precision, Recall and F1-score

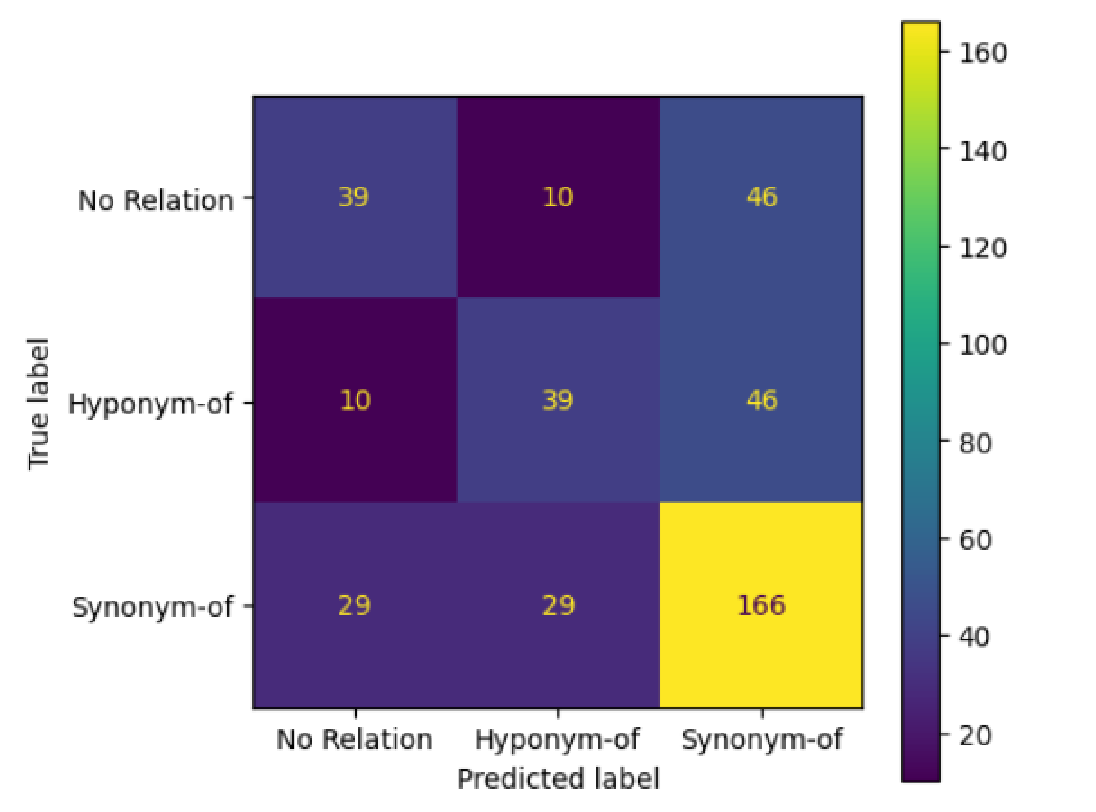


For subtask1 and subtask2

Score Type	Values
F1 Score	0.31
Precision Score	0.48
Recall Score	0.26

Evaluation (Subtask 3)

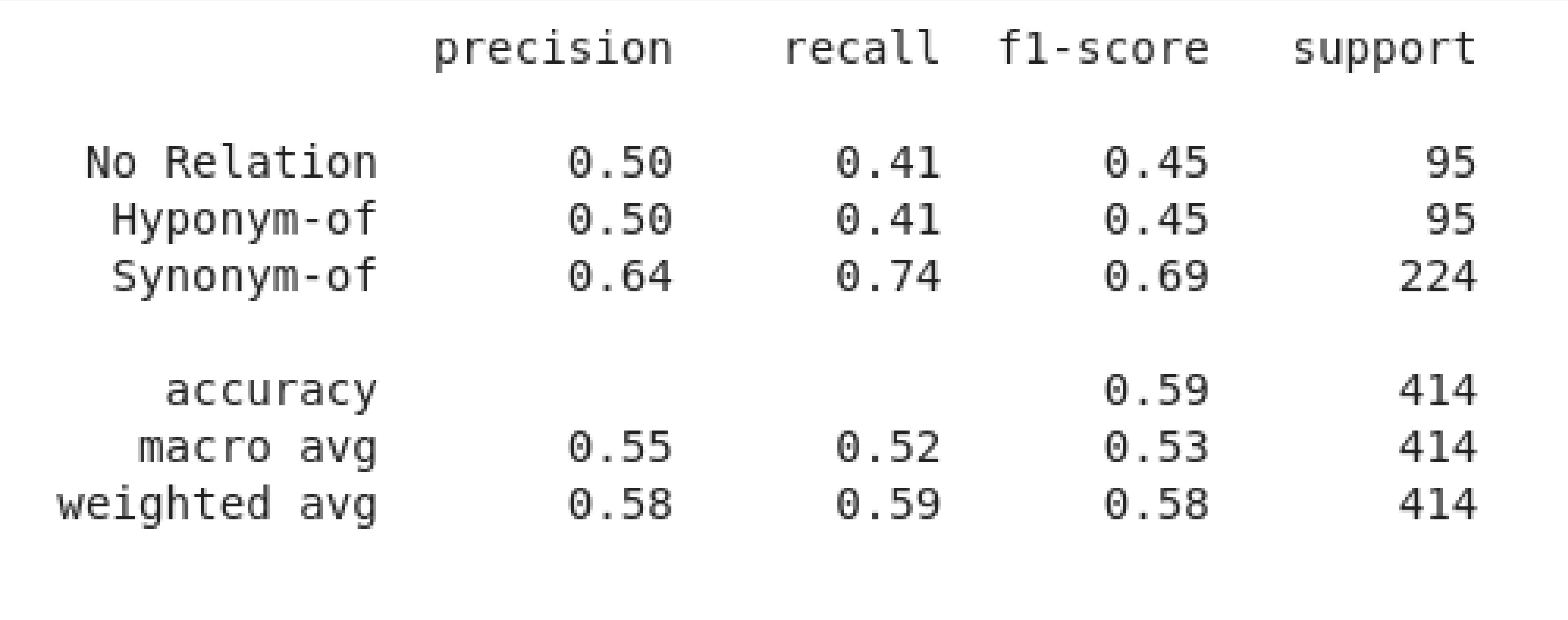
- Confusion Matrix



True label \ Predicted label	No Relation	Hyponym-of	Synonym-of
No Relation	39	10	46
Hyponym-of	10	39	46
Synonym-of	29	29	166

EVALUATION (SUBTASK 3)

- Classification Report



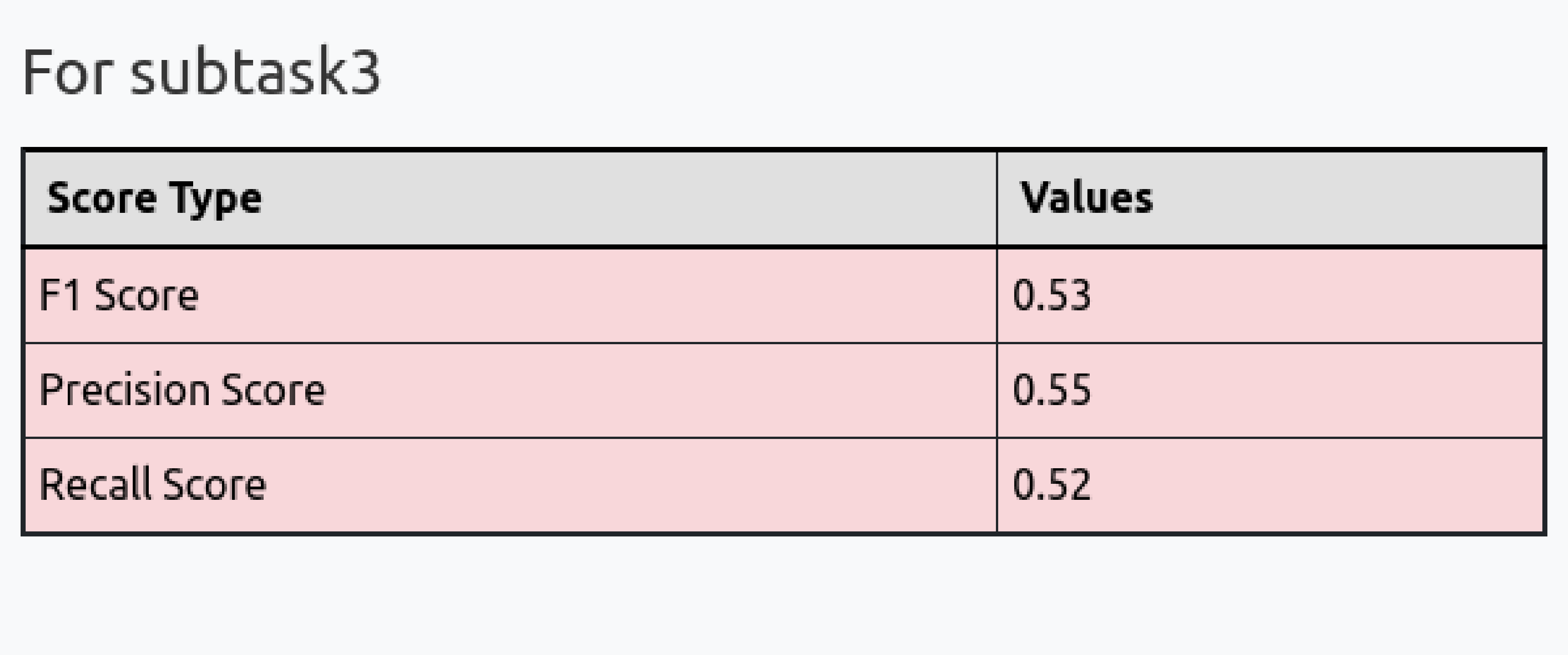
```
              precision    recall  f1-score   support

 No Relation       0.50      0.41      0.45        95
  Hyponym-of       0.50      0.41      0.45        95
  Synonym-of       0.64      0.74      0.69       224

    accuracy                           0.59       414
   macro avg       0.55      0.52      0.53       414
weighted avg       0.58      0.59      0.58       414
```

EVALUATION (SUBTASK 3)

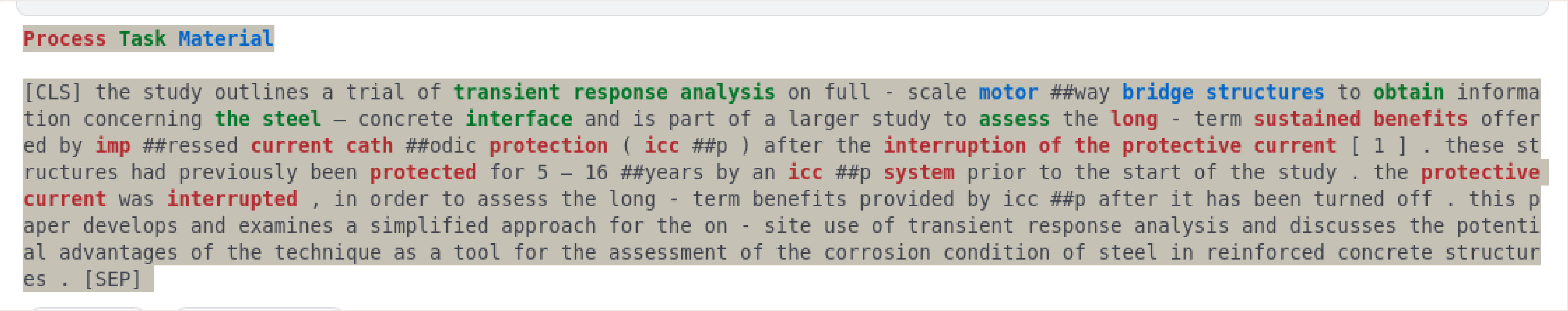
- Precision, Recall and F1-score



For subtask3

Score Type	Values
F1 Score	0.53
Precision Score	0.55
Recall Score	0.52

SAMPLE RESULTS



```
Process Task Material

[CLS] the study outlines a trial of transient response analysis on full - scale motor ##way bridge structures to obtain informa
tion concerning the steel – concrete interface and is part of a larger study to assess the long - term sustained benefits offer
ed by imp ##ressed current cath ##odic protection ( icc ##p ) after the interruption of the protective current [ 1 ] . these st
ructures had previously been protected for 5 – 16 ##years by an icc ##p system prior to the start of the study . the protective
current was interrupted , in order to assess the long - term benefits provided by icc ##p after it has been turned off . this p
aper develops and examines a simplified approach for the on - site use of transient response analysis and discusses the potenti
al advantages of the technique as a tool for the assessment of the corrosion condition of steel in reinforced concrete structur
es . [SEP]
```

CONCLUSION

- The Scientific domain has very few annotated datasets available, SemEval 2017 task 10 was a sweet and short attempt to make dataset available for research purposes in scientific research domain.
- All the submissions of the task are based on RNNs and LSTMs, so we tried to solve the problem using transformers.
- For the same purpose we have used the pretrained SciBERT which is a scientific domain variation of BERT to solve the first 2 subtasks and also used ever reliable SVM to solve the 3rd subtask.

THANK YOU