NER Extraction

Team Name: Cheese Maggi

Aryan Jain (2019101056)

Palash Sharma (2019101082)

Arth Raj (2019101094)

About the Project

- Preparing Named Entity tagged data using weak supervision with help of snorkel.

- Performing Fine Tuning for F1 score improvement.

What is NER and Snorkel?

NER is an information extraction technique to identify and classify named entities in text.

Snorkel is a system that facilitates the process of building and managing training datasets without manual labeling.

Motivation

- NER has a wide variety of use cases in the business.

- Applications of NER include:
  - Extracting important named entities from legal, financial, and medical documents
  - Classifying content for news providers
  - Improving the search algorithms, and etc.

Datasets used

- For Snorkel labelling: We have used re3d defense dataset for our NER extraction.

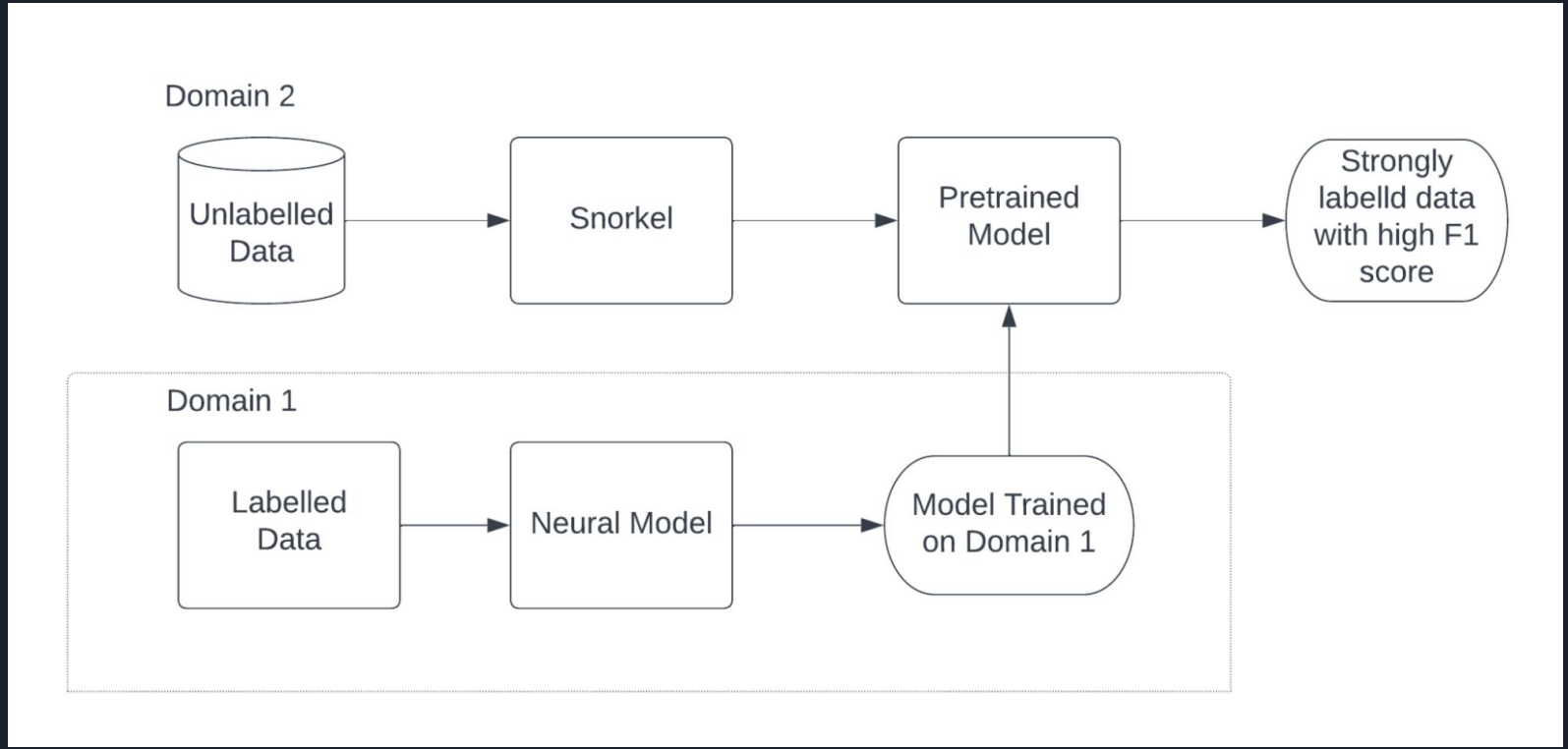
- For training Neural Network: We have used the Conll2003 News dataset.

Procedure

The architecture is designed in three steps:

- Firstly we build labelled data using weak supervision with the help of snorkel.
- In next step we pass this data to a pretrained model(which is trained in some other domain).
- After this we finetune(transfer learning via bootstrapping) it which results in significant improvements in accuracy and F1 score.

Pipeline


Domain 2
Unlabelled Data
Snorkel
Pretrained Model
Strongly labelld data with high F1 score
Domain 1
Labelled Data
Neural Model
Model Trained on Domain 1

First Step :- Weak Supervision using Snorkel

To label unlabelled data, we can turn to a weak supervision approach, using labeling functions (LFs) in Snorkel: noisy, programmatic rules and heuristics that assign labels to unlabelled training data. We have used 8-10 labeling functions for the re3d dataset to classify them into classes: Money, Quantity, Document Reference, Location, Weapon, Nationality, Organization, Temporal. The tokens which do not fall in any of these classes are labeled as ‘O’, which means Out of Tag.

Second Step :- Learning using Neural Model

This part of the architecture tries to learn the intuition from this weak labeling to label the same dataset with a much better accuracy. The main architecture of the neural model consists of:

Layer 1: Elmo Pretrained embedding model

Layer 2: Bi-LSTM with dropout

Layer 3: Bi-LSTM with dropout

Layer 4: Fully connected layer with softmax activation

This model is trained on a sufficiently large dataset first so that the inherent nature of the problem and understanding required is gained by the LSTM layers.

Third Step :- Transfer learning via Bootstrapping (Fine-tuning)

This is the last part of the architecture where we need the pre-trained neural model to learn new domain specific information. Hence, we remove the last layer and replace it with 2 new FC layers, freeze the old layers and train the model on new, weakly labeled (by snorkel) dataset so that the new layers learn how to classify the new tags needed. Finally, all the layers are unfrozen and fine-tuned on 100-200 epochs to get a supreme F1-score.

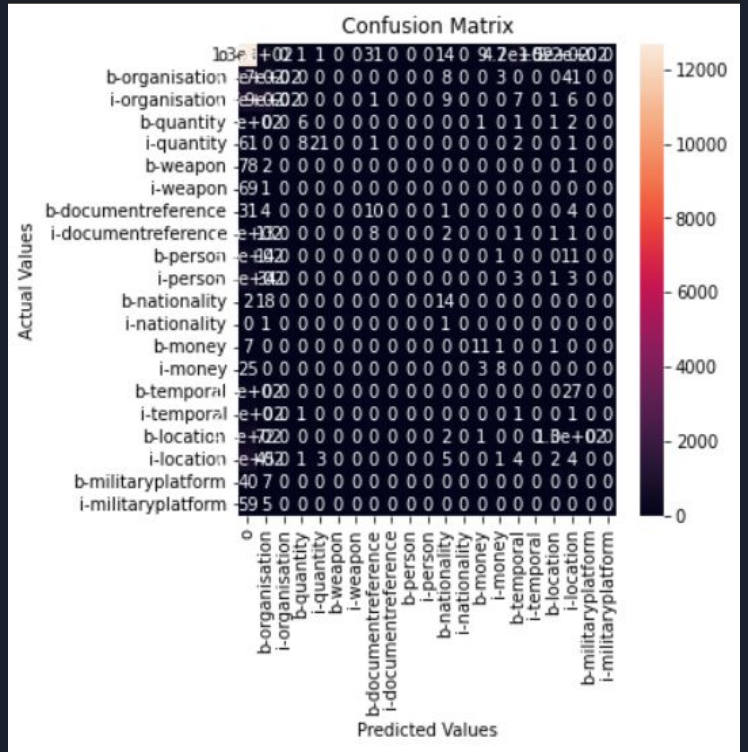
Results and Analysis

We ran our snorkel based weak supervision approach on re3d data and obtained an accuracy of ~65%.

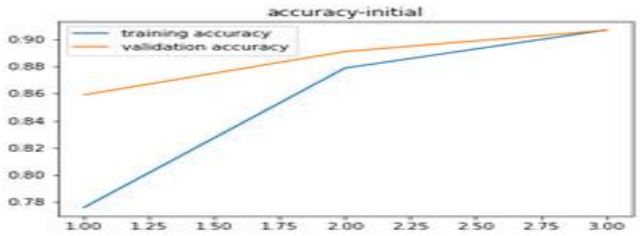
Cumulative F1 score came out to be around 0.58.

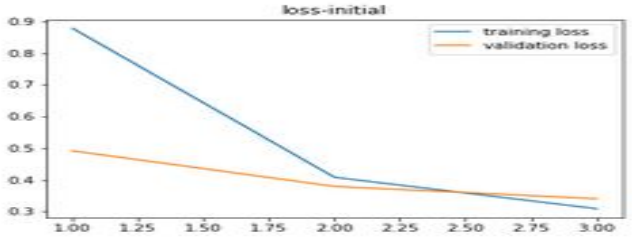
Classification Report & Confusion Matrix on the snorkel labelled dataset

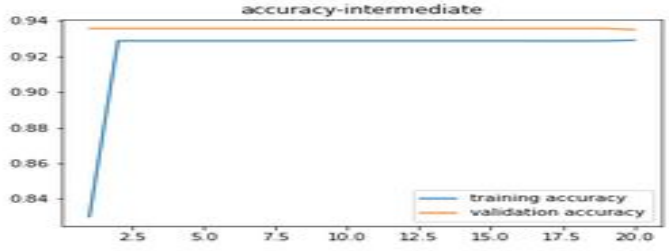
	precision	recall	f1-score	support
o	0.73	0.92	0.82	13747
b-organisation	0.18	0.24	0.21	1104
i-organisation	1.00	0.00	0.00	1539
b-quantity	0.35	0.04	0.07	150
i-quantity	0.84	0.22	0.35	94
b-weapon	1.00	0.00	0.00	81
i-weapon	1.00	0.00	0.00	70
b-documentreference	0.20	0.20	0.20	50
i-documentreference	1.00	0.00	0.00	146
b-person	1.00	0.00	0.00	411
i-person	1.00	0.00	0.00	605
b-nationality	0.25	0.41	0.31	34
i-nationality	1.00	0.00	0.00	2
b-money	0.44	0.55	0.49	20
i-money	0.38	0.22	0.28	36
b-temporal	0.00	0.00	0.00	177
i-temporal	1.00	0.00	0.00	225
b-location	0.00	0.00	0.00	644
i-location	0.01	0.01	0.01	784
b-militaryplatform	1.00	0.00	0.00	47
i-militaryplatform	1.00	0.00	0.00	64
accuracy			0.65	20030
macro avg	0.64	0.13	0.13	20030
weighted avg	0.68	0.65	0.58	20030

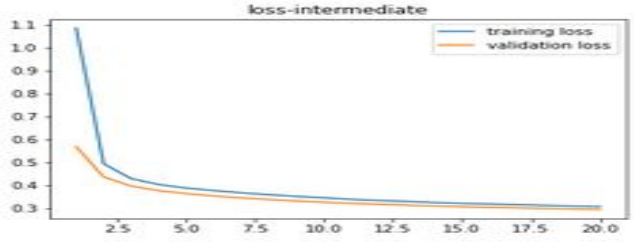


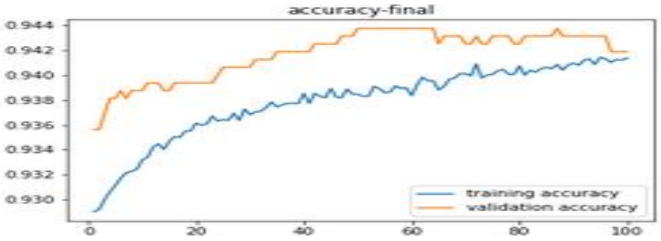
Neural Model Training & Transfer Learning Results

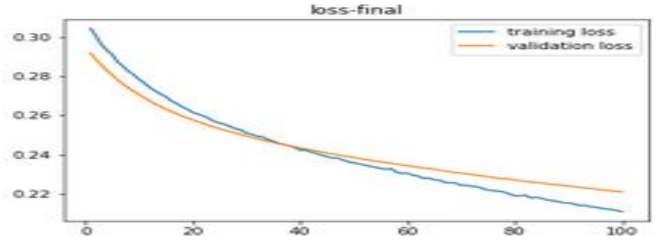










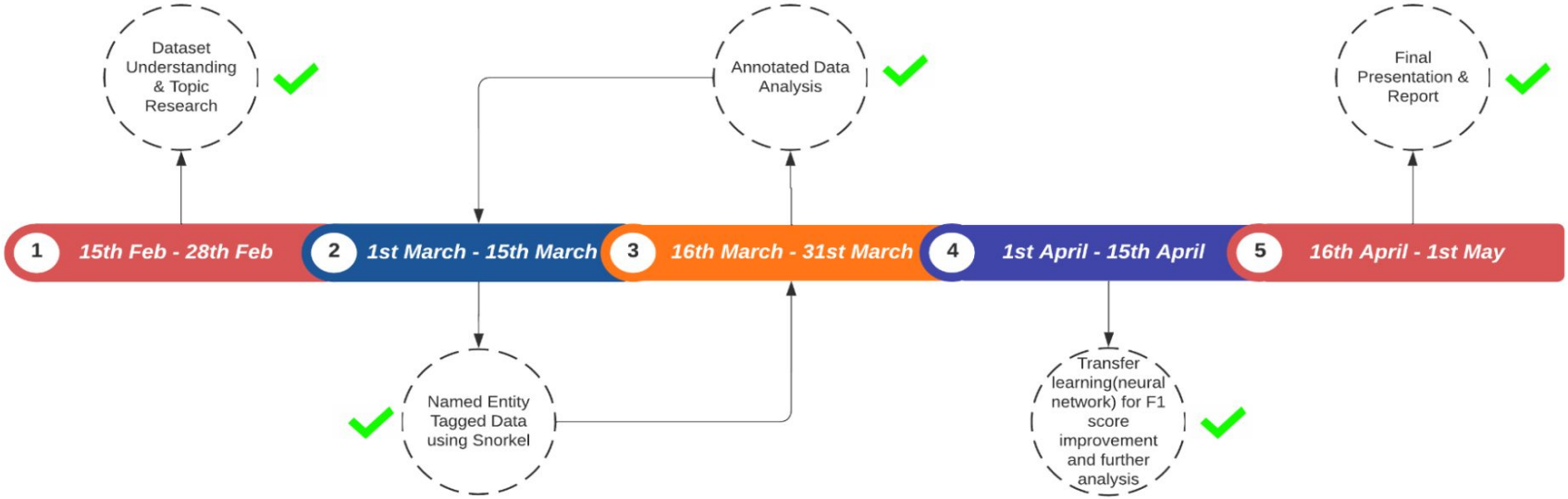


It is visible above that the accuracies have reached a good value at the end of each training for all training steps involved in the architecture. Accuracies upto 96% were reached consistently.

Conclusion

Manual labeling has been a big problem in the field of Machine Learning. To solve this problem, one can label any unlabelled dataset, firstly using weak supervision with the help of snorkel and then, the tagged data can be passed to a pre-trained neural model with some fine tuning, results in significant accuracy improvement and high F1-score. This is achieved due to the transfer learning approach used. Hence, unlabelled datasets can be labeled with sufficiently good F1 scores for use in various ML tasks.

Timeline


Dataset Understanding & Topic Research
Annotated Data Analysis
Final Presentation & Report
1
15th Feb - 28th Feb
2
1st March - 15th March
3
16th March - 31st March
4
1st April - 15th April
5
16th April - 1st May
Named Entity Tagged Data using Snorkel
Transfer learning(neural network) for F1 score improvement and further analysis

Thank You