Univ.AI

# Project NER

*19th June 2022*

## Team: Towards_NLP

Rachit Jain    Anshul Bhardwaj

NLP Scholars
AI-3 C-2 @ Univ.AI

## Executive Summary

- **About** | Understanding the motivation behind the project and the task at hand
- **Data & Labels** | Visualizing what the data looks like and how it can be processed
- **Models & Results** | Developing models for the task and iterating to improve the results
- **Conclusion & Scope of Improvement** | Listing the insights gained from the project and noting down areas for further improvement

## Understanding the motivation behind the project and the task at hand

The project focuses on creating a NER model that identifies key tokens and classifies them into set of predefined entities. The data would involve scientific publications in the WIESP Dataset.

NER helps us extract key information from scientific papers which can help search engines to better select and filter articles.

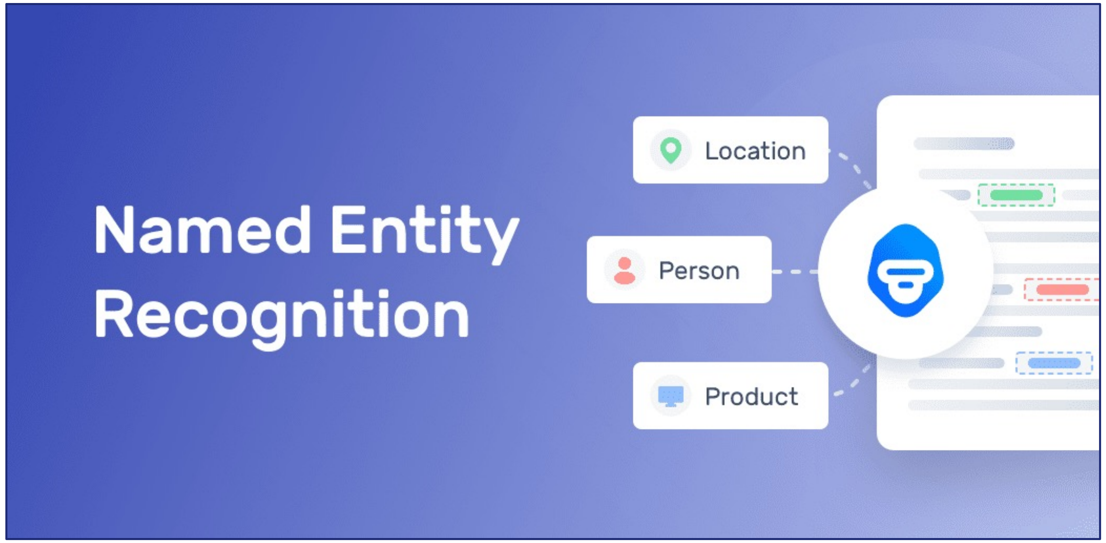



Fig: NER

```
O                      I-ComputingFacility
I-Organization         I-Model
I-Citation             I-Location
I-Formula              B-Survey
B-Organization         I-Archive
B-Citation             I-Collaboration
B-Person               B-Instrument
B-Grant                B-Fellowship
I-Grant                B-ComputingFacility
I-Person               I-Telescope
B-CelestialObject      I-Software
B-Wavelength           I-Survey
B-Formula              B-Collaboration
I-CelestialObject      B-Database
B-Location             I-Database
B-Telescope            B-URL
I-Observatory          B-Archive
B-Model                B-Dataset
I-Wavelength           I-Dataset
I-Fellowship           I-CelestialObjectRegion
B-Observatory          I-Proposal
B-Software
```

Fig: The NER Tags present within the dataset


Fig: WIESP dataset

# Visualizing what the data looks like and how it can be processed

## EDA

- Total tokens/words = 573132
- Total sentences = 1753
- Mean sentence length = 326
- Number of unique NER-Tags = 63

| bibcode | label_studio_id | ner_ids | ner_tags | section | tokens | unique_id |
|---|---|---|---|---|---|---|
| 2019MNRAS.486.5558S | 487 | 62 | O | fulltext | fit | fulltext_487_2019MNRAS.486.5558S |
| 2019MNRAS.486.5558S | 487 | 62 | O | fulltext | uncertainty. | fulltext_487_2019MNRAS.486.5558S |
| 2019MNRAS.486.5558S | 487 | 62 | O | fulltext | Photometric | fulltext_487_2019MNRAS.486.5558S |
| 2019MNRAS.486.5558S | 487 | 62 | O | fulltext | data | fulltext_487_2019MNRAS.486.5558S |
| 2019MNRAS.486.5558S | 487 | 62 | O | fulltext | are | fulltext_487_2019MNRAS.486.5558S |
| 2019MNRAS.486.5558S | 487 | 62 | O | fulltext | from | fulltext_487_2019MNRAS.486.5558S |
| 2019MNRAS.486.5558S | 487 | 17 | B-Mission | fulltext | K2, | fulltext_487_2019MNRAS.486.5558S |
| 2019MNRAS.486.5558S | 487 | 15 | B-Instrument | fulltext | SMEI, | fulltext_487_2019MNRAS.486.5558S |
| 2019MNRAS.486.5558S | 487 | 27 | B-Telescope | fulltext | Hipparcos | fulltext_487_2019MNRAS.486.5558S |

| |
|---|
| bibcode |
| label_studio_id |
| ner_ids |
| ner_tags |
| section |
| tokens |
| unique_id |

Fig: WIESP Dataset & Columns Names

## Pre-processing

- Words about years were replaced with <YEAR> token
- Words containing numbers were replaced with <NUM> token

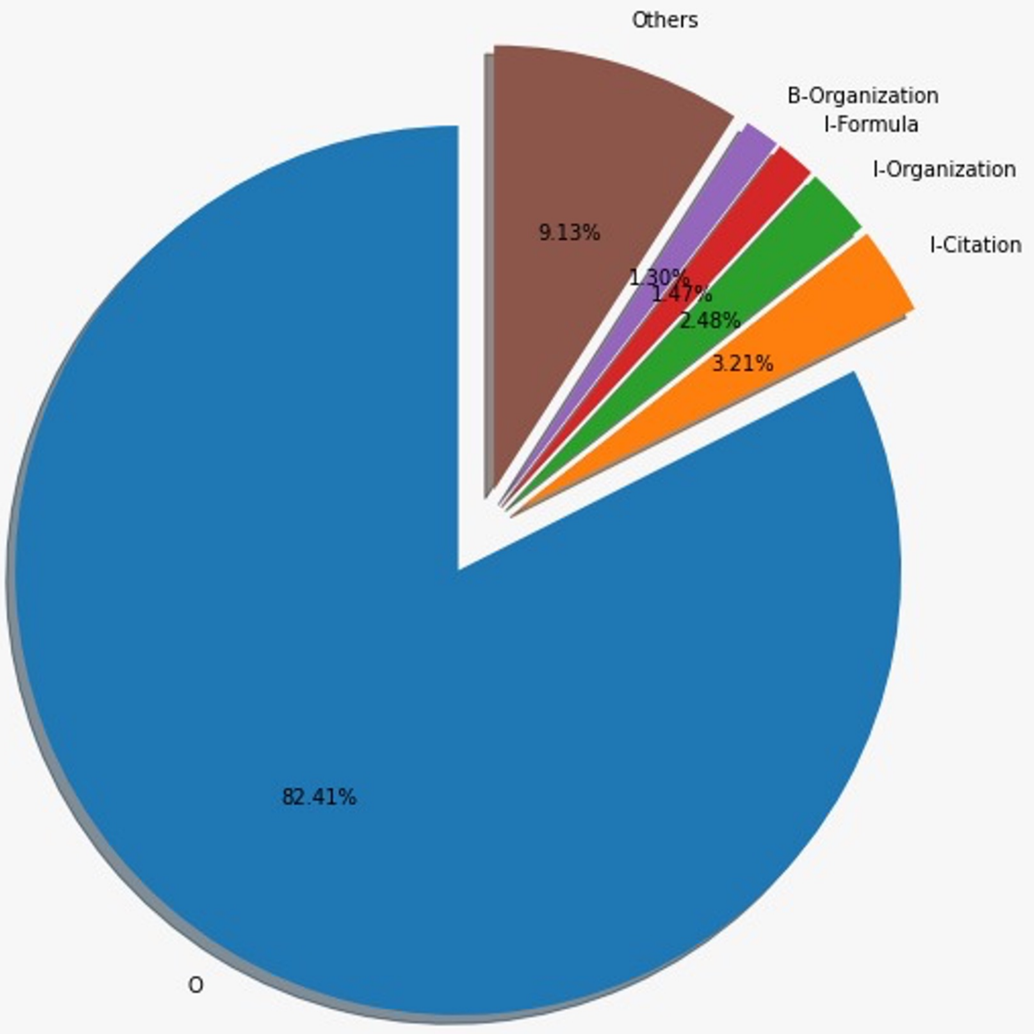



Fig: NER Tags pie-chart

# Developing models for the task and iterating to improve the results

## 1 Stacking SimpleRNN Layers

- Split sentences on '.' for shorter sentence length
- Varied batch size, sentence length, # layers, etc.
- Pre-processing data didn't help the model
- ~92% accuracy; 10% improvement over naive

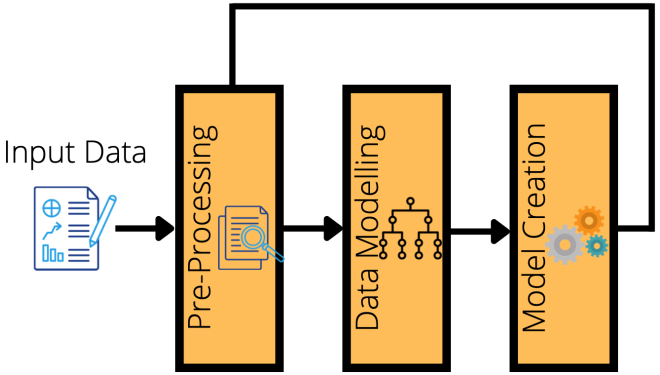



Fig: Basic Architecture for any model for the task

## 1.5 Using Word2Vec Embeddings

Trained embeddings didn't improve the model metrics due to lack of trainable parameters and embeddings not trained for this specific task.

```
## Embedding Layer for forward.
embedding_layer_f = tf.keras.layers.Embedding(input_dim=2000,
                                              output_dim=300,
                                              weights=[embedding_matrix],
                                              trainable=False,    # pre-trained.
                                              mask_zero=True)
```

Fig: Using Word2Vec Embeddings

| Model Type | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Mainframe | Layers | Sentence Length | Bidirectional | Hidden Size | Batch Size | Epochs | Dataset | Loss | Accuracy | Validation Loss | Validation Accuracy |
| SimpleRNN | 3 | 10 to 30 | FALSE | 32 | 8 | 30 | After extra pre-processing | 0.1811 | 0.9378 | 0.2933 | 0.9195 |
| SimpleRNN | 3 | 10 to 30 | FALSE | 32 | 8 | 30 | Without extra pre-processing | 0.2388 | 0.9249 | 0.2897 | 0.9166 |
| SimpleRNN | 3 | 10 to 30 | TRUE | 32 | 8 | 30 | After extra pre-processing | 0.2414 | 0.9228 | 0.287 | 0.9155 |
| SimpleRNN | 3 | 10 to 30 | TRUE | 32 | 8 | 30 | Without extra pre-processing | 0.2332 | 0.9255 | 0.284 | 0.9142 |
| SimpleRNN | 3 | 10 to 30 | TRUE | 32 | 8 | 8 | After extra pre-processing | 0.2031 | 0.9308 | 0.2797 | 0.9169 |
| SimpleRNN | 3 | 10 to 30 | TRUE | 32 | 8 | 10 | Without extra pre-processing | 0.218 | 0.9289 | 0.2844 | 0.9182 |
| SimpleRNN | 3 | 10 to 60 | TRUE | 32 | 8 | 9 | Without extra pre-processing | 0.1352 | 0.9278 | 0.1651 | 0.9191 |

Fig: Results using SimpleRNN architecture with varying parameters

# Developing models for the task and iterating to improve the results

## 2 Baseline: Two LSTMs stacked

- Return all sequences for this seq2seq task
- Longer sentence length allowed; removed outliers
- Spiky plots but change in learning rate not useful
- ~92.5% accuracy even with Word2Vec

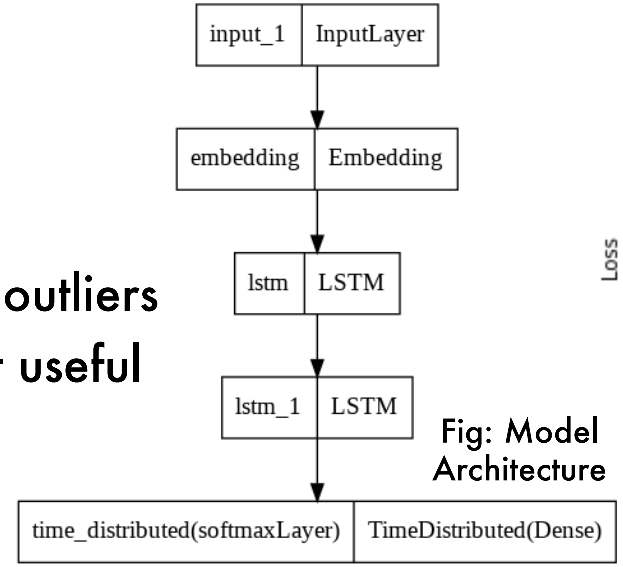



Fig: Model Architecture

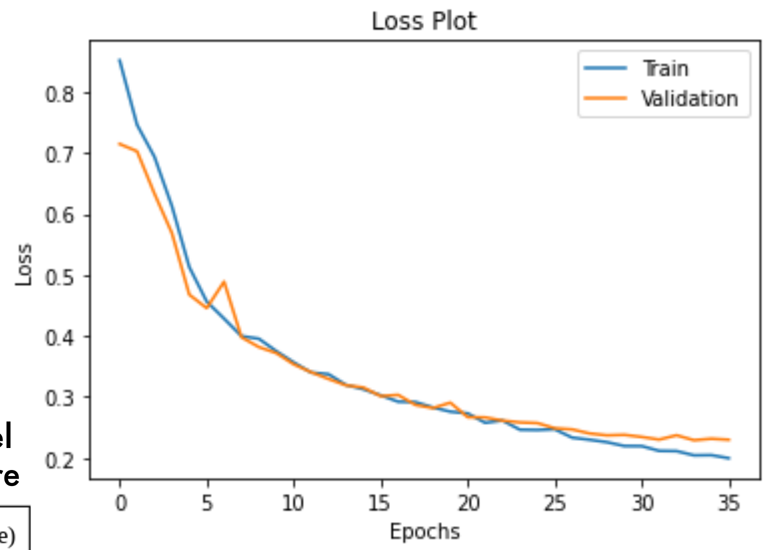
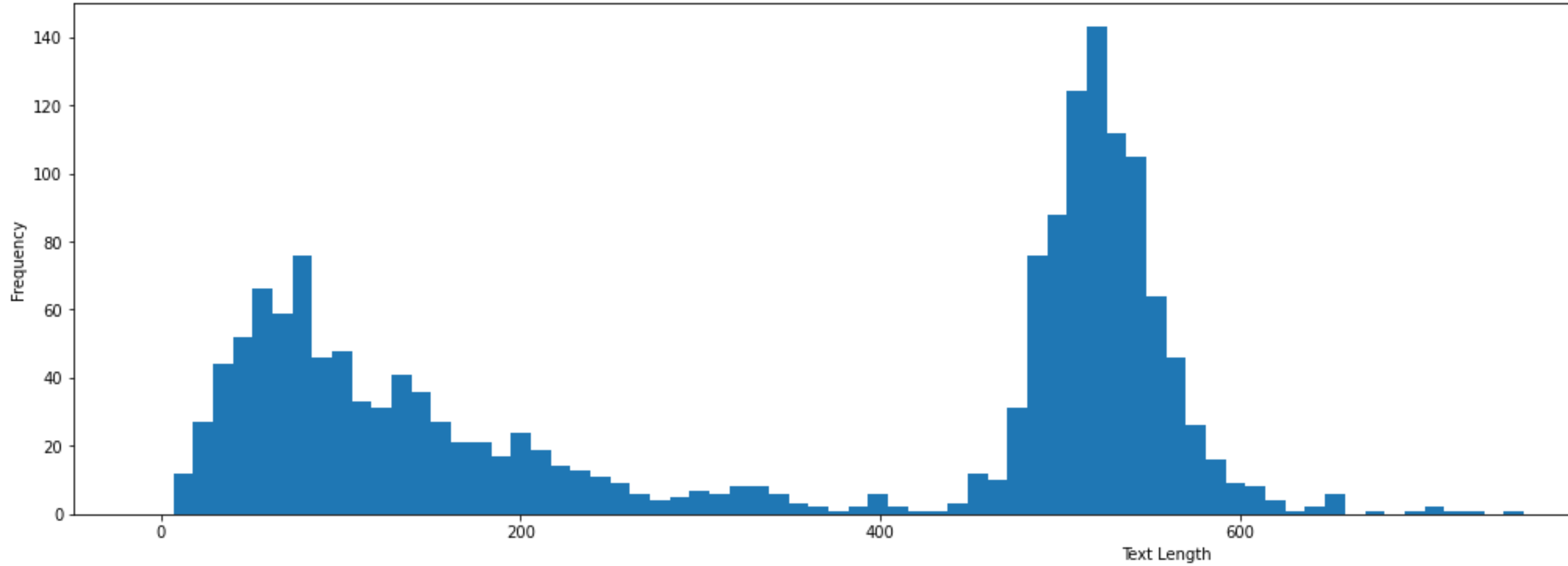



Fig: Histogram for Text Length for each unit in data




Developing models for the task and iterating to improve the results

## 3 Two Bidirectional LSTMs with Dense Layers

- Achieved previous accuracy faster because of bidirectional access
- Learned context over longer sentences; thanks to LSTM
- More trainable parameters improved data fitting
- 95.75% accuracy without any pre-training!

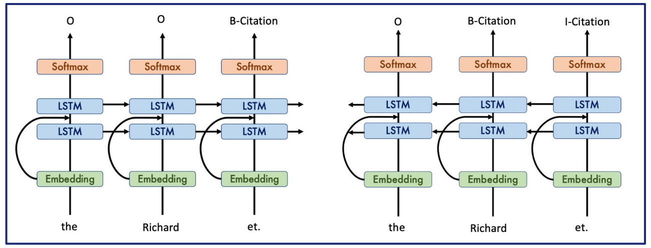
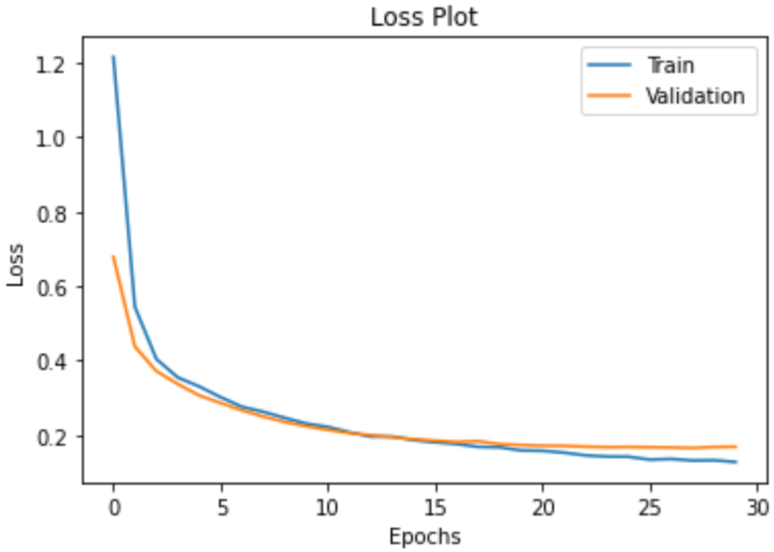
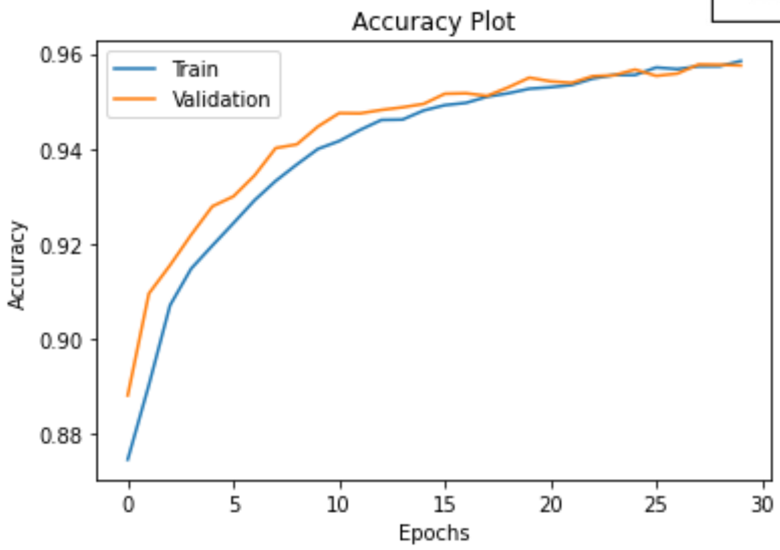



Fig: Understanding Bi-LSTM Model for NER



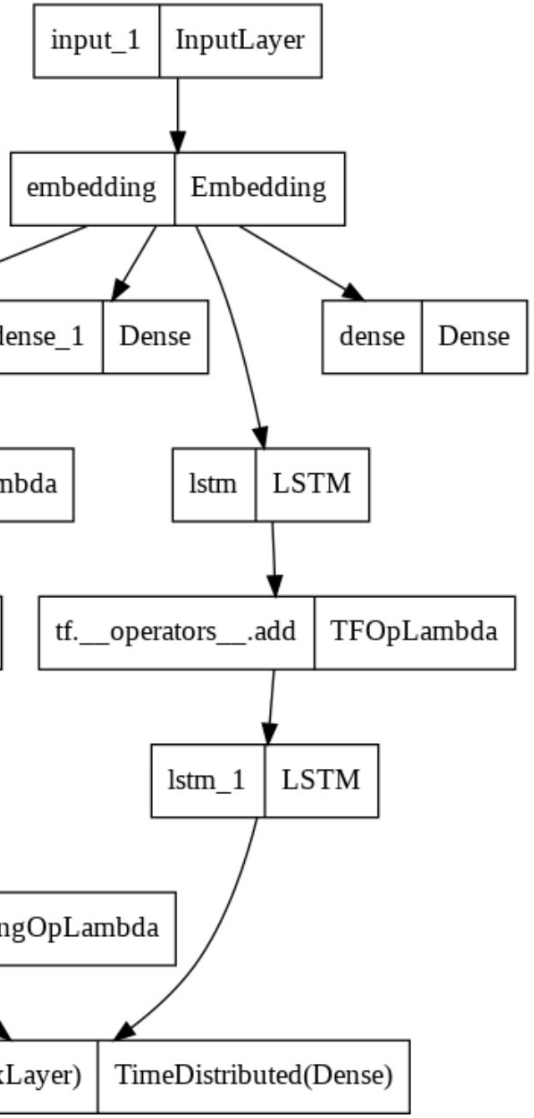




Fig: Model Architecture

Developing models for the task and iterating to improve the results

## 4 SciBERT – A BERT based Model

- Used SimpleTransformers library to custom train SciBERT
- SciBERT is BERT based model pretrained on millions of scientific papers
- 97.2% accuracy with transfer learning framework!
- Other models based on Roberta and Longformers were also tried.
- Accuracy & F1 score was the highest with SciBERT
- BERT Limitation: Supports maximum sequence length of 512.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| weighted avg | 0.97 | 0.97 | 0.97 | 89990.00 |
| macro avg | 0.60 | 0.65 | 0.61 | 89990.00 |
| O | 0.99 | 0.99 | 0.99 | 73164.00 |
| I-Organization | 0.94 | 0.94 | 0.94 | 2787.00 |
| I-Citation | 0.99 | 0.99 | 0.99 | 2743.00 |
| I-Formula | 0.95 | 0.90 | 0.93 | 1532.00 |
| B-Organization | 0.91 | 0.91 | 0.91 | 1383.00 |
| B-Citation | 0.99 | 0.91 | 0.95 | 1020.00 |
| B-Person | 0.98 | 0.99 | 0.99 | 755.00 |
| B-Grant | 0.85 | 0.85 | 0.85 | 645.00 |
| I-Grant | 0.70 | 0.82 | 0.76 | 623.00 |
| I-Person | 1.00 | 0.99 | 0.99 | 474.00 |
| B-CelestialObject | 0.92 | 0.86 | 0.89 | 473.00 |
| B-Wavelength | 0.83 | 0.84 | 0.83 | 434.00 |
| B-Formula | 0.89 | 0.88 | 0.88 | 387.00 |
| I-CelestialObject | 0.95 | 0.93 | 0.94 | 268.00 |
| B-Location | 0.85 | 0.84 | 0.85 | 264.00 |
| B-Telescope | 0.80 | 0.69 | 0.74 | 253.00 |
| I-Observatory | 0.98 | 0.89 | 0.93 | 248.00 |
| B-Model | 0.48 | 0.57 | 0.52 | 247.00 |
| I-Wavelength | 0.83 | 0.77 | 0.80 | 202.00 |
| I-Fellowship | 0.84 | 0.75 | 0.79 | 195.00 |
| B-Observatory | 0.89 | 0.85 | 0.87 | 182.00 |
| B-Software | 0.76 | 0.77 | 0.76 | 146.00 |
| I-ComputingFacility | 0.74 | 0.77 | 0.76 | 145.00 |
| I-Model | 0.36 | 0.54 | 0.44 | 125.00 |
| I-Location | 0.93 | 0.79 | 0.86 | 102.00 |

Fig: Classification Report

Listing the insights gained from the project and noting down areas for further improvement

# Current Insights and Scope of Work

- Right batch size of 4 or 8 allows sufficient training without loosing onto data patterns
- More trainable parameters at right locations allows the model to be more flexible
- Deep Architectures allow for subtle nuances to be explored instead of predicting from a glance
- Transfer Learning allows complex models trained over huge corpus to be extended for specific tasks

- Use of custom loss function to handle class imbalance in dataset
- Increase data on which the model is trained, especially for under-represented NER tags
- Use embeddings trained for Scientific NER like ELMo or other pretrained embeddings

Univ.AI

# Named Entity Recognition (NER)

Rachit Jain, Anshul Bhardwaj



## ABOUT

The project focuses on creating a NER model that identifies key tokens and classifies them into set of predefined entities. The data would involve scientific publications in the **WIESP Dataset**.

NER helps us extract key information from scientific papers which can help search engines to better select and filter articles.

## MODELS & RESULTS

The basic architecture of any model remains the same. What defines its success is how well it is put into use!

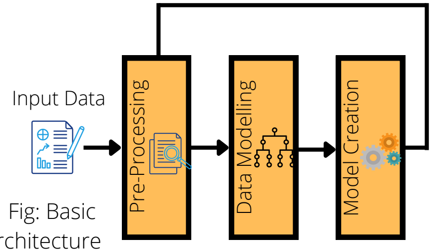



Fig: Basic Architecture

**USING SIMPLE RNN** — Accuracy maxed out at **92%** with multiple design changes

**USING WORD2VEC** — The trained embeddings **do not do well** for our task

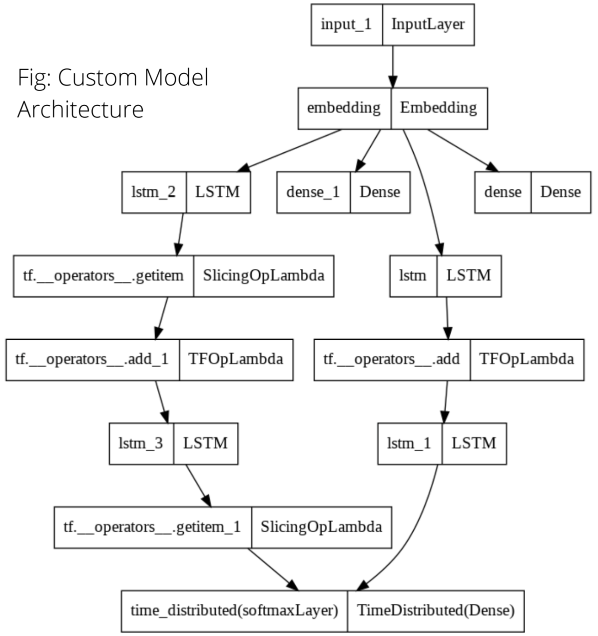
**BASELINE: 2 LSTM STACKED** — Improvement over RNNs; **92.5%** accuracy after fine-tuning

| Model Type | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Mainframe | Layers | Sentence Length | Bidirectional | Hidden Size | Batch Size | Epochs | Dataset | Loss | Accuracy | Validation Loss | Validation Accuracy |
| SimpleRNN | 3 | 10 to 30 | FALSE | 32 | 8 | 30 | After extra pre-processing | 0.1811 | 0.9378 | 0.2933 | 0.9195 |
| SimpleRNN | 3 | 10 to 30 | FALSE | 32 | 8 | 30 | Without extra pre-processing | 0.2388 | 0.9249 | 0.2897 | 0.9166 |
| SimpleRNN | 3 | 10 to 30 | TRUE | 32 | 8 | 30 | After extra pre-processing | 0.2414 | 0.9228 | 0.287 | 0.9155 |
| SimpleRNN | 3 | 10 to 30 | TRUE | 32 | 8 | 30 | Without extra pre-processing | 0.2332 | 0.9255 | 0.284 | 0.9142 |
| SimpleRNN | 3 | 10 to 30 | TRUE | 32 | 8 | 8 | After extra pre-processing | 0.2031 | 0.9308 | 0.2797 | 0.9169 |
| SimpleRNN | 3 | 10 to 30 | TRUE | 32 | 8 | 10 | Without extra pre-processing | 0.218 | 0.9289 | 0.2844 | 0.9182 |
| SimpleRNN | 3 | 10 to 60 | TRUE | 32 | 8 | 9 | Without extra pre-processing | 0.1352 | 0.9278 | 0.1651 | 0.9191 |

Fig: Different experiments with basic model architectures by tuning parameters

### 2 BIDIRECTIONAL LSTM LAYERS

Fig: Custom Model Architecture




This allowed model to **learn context over long sentences** and have enough trainable parameters to increase its robustness. Accuracy rose to **95.75%** without any pre-training system

### BERT-BASED MODEL

After multiple BERT versions, the best accuracy of **97.2%** came on the model based on **SciBERT**: A Pretrained Language Model for Scientific Text

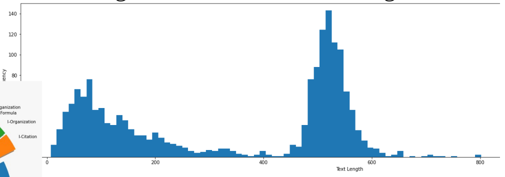
## DATA & LABELS

1. **bibcode**: Can be used to extract the texts.
2. **label_studio_id**: Can also be used to extract texts.
3. **ner_ids**: Can be used as label encoded values for the ner_tags.
4. **section**: Two types: fulltext or acknowledgement.
5. **tokens**: The units whose entities need to be found by the model.
6. **unique_id**: This could also be used to separate out units of texts.

Fig: Raw Data Representation

| bibcode | label_studio_id | ner_ids | ner_tags | section | tokens | unique_id |
|---|---|---|---|---|---|---|
| 2019MNRAS.486.5558S | 487 | 62 | O | fulltext | fit | fulltext_487_2019MNRAS.486.5558S |
| 2019MNRAS.486.5558S | 487 | 62 | O | fulltext | uncertainty. | fulltext_487_2019MNRAS.486.5558S |
| 2019MNRAS.486.5558S | 487 | 62 | O | fulltext | Photometric | fulltext_487_2019MNRAS.486.5558S |
| 2019MNRAS.486.5558S | 487 | 62 | O | fulltext | data | fulltext_487_2019MNRAS.486.5558S |
| 2019MNRAS.486.5558S | 487 | 62 | O | fulltext | are | fulltext_487_2019MNRAS.486.5558S |
| 2019MNRAS.486.5558S | 487 | 62 | O | fulltext | from | |
| 2019MNRAS.486.5558S | 487 | 17 | B-Mission | fulltext | K2, | |
| 2019MNRAS.486.5558S | 487 | 15 | B-Instrument | fulltext | SMEI, | |
| 2019MNRAS.486.5558S | 487 | 27 | B-Telescope | fulltext | Hipparcos | |
| 2019MNRAS.486.5558S | 487 | 62 | O | fulltext | ( | |
| 2019MNRAS.486.5558S | 487 | 30 | B-Wavelength | fulltext | H | |
| 2019MNRAS.486.5558S | 487 | 61 | I-Wavelength | fulltext | p | |

Fig: Extracted Sentence Lengths



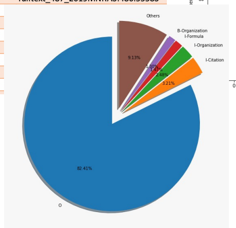



Fig:Distribution of NER Tags in the dataset

## CONCLUSION

Right Batch Size, Early Stopping, More Trainable Parameters, Deep Architectures and Transfer Learning truly boost the performance of models.

## SCOPE OF IMPROVEMENT

1. Class Imbalance: use custom loss function
2. More data for under-represented tokens
3. Use embeddings trained for Scientific NER

# Project NER

*19th June 2022*

## Team: Towards_NLP

Rachit Jain Anshul Bhardwaj
NLP Scholars, AI-3 C-2
Univ.AI

## Thank You!