Univ.AI

Name Entity Recognition (NER)

Task Force X : Abhishek, Ishani, Shruti and Tharun

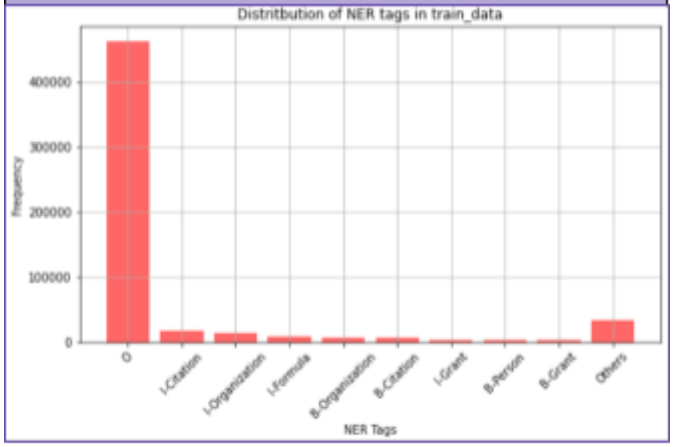
Motivation and About the Project

It is difficult to navigate in this ocean of data; finding articles rely heavily on string matching searches and following citations/references. NER helps us extract key information from scientific papers which can help search engines to better select and filter articles.

The project revolves around Name Entity Recognition (NER) on WIESP dataset to extract key information from scientific papers. It has text fragments from astrophysics papers with manually tagged astronomical facilities and other entities of interest. The train data has 1753 sentences having NER tags and ids for each sentence tokens.

Data and Labels

Huge imbalance in dataset (majority of 'O' tags)

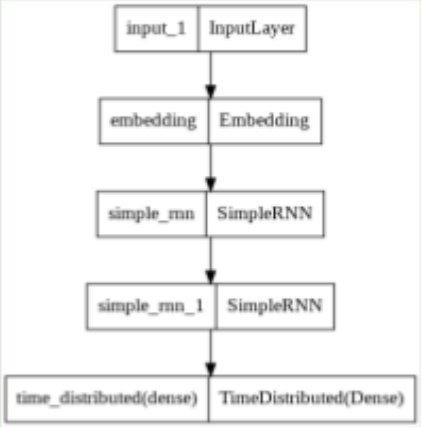


References

[1].https://towardsdatascience.com/named-entity-recognition-ner-using-keras-bidirectional-lstm-28cd3f301f54

Model

Baseline: Simple RNN layers

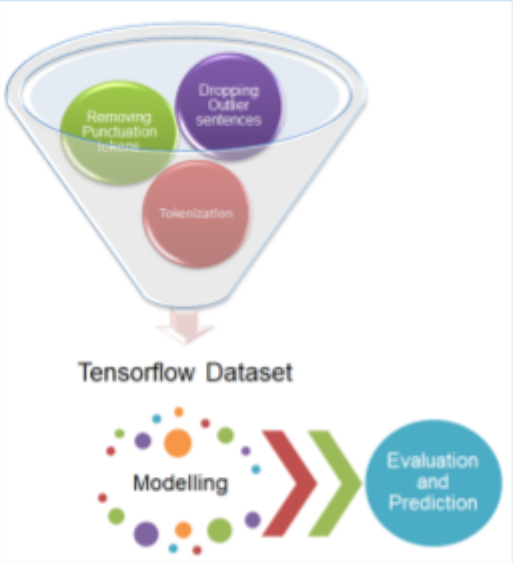


1. Embeddings. from scratch
2. Word2Vec Embeddings

Bidirectional LSTM Model

1. Embeddings. from scratch
2. Word2Vec Embeddings

Methodology

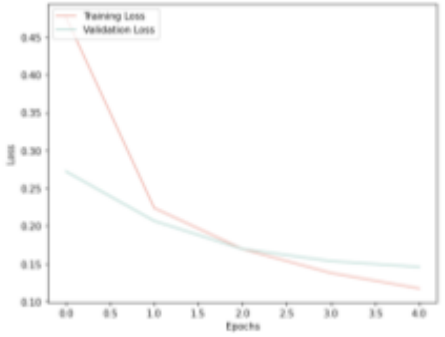


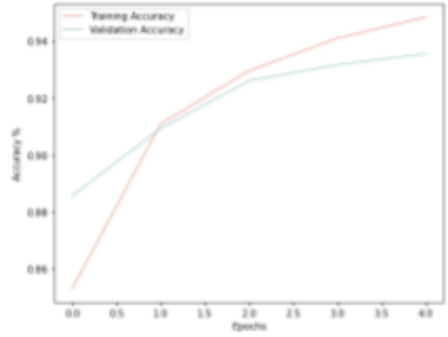
Results

Comparison of models

Model	Train Accuracy	Validation Accuracy
Baseline Model (2 SimpleRNN layers)	0.936	0.916
Bidirectional LSTM model	0.954	0.936
Baseline Model (Using Word2Vec Embeddings)	0.885	0.873
Bidirectional LSTM (Using Word2Vec Embeddings) model	0.908	0.895

Accuracy and Loss plots for Best Model





Conclusion

- Bidirectional LSTM model having train accuracy of 95.4% and validation accuracy of 93.6%.
- As there is class imbalance in our dataset, we would chose F1-Score as evaluation metric
- For the other tags which are higher in counts after 'O' also have high F1-scores indicating that our model would predict entities well rather than just predicting 'O' tag.

Future Work

- Improving the F1_scores of rest of the tags by getting more data (tokens) of the tags which are less in the current dataset
- Use Transformers: Bert models
- Changing the model hyperparameters like epochs and batch size.