Univ.AI

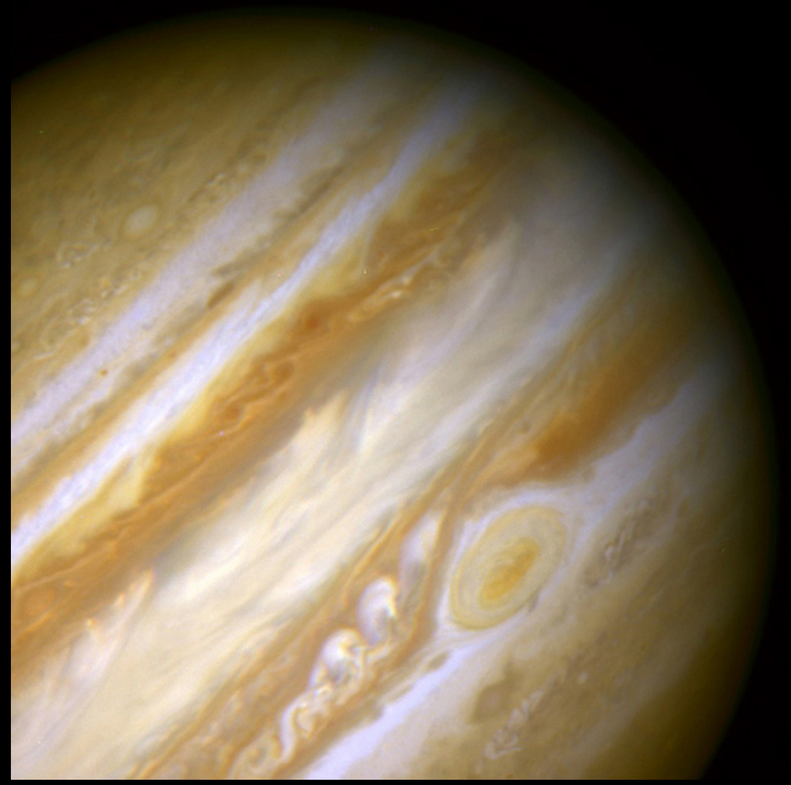
# NAMED ENTITY RECOGNITION

*Information Extraction From Scientific Publication*
WIESP -2022

Submitted to: Prof. Pavlos Protopapas

Submitted by:
Ajinkya, Hemani, Rohit, Shibani

# PROBLEM STATEMENT



- As the number of research papers being published have become very high, the problem to search for the relevant papers has become significant.

- With this project, our aim is to extract the key information from scientific papers to better select and filter the articles on the search engines.

# WIESP 2022

## OBSERVATIONS

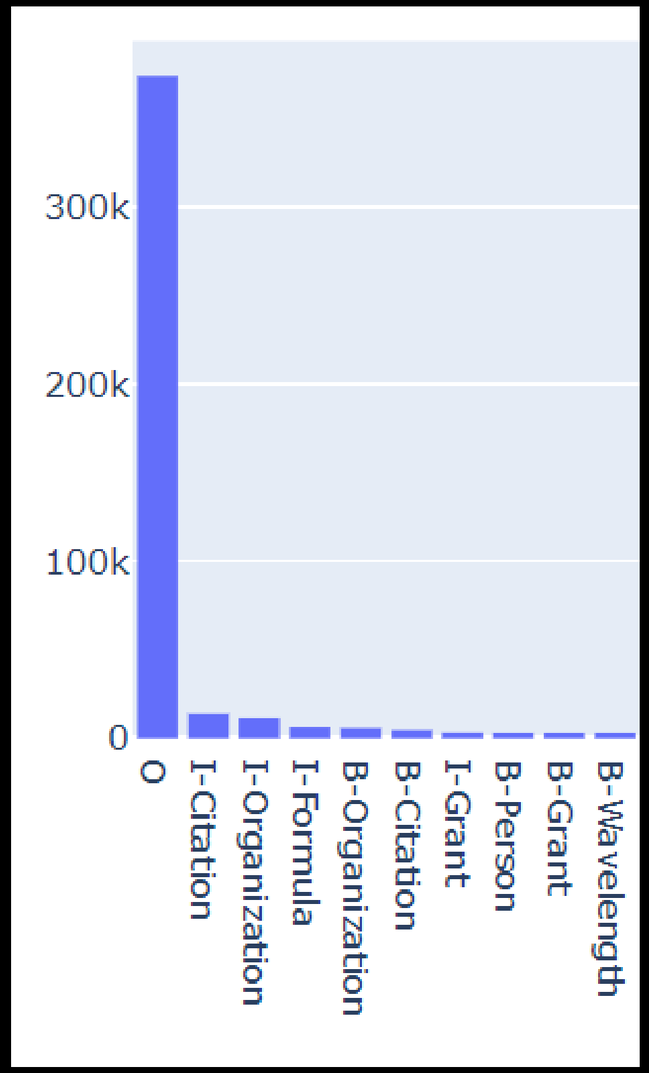
- Around 83% of the tags are 'O' tags comprising of 373,897 occurrences.
- The length of texts range from 7 to 795.
- Many tags have very few occurrence, for eg. - I-URL has just 2 occurrences in the whole dataset.

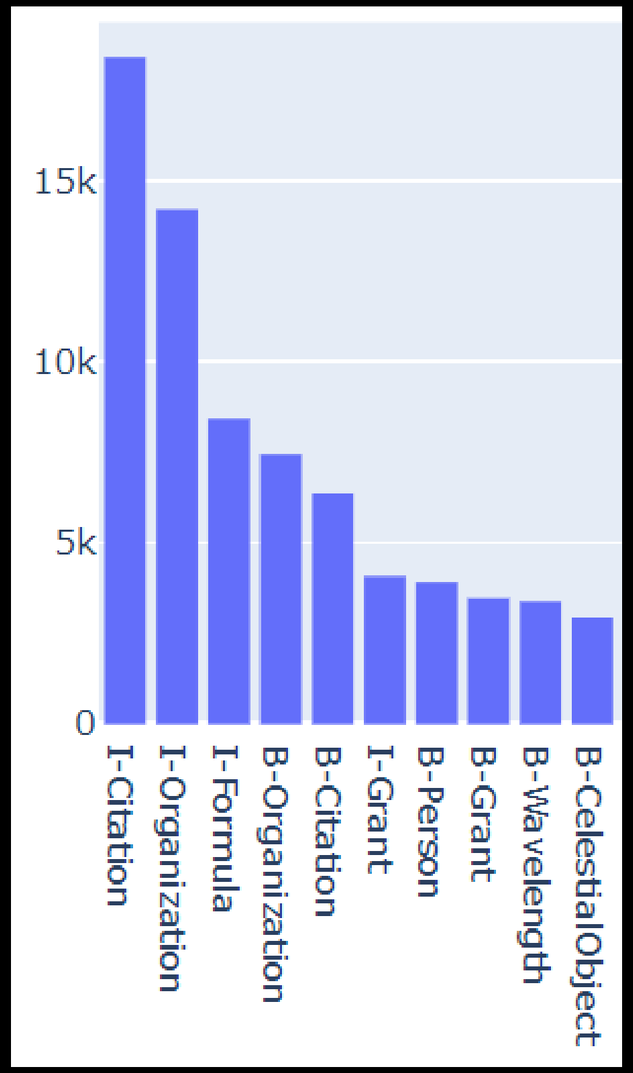
## Dataset Description

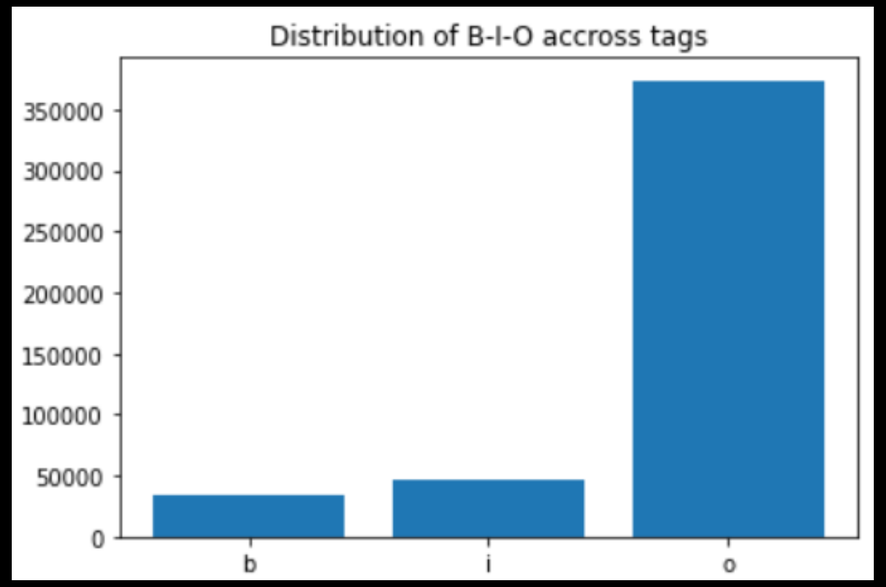
- Datasets with text fragments from astrophysics papers, provided by the NASA Astrophysical Data System with manually tagged astronomical facilities and other entities of interest
- Examples : [BDM, gratefully, acknowledges, support, from,...]
- IOB2 syntax : [B-Person, O, O, O, O, B-Organization, B-Grant...]

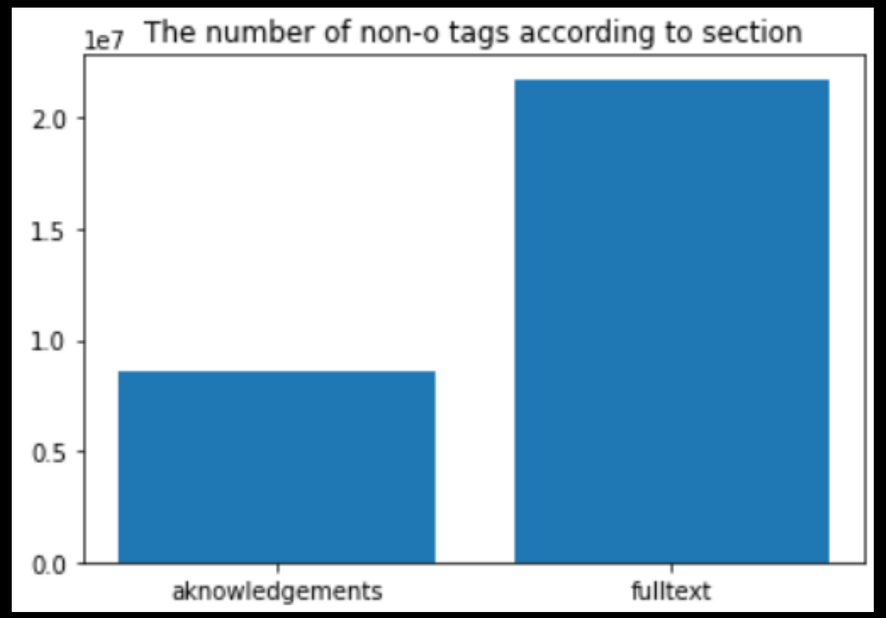
# EDA ON THE WIESP DATASET

## Comparison of O tags with other tags

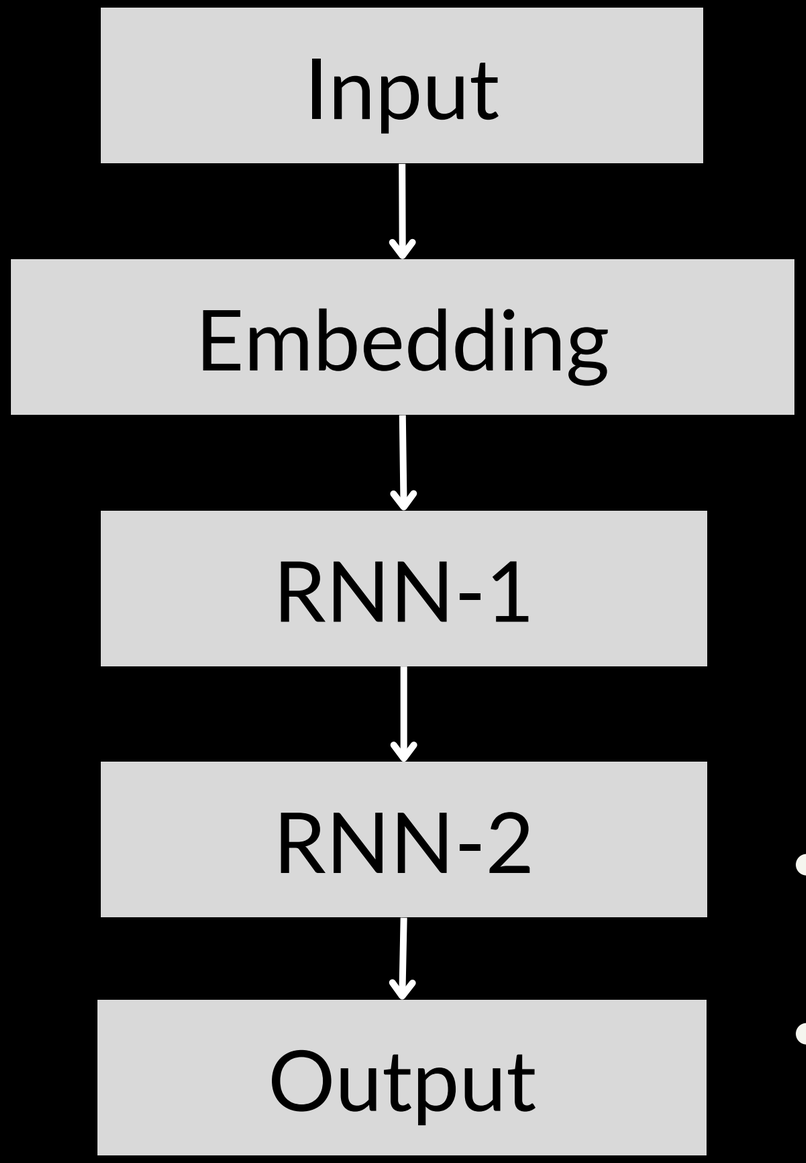


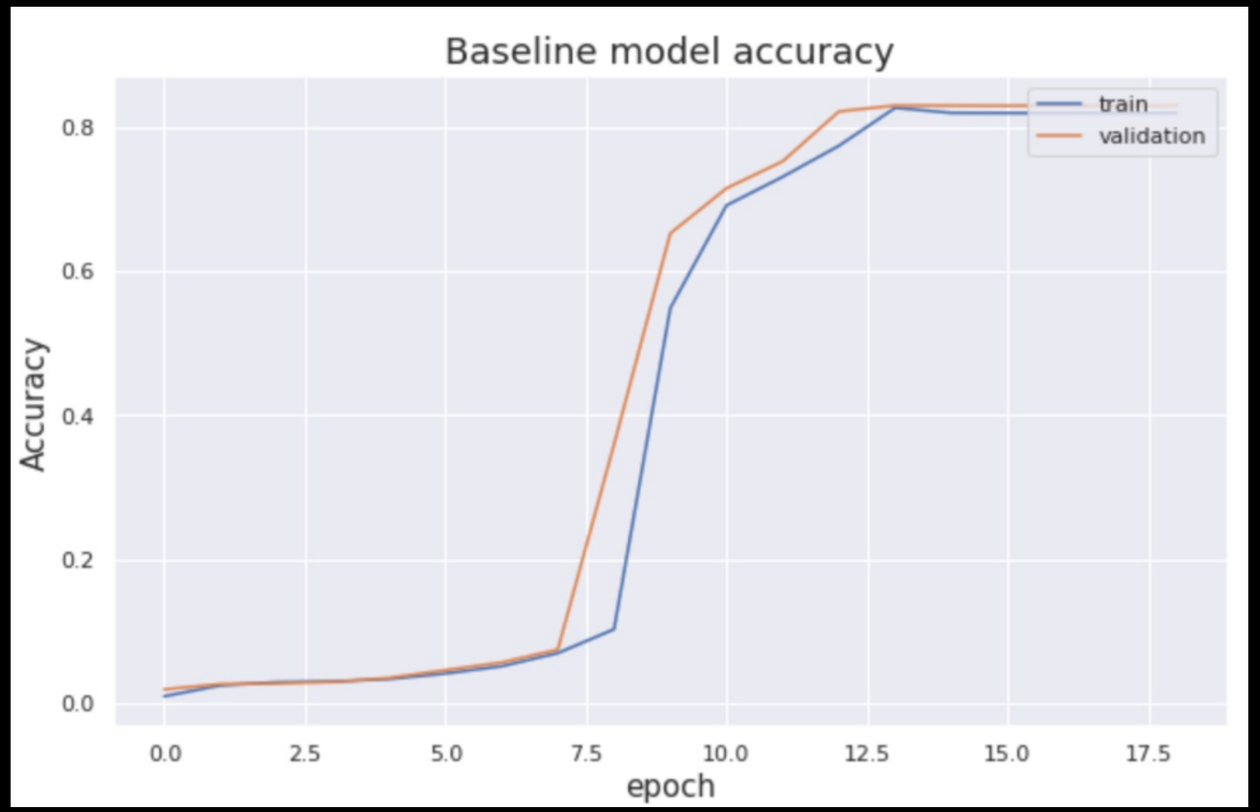



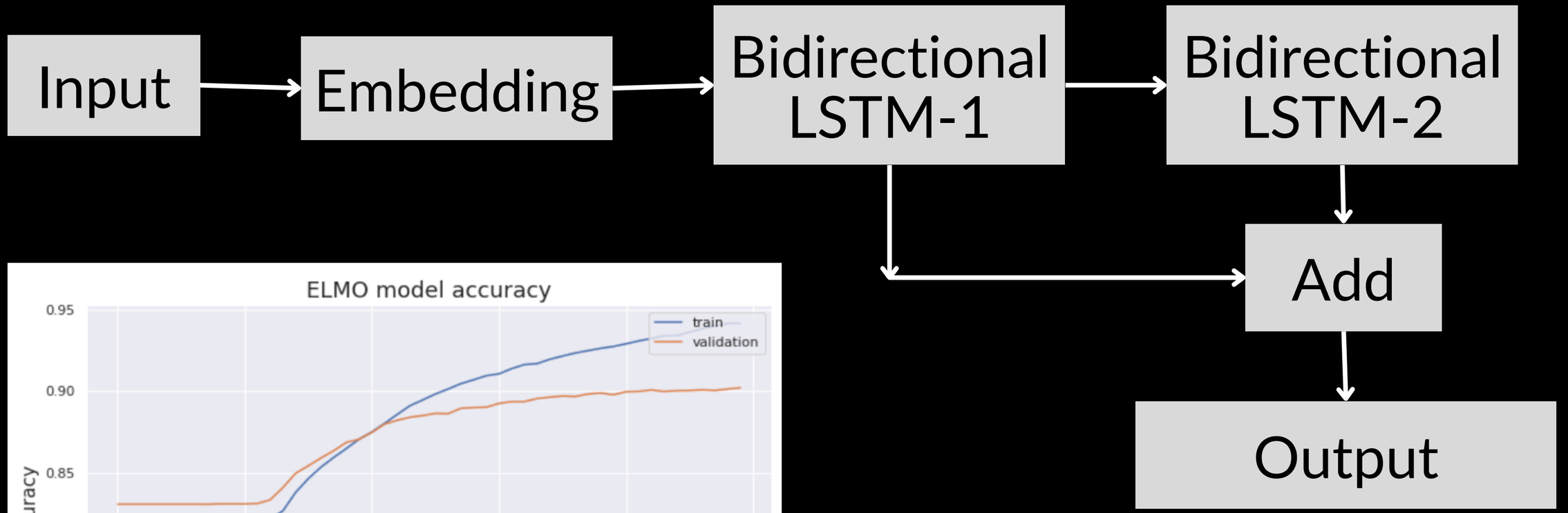
# BASELINE MODEL




- The accuracy of the baseline model converged at 83% for validation set, and 82% for training set.
- On checking, we got that it was only predicting 'O's, with a few exceptions at certain times.

# ELMO MODEL

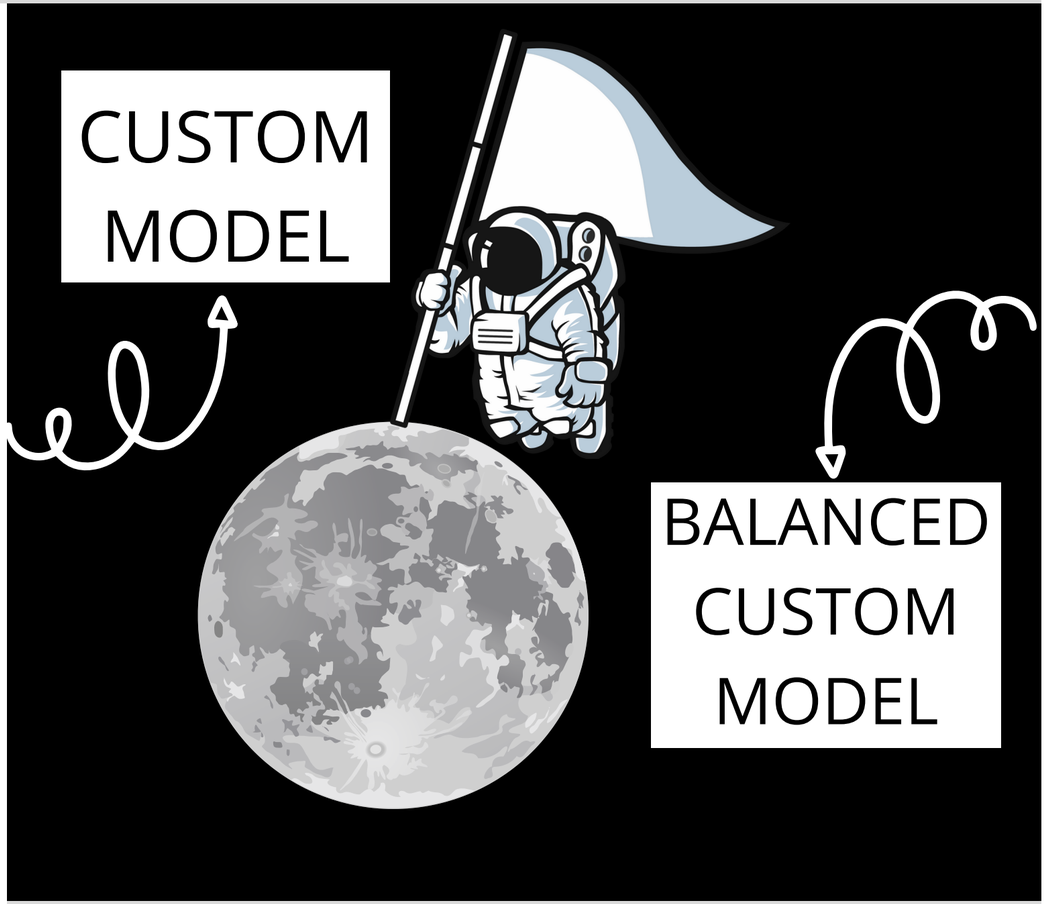






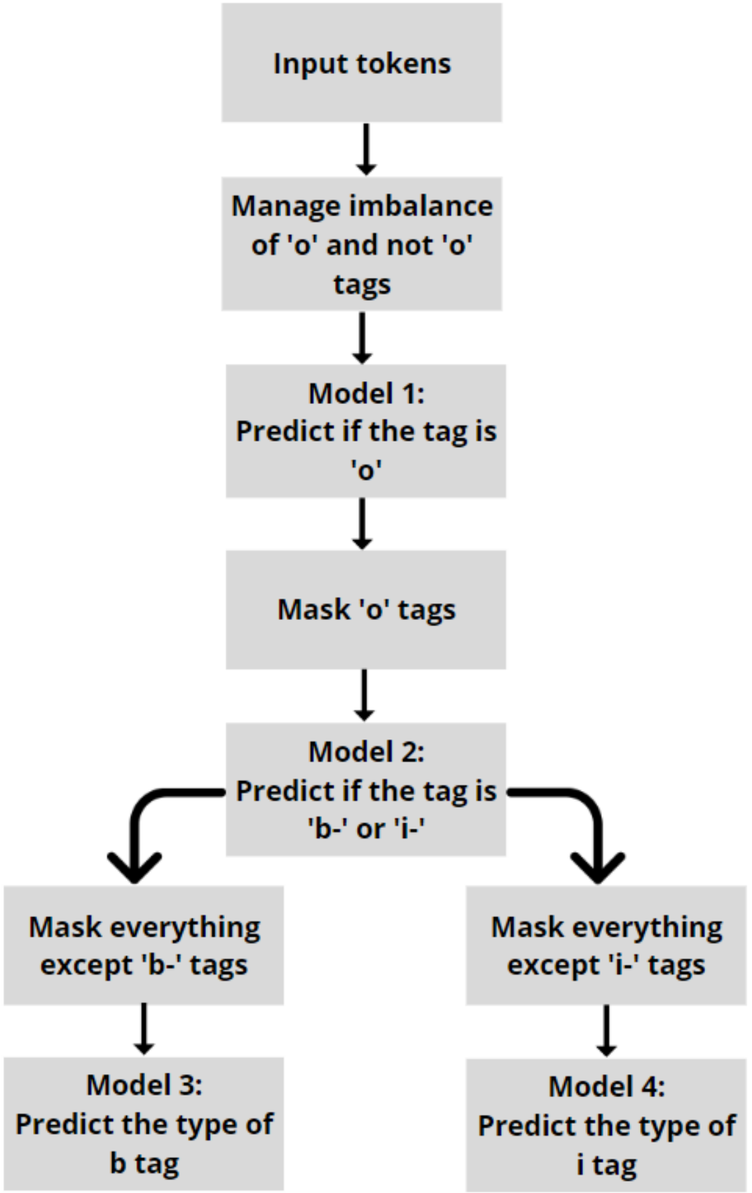
- The accuracy of the ELMO model converged at around 90% for validation set, the learning had become very slow after reaching 88%.

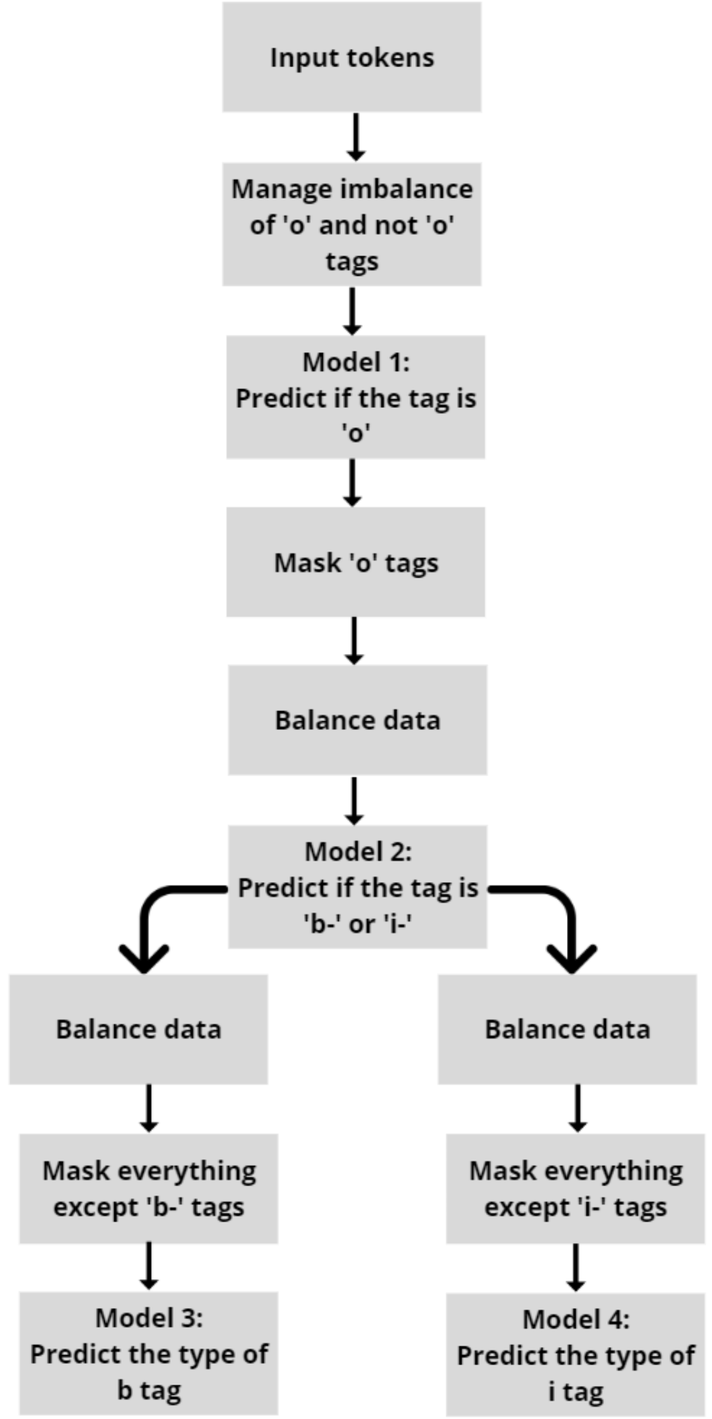
# CUSTOM MODEL & BALANCED CUSTOM MODEL




**Accuracy:**
**Custom_model: 83%**
**(predicting 26 classes)**

**Balanced_Custom_model:84%**
**(predicting 7 classes)**



# CONCLUSIONS AND FUTURE WORK

## Conclusions

- We explored and looked into the difference the predictions that we got after seeing the ELMO model which had context incorporated into it, as compared to the baseline model.
- We also examined the results that we got with the whole data, and when the 'O' tags were balanced to make the frequency comparable.

## Future Work

- To work on the custom model and the balanced custom model for better predictions. We think that there is a lot of scope for improvement, if the combination works fine.
- To improve the autocorrect that we tried to implement on this corpus, and extend it to a larger corpus.
- Investigate more about the contextual embeddings, and to what level they play a role in the predictions.
- Explore better ways to deal with the extensively large number of 'O' tags.


THANK YOU !!!