Shallow Parsing (Chunking)

Team 6:
Shreya Pathak
Neel Aryan Gupta
Mohammad Ali Rehan

Features used in MEMM

The features used included:

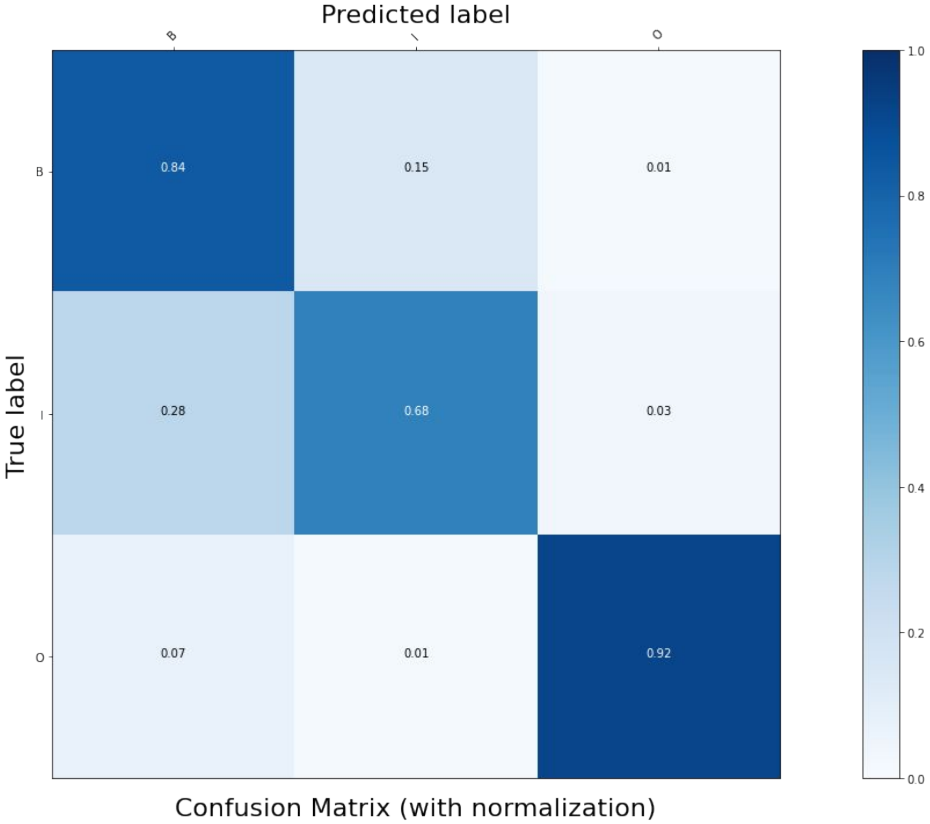
- POS tags with a window of 5 around current word
- Glove vectors with a window of 5 around current word
- Some common suffixes
- Bits to identify if the word is the first or last word in the sentence
- Bit to indicate capitalisation
- Chunking tag of the previous word

Description of MEMM model

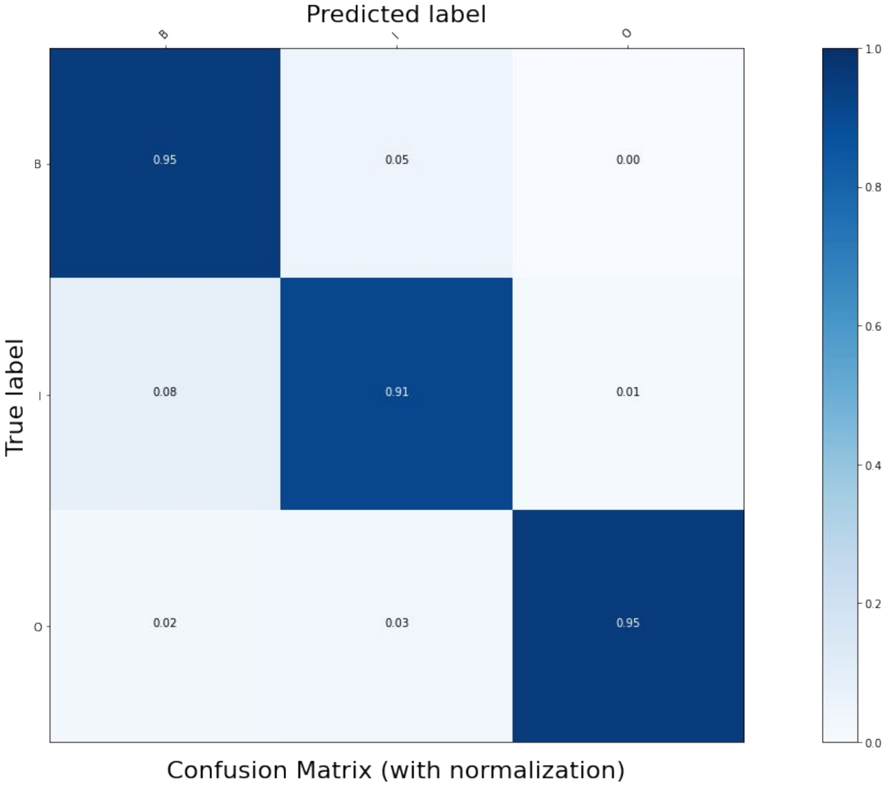
As has been proved in LogisticRegressionMaxEntropy, logistic regression and maximum entropy models are equivalent. Thus, we use a simple logistic regression model to calculate the weights corresponding to the different features.

We then use these features+weights to calculate transition probabilities for our MEMM on which we then apply the Viterbi algorithm to obtain the most probable sequence of the BIO tags

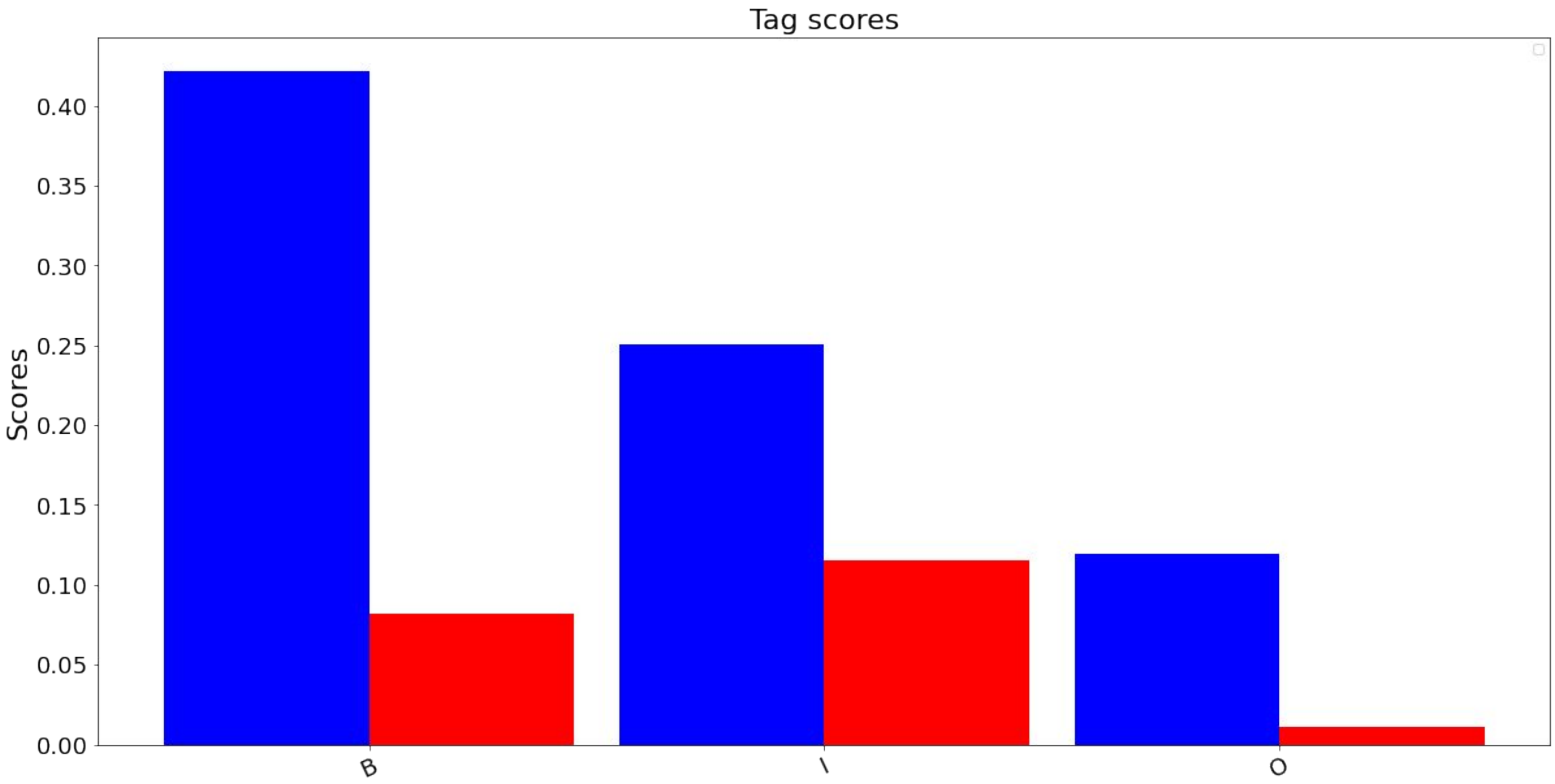
Confusion matrix without POS tags



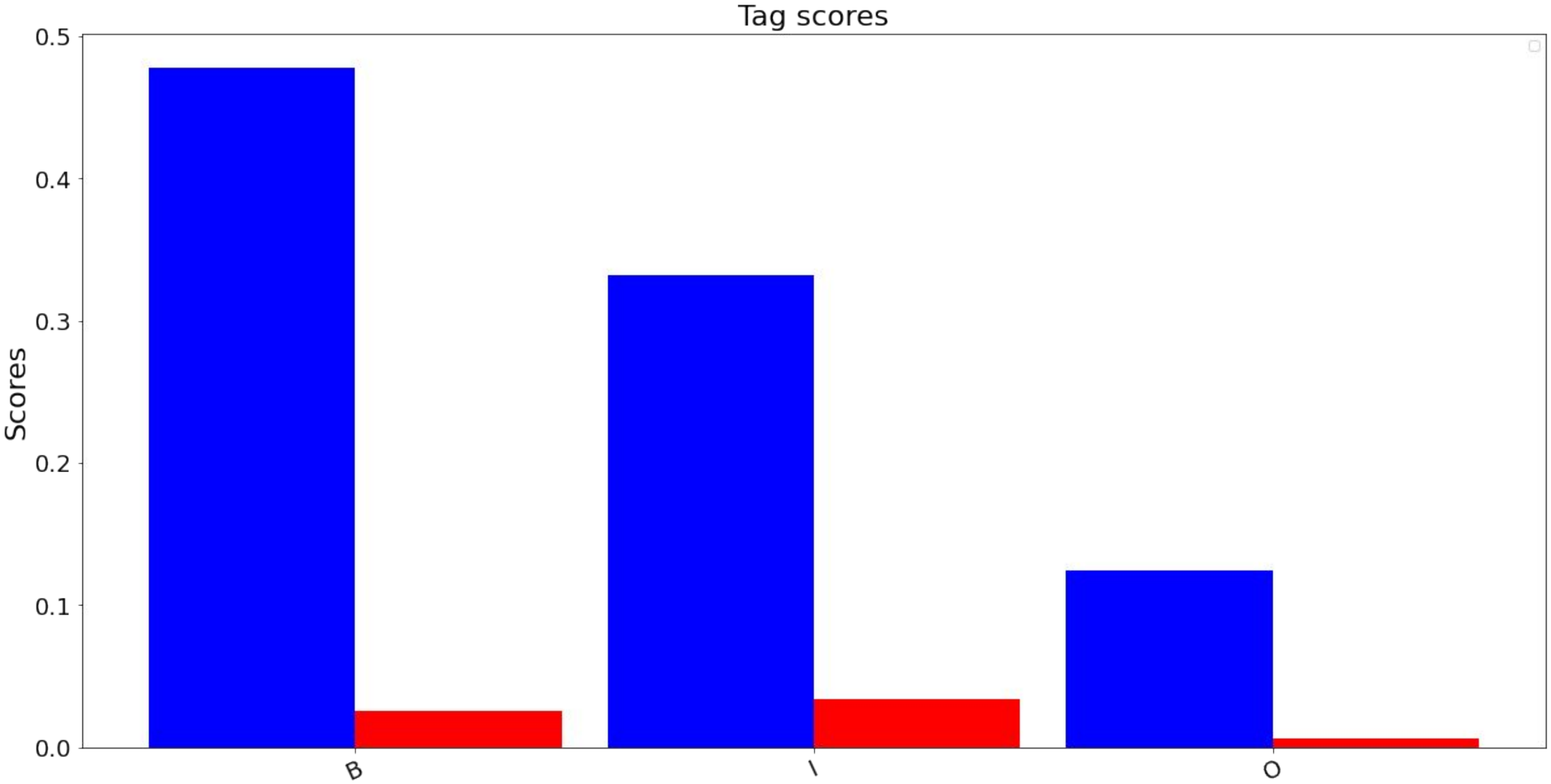
Confusion matrix with POS tags



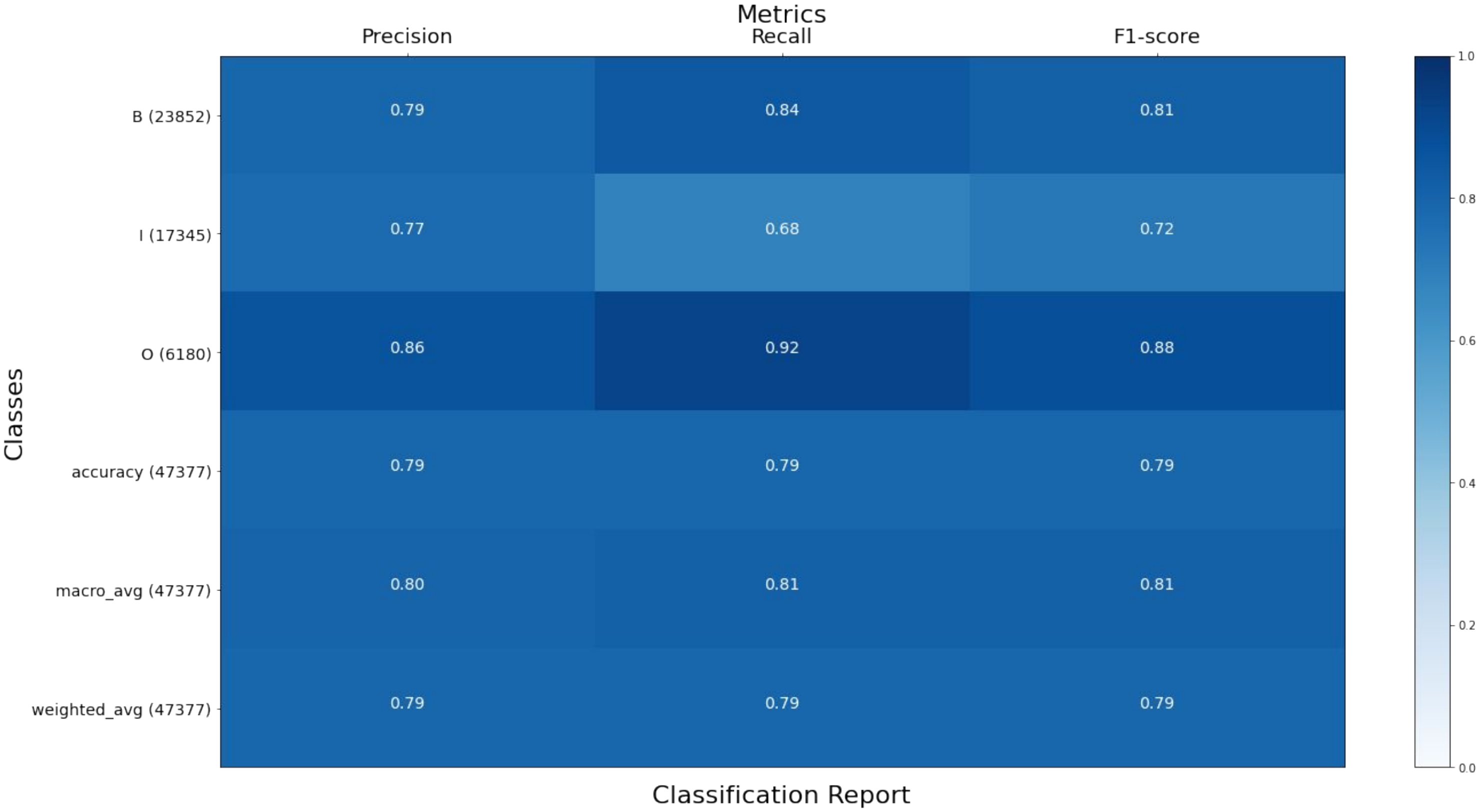
Tag scores without POS tags



Tag scores with POS tags

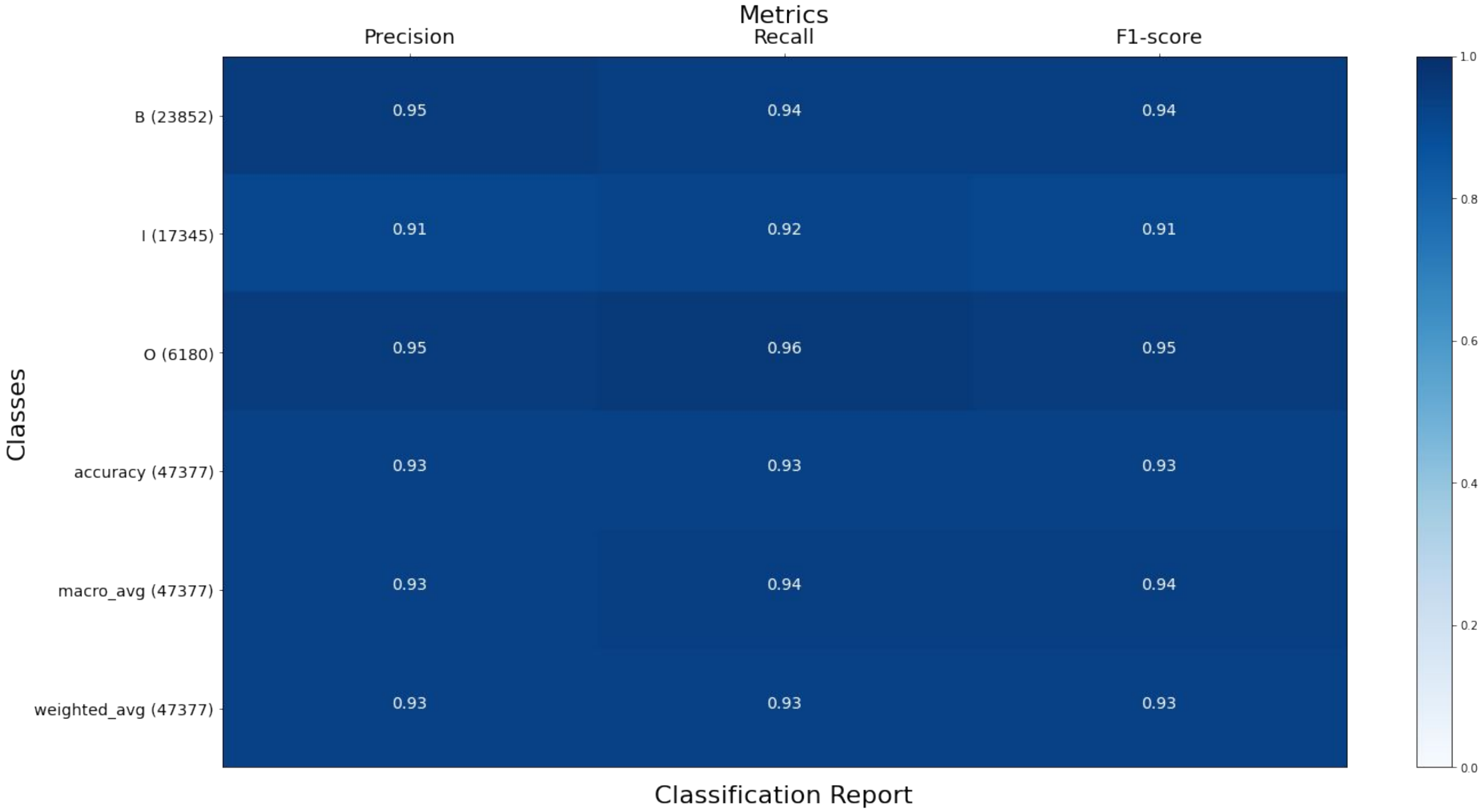


Classification report without POS tags



Classes	Precision	Recall	F1-score
B (23852)	0.79	0.84	0.81
I (17345)	0.77	0.68	0.72
O (6180)	0.86	0.92	0.88
accuracy (47377)	0.79	0.79	0.79
macro_avg (47377)	0.80	0.81	0.81
weighted_avg (47377)	0.79	0.79	0.79

Classification report with POS tags



Classes	Precision	Recall	F1-score
B (23852)	0.95	0.94	0.94
I (17345)	0.91	0.92	0.91
O (6180)	0.95	0.96	0.95
accuracy (47377)	0.93	0.93	0.93
macro_avg (47377)	0.93	0.94	0.94
weighted_avg (47377)	0.93	0.93	0.93

Error Analysis (without POS)

- Mistagging due to punctuations being used within a chunk. The model mistakes them for 'O' tag. Eg:
    a. , electronics , (correct: I I I predicted: O B O)
    b. , Ill. , (correct: I I I predicted: O B O)
- Numbers inside chunks were often misclassified. The model mistakes non-words as 'B' tags as that is a common occurrence in English. Eg,
    a. about 60\%-held by (correct: B I B predicted: B B B)
    b. 's 747 jetliners (correct: B I I predicted: B B I)

- Conjunctions and their following word (within chunks) were often mistagged. Conjunctions were mistagged as 'O' and due to this the following word was considered as a 'B' tag. Eg,
    a. investment and asset-management (correct: I I I predicted: I O B)
    b. automotive and graphics (correct: I I I predicted: B O B)
- Multiple Proper nouns occurring together (as in a name) were mistagged as separate chunks. The model was unable to group them together as they are less common words. Eg,
    a. Messrs. Wolf and (correct: B I I predicted: B B O)
    b. Van Pell , (correct: B I O predicted: B B O)

Error Analysis (with POS)

- Mistagging due to punctuations being used within a chunk. Again the model mistakes them to be 'O' tagged. Eg,
  a. , electronics , (correct: I I I predicted: O B O)
  b. , Ill. , (correct: I I I predicted: O B O)
- Conjunctions and their following word (within chunks) were often mistagged. Eg,
  a. loan and real (correct: I I I predicted: I O B)
  b. Wolf and Pope (correct: I I I predicted: I O B)

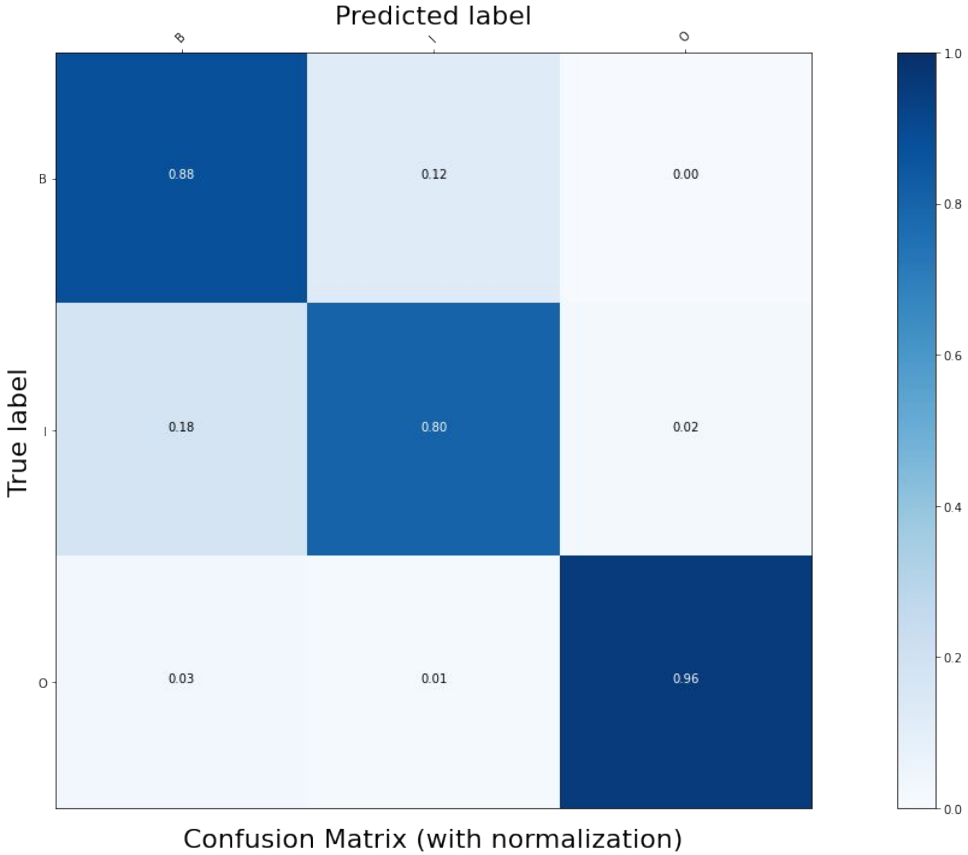
- The word 'to' occuring in a chunk is often mistagged when it is present within a chunk as it is present as a ‘B’ tag in considerably more examples in the training data. Eg,
  1. continue to plummet (correct: B I I predicted: B B I)
  2. stood to gain (correct: B I I predicted: B B I)
- When a predominantly common noun is used as part of a proper noun chunk, eg, US Facilities closed (correct: B I B predicted: B B B)

Features used in CRF

The features used included:

- POS tags with a window of 5 around current word
- Glove vectors with a window of 5 around current word
- The suffix as obtained using porter stemmer
- Bits to identify if the word is the first or last word in the sentence

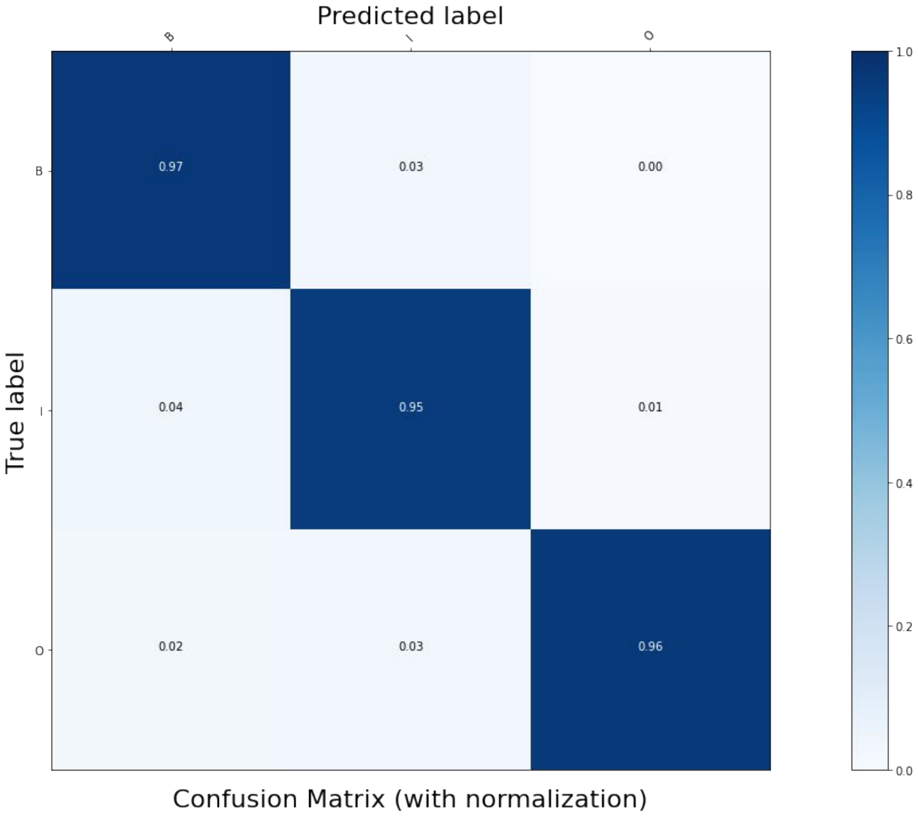
Confusion matrix without POS tags



	B	I	O
B	0.88	0.12	0.00
I	0.18	0.80	0.02
O	0.03	0.01	0.96

Confusion Matrix (with normalization)

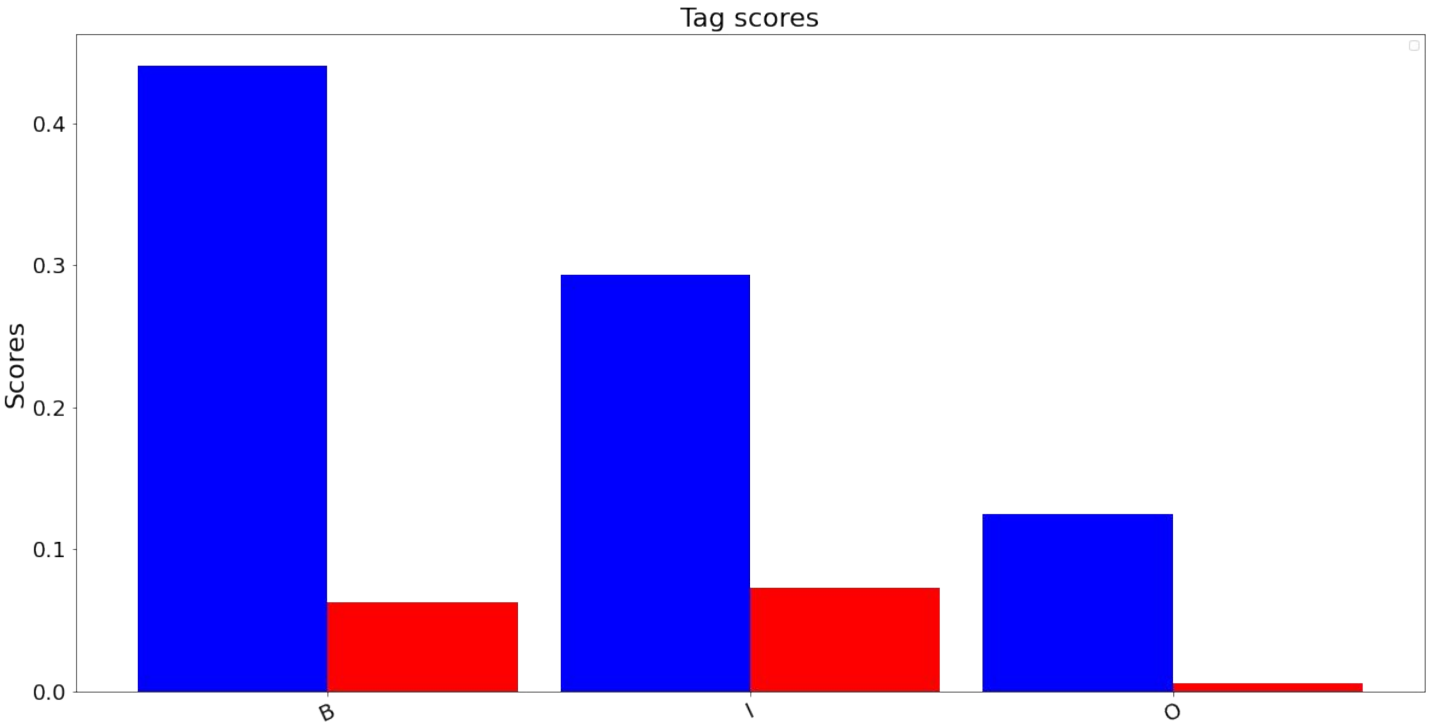
Confusion matrix with POS tags



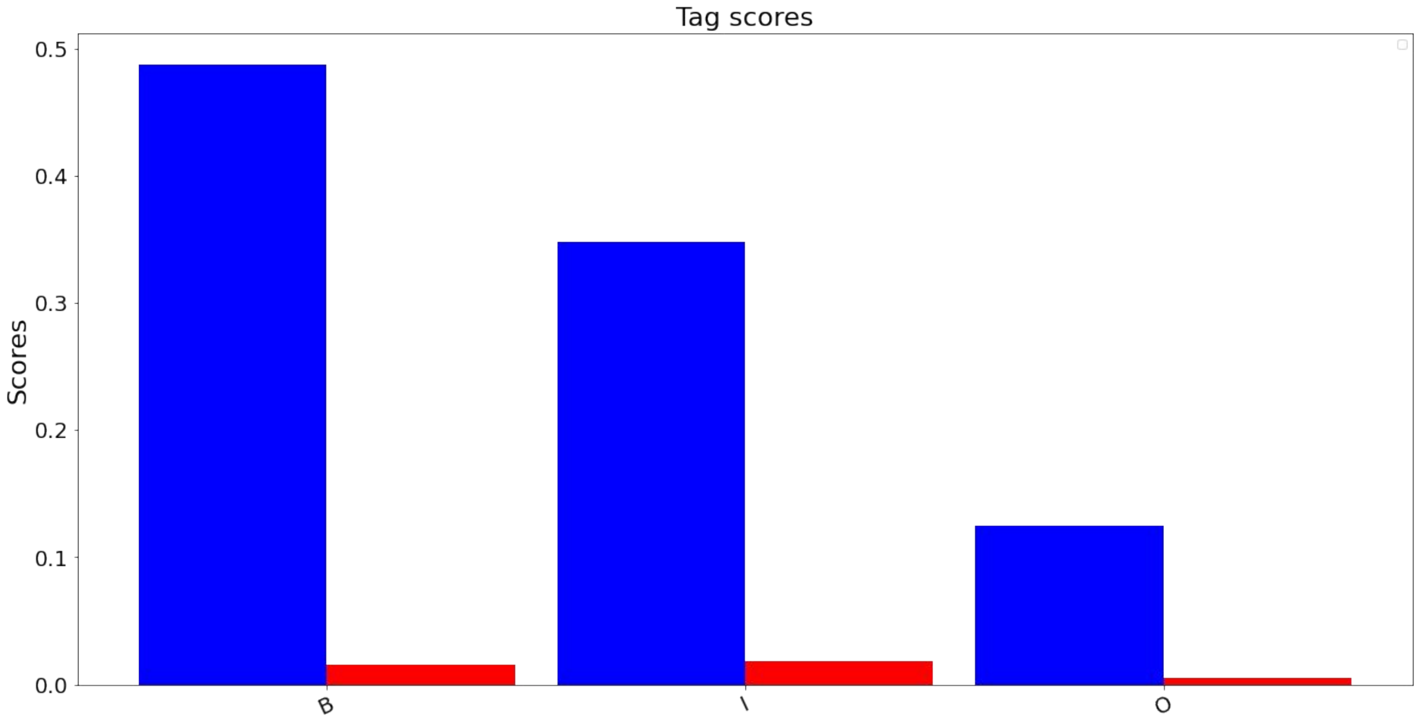
	B	I	O
B	0.97	0.03	0.00
I	0.04	0.95	0.01
O	0.02	0.03	0.96

Confusion Matrix (with normalization)

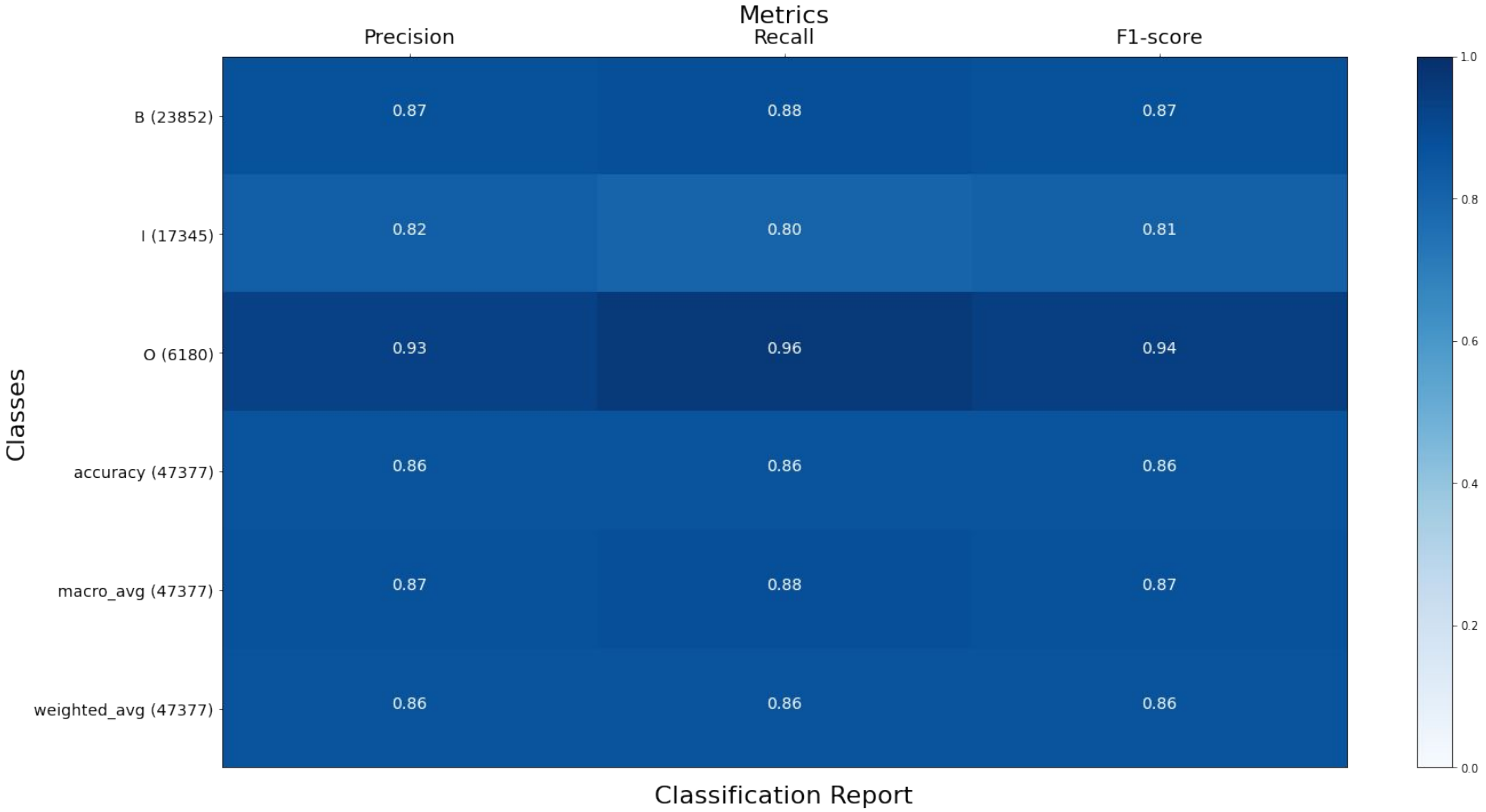
Tag scores without POS tags



Tag scores with POS tags

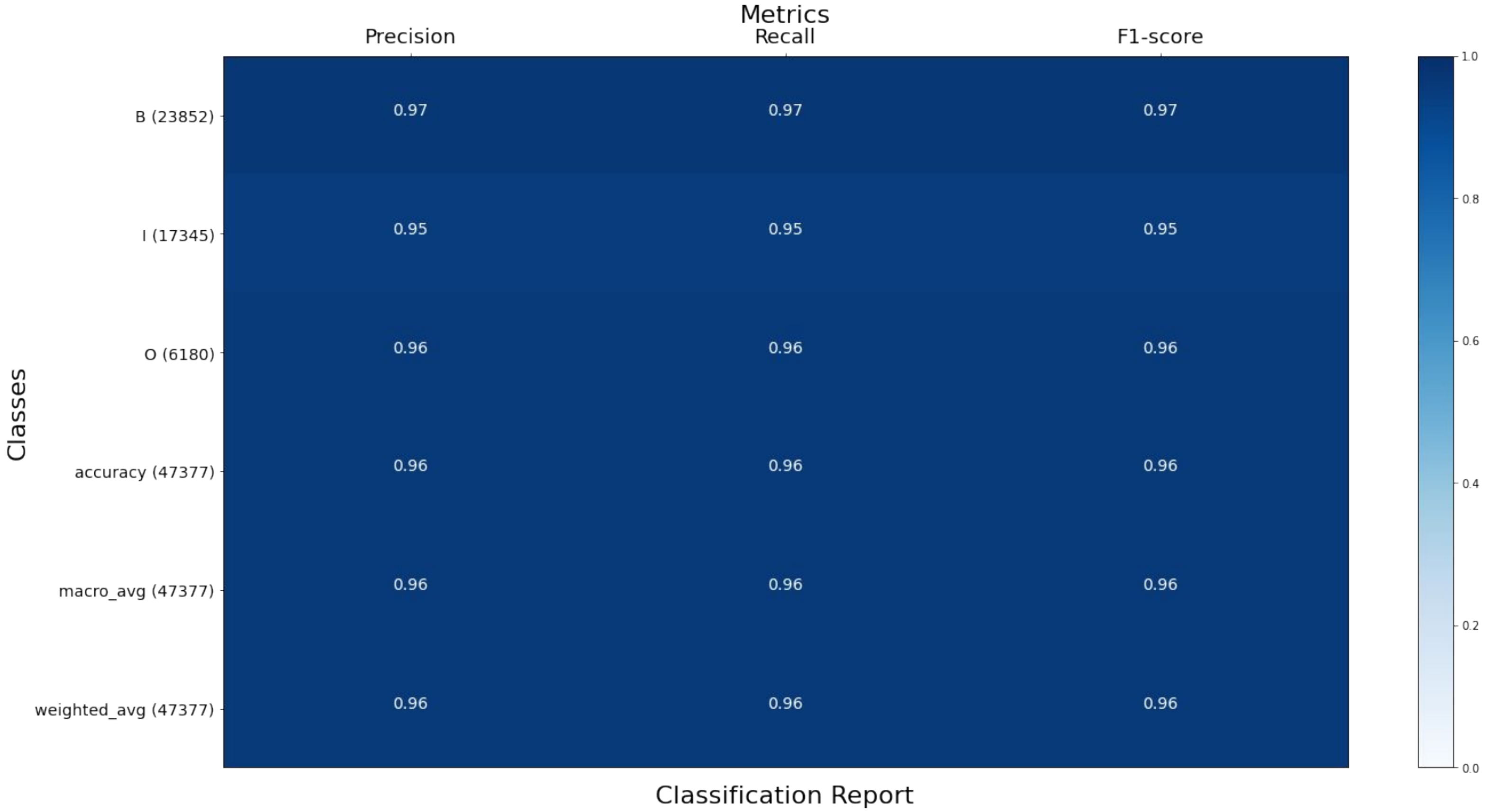


Classification report without POS tags



Classes	Precision	Recall	F1-score
B (23852)	0.87	0.88	0.87
I (17345)	0.82	0.80	0.81
O (6180)	0.93	0.96	0.94
accuracy (47377)	0.86	0.86	0.86
macro_avg (47377)	0.87	0.88	0.87
weighted_avg (47377)	0.86	0.86	0.86

Classification report with POS tags



Metrics

Classification Report

Classes	Precision	Recall	F1-score
B (23852)	0.97	0.97	0.97
I (17345)	0.95	0.95	0.95
O (6180)	0.96	0.96	0.96
accuracy (47377)	0.96	0.96	0.96
macro_avg (47377)	0.96	0.96	0.96
weighted_avg (47377)	0.96	0.96	0.96

Error Analysis (without POS)

- Because of lack of punctuation in data compared to other words, chunk are mistagged near commas and apostrophes. Eg:
  a. These include , (correct: B B O predicted: B I O)
  b. each jetliner 's (correct: B I B predicted: B B B)
  c. aerospace , electronics (correct: I I I predicted: I O B)
- At times the model cannot figure whether a very rarely occurring word is a noun or adjective. Eg:
  a. keel beam . (correct: I I O predicted: I B O) Here 'keel' is a rare word.
- Proper nouns are wrongly classified, we expect this to be improved by POS tagging as it would have NNP tag. Eg
  a. Frank Carlucci III (correct: B I I predicted: B B I)
  b. to Boeing . (correct: B B O predicted: B I O)

- Numbers are also wrongly classified sometimes. Model is not able to figure that numeral followed by year is one chunk, perhaps because numbers occur rarely in corpus. This should improve when POS tags are available.
  a. 42 years old (correct: B I B predicted: B B B)
  b. 59 years old (correct: B I B predicted: B B B)
- Longer chunks are mishandled at times. Eg:
  a. loan and real estate (correct: I I I I predicted: I O B I)

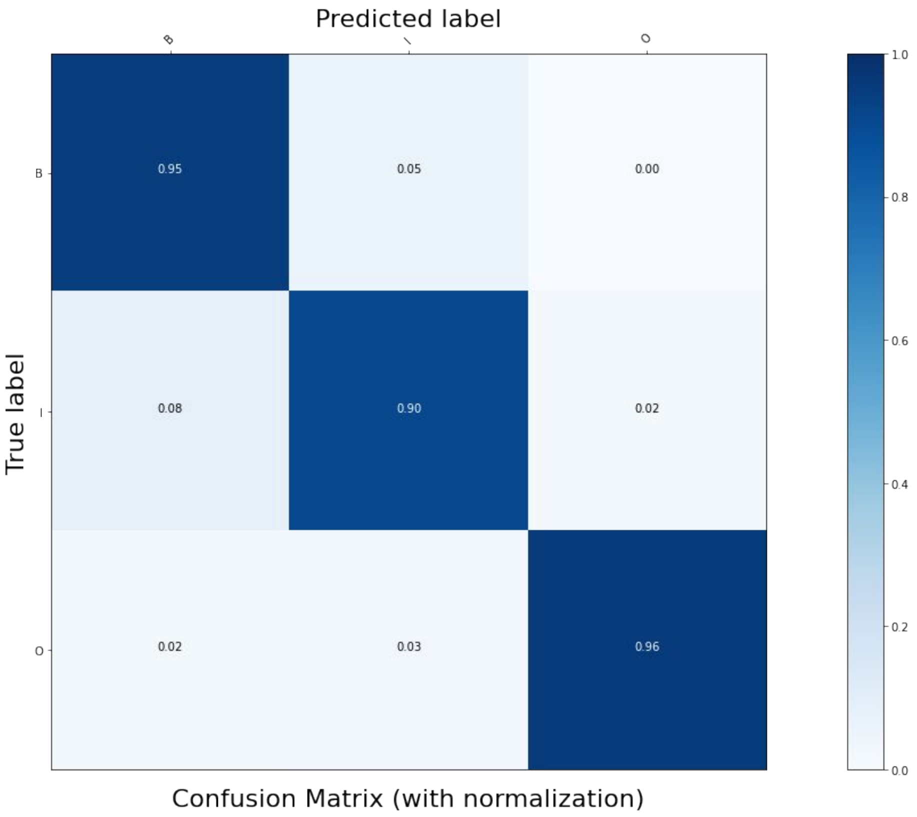
Error Analysis (with POS)

- Long names composed of common nouns were misclassified.
    a. US Facilities Corp. (correct: B I I predicted: B B I)
- After classifying one word as 'O' the next word is misclassified,this could be because in the corpus 'O' tag is given to a variety of things like punctuation marks and conjunctions ('and' being an example) Also the total number of them is quite less because of which the model couldn't learn useful information about it.
- When signs such as '\$' or '\%' appear in the corpus, it can lead to mis classified words many a times.
- Some hyphenated words were misclassified when they appeared with 'I' tag. Eg:
    a. and personal-care businesses (correct: I I I predicted: O B I)
- The accuracy is quite high and rest of the misclassifications dont follow a definitive pattern

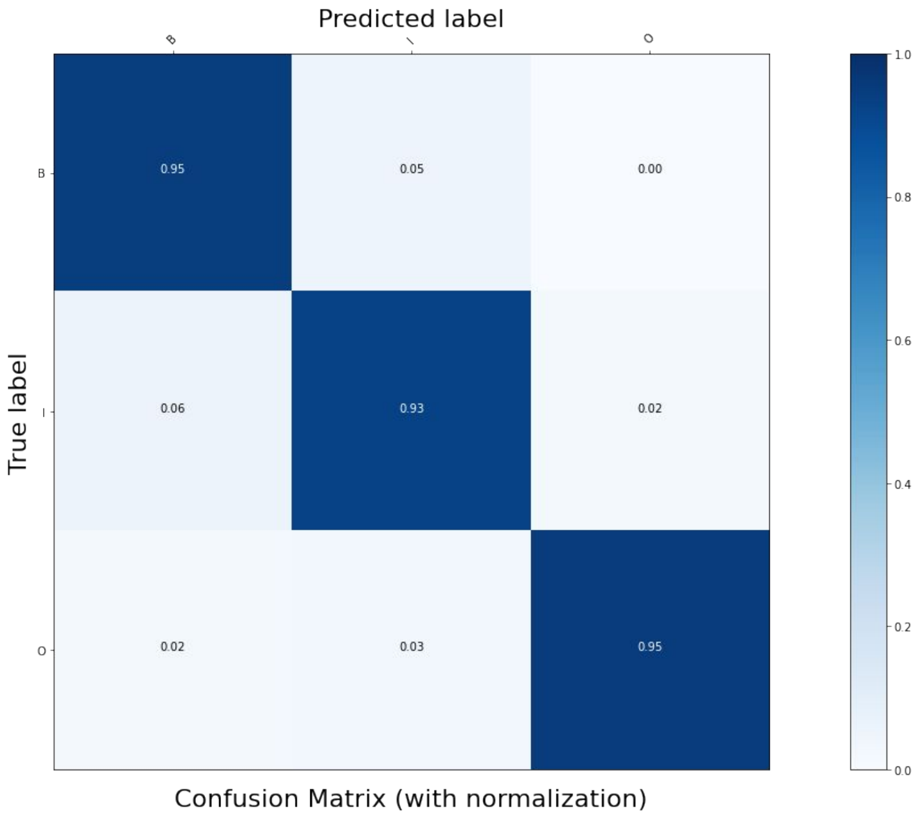
BiLSTM Model Architecture

- 300 dimensional GloVe embeddings used as initialization
- Embedding layer used for word features
- Additional Embedding layer to analyse the effects of POS tags
- TimeDistributed Layer followed by LSTM architecture
- As expected (in the following slides), the inclusion of POS tags along with the sentences doesn’t have much effect on the accuracy as opposed to other models
- Better performance with GloVe embeddings as compared to Word2Vec

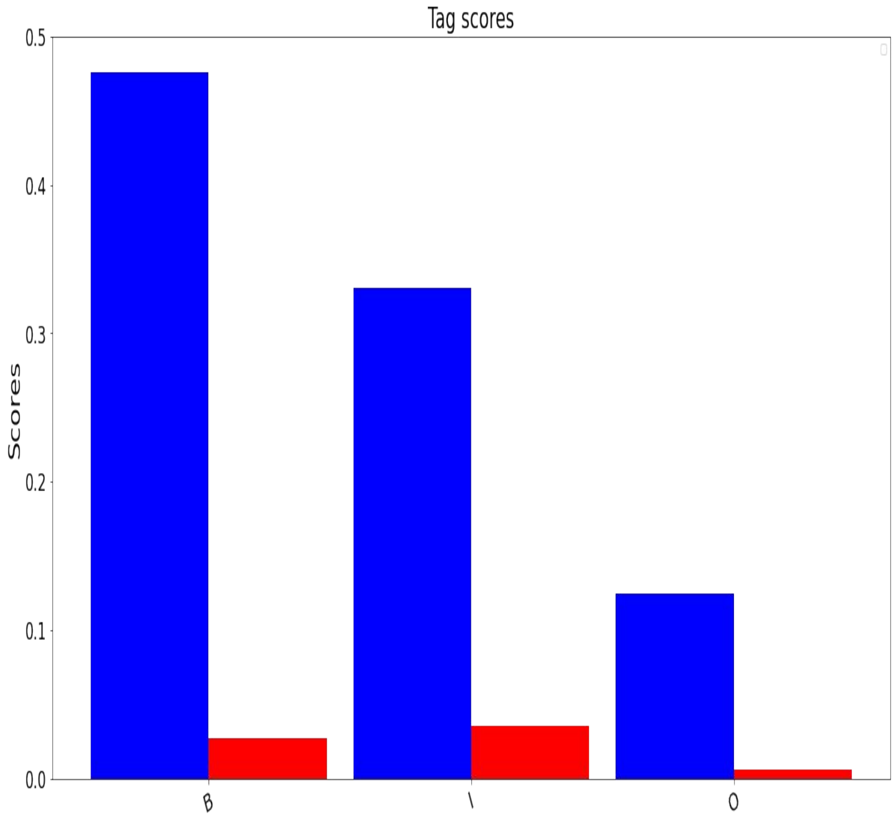
Confusion matrix without POS tags



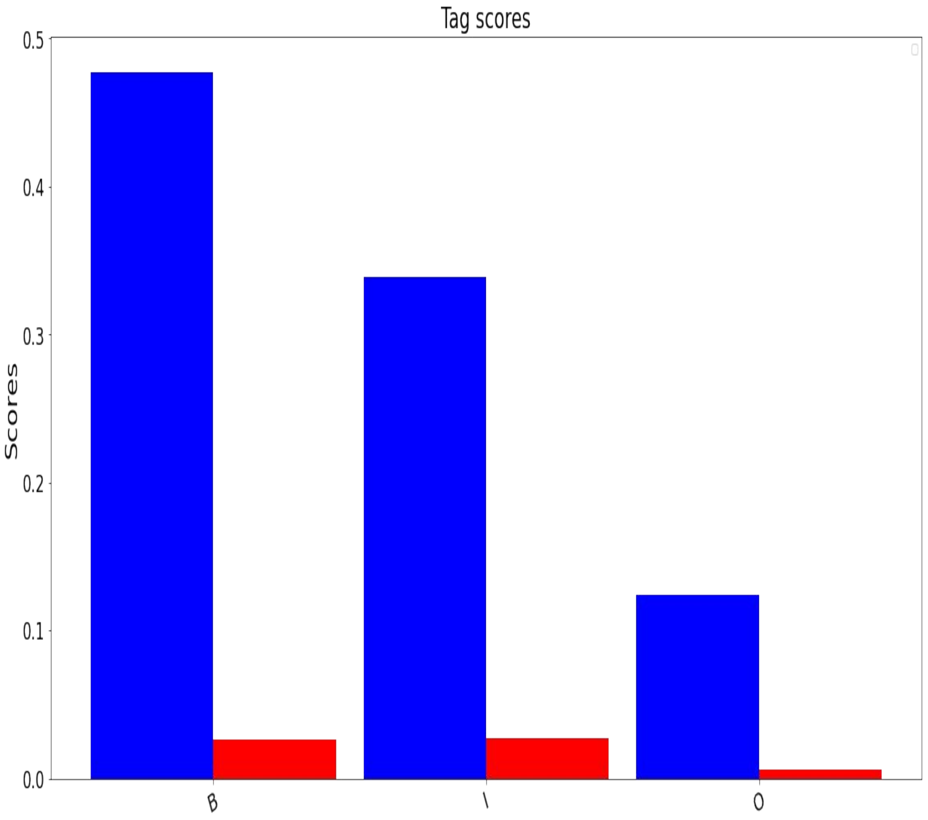
Confusion matrix with POS tags



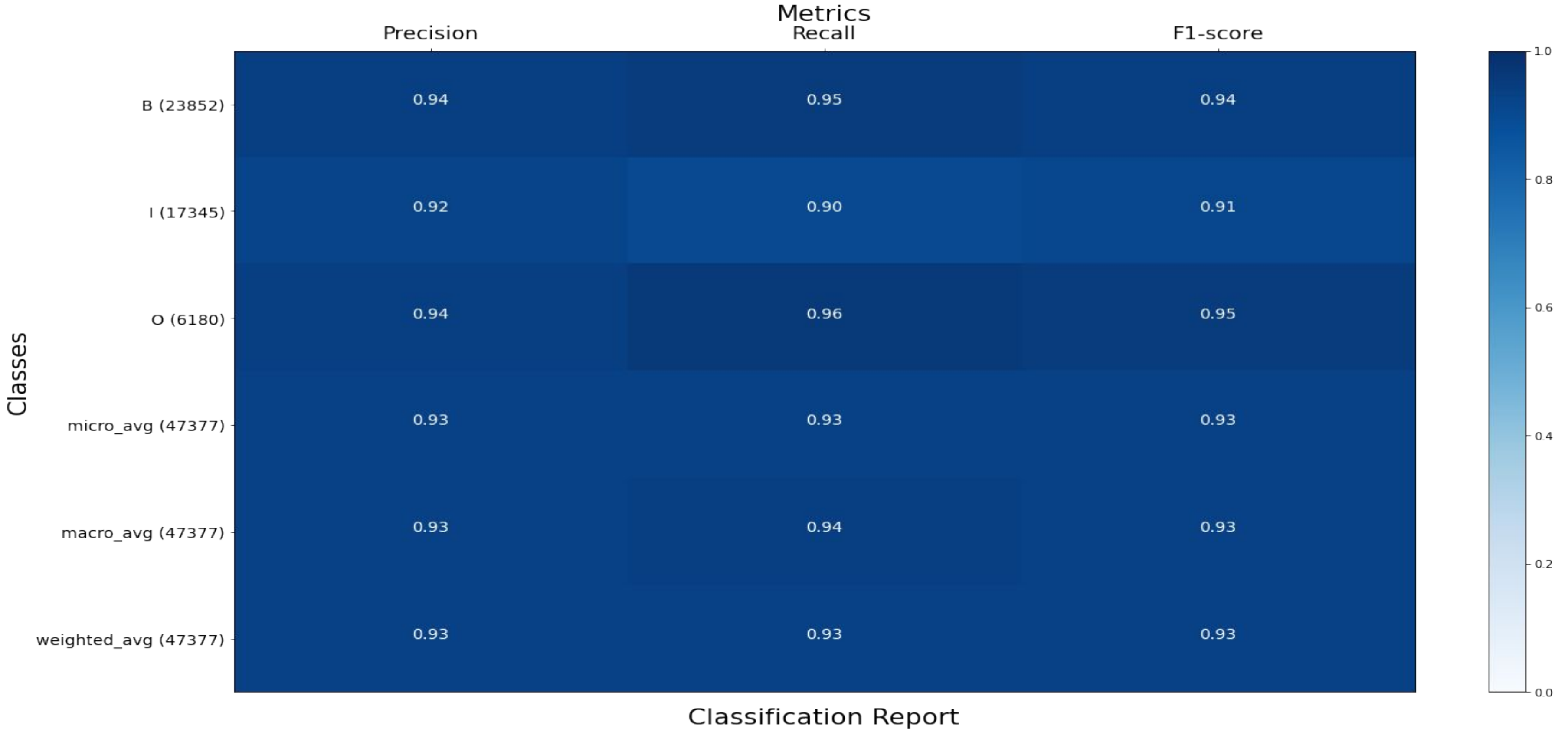
Tag Scores without POS tags



Tag Scores with POS tags

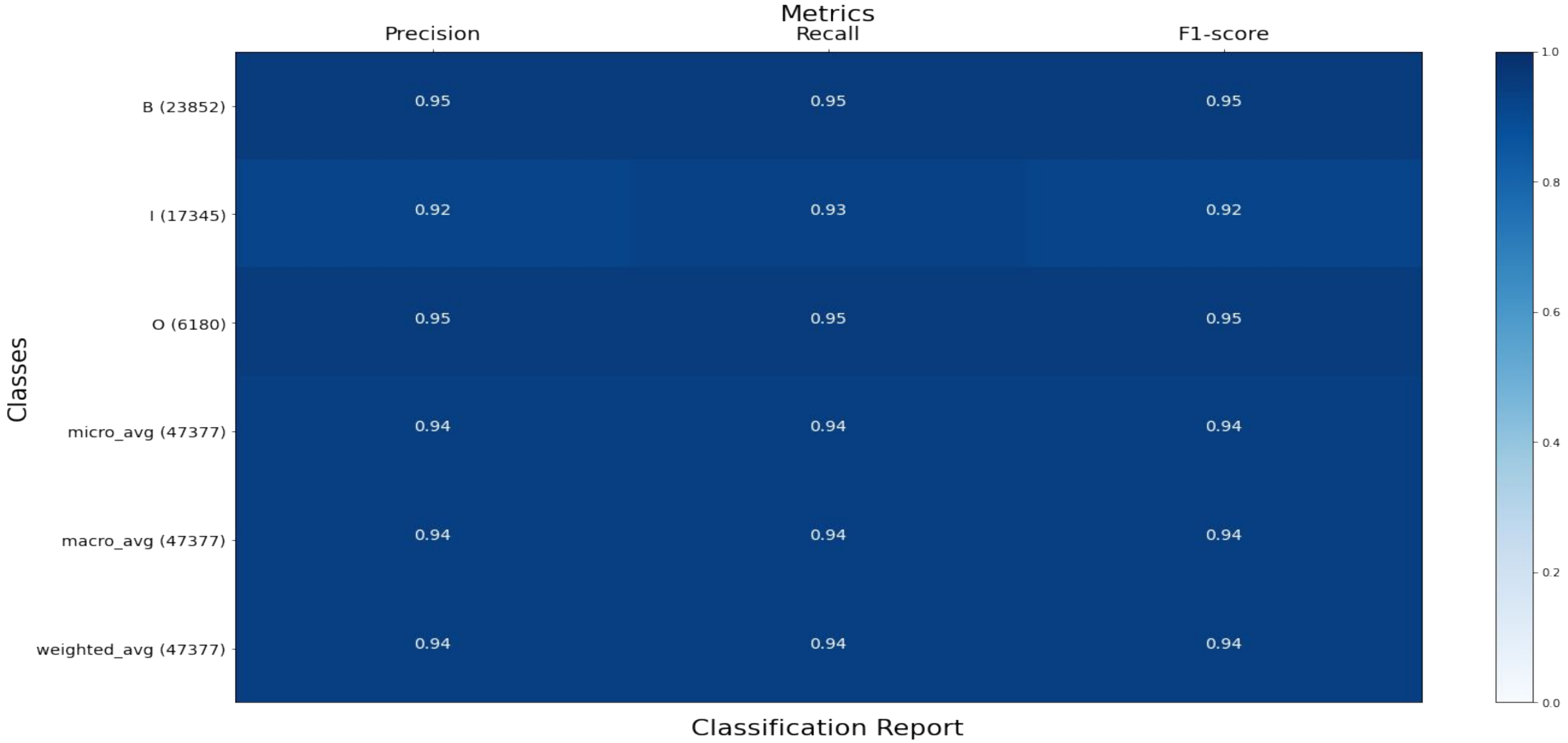


Classification report without POS tags



Classes	Precision	Recall	F1-score
B (23852)	0.94	0.95	0.94
I (17345)	0.92	0.90	0.91
O (6180)	0.94	0.96	0.95
micro_avg (47377)	0.93	0.93	0.93
macro_avg (47377)	0.93	0.94	0.93
weighted_avg (47377)	0.93	0.93	0.93

Classification report with POS tags



Classes	Precision	Recall	F1-score
B (23852)	0.95	0.95	0.95
I (17345)	0.92	0.93	0.92
O (6180)	0.95	0.95	0.95
micro_avg (47377)	0.94	0.94	0.94
macro_avg (47377)	0.94	0.94	0.94
weighted_avg (47377)	0.94	0.94	0.94

Error Analysis (with POS)

- Many word are present in the test data which are unknown to the train data vocabulary, which results in the model to map this <UNK> to I, most of these unknown words are nouns at the beginning of a chunk
- Most punctuations such as comma when used alongside the a noun or verb are actually part of the chunk, but are mapped to O tag by the model
- Model sometimes fails to identify conjunctions in the same chunk, similar to comma misclassification
  a. chairman and president (correct: I I I predicted: I O B)
  b. formulate and execute (correct: I I I predicted: I O B)
- Symbols for currency like '$' are tagged as B when it occurs as a part of compound/longer chunk
  a. paltry $ 43.5 (correct: I I I predicted: I B I)

Error Analysis (without POS)

- This model also suffers with the same misclassifications as the one with postags
- ‘to’ word is misclassified often as it is sometimes the VP chunk beginning as a part of infinitives, and sometimes occurs inside the VP chunk
  1. stood to gain (correct: B I I   predicted: B B I)
- Misclassifies the words which follow O tag in the sentences as a single word chunk, this did not occur in model with POS tags, as NN tag would help the model to identify the type of chunk
  1. general and administrative (correct: B I I   predicted: B O B)
- Some verbs which are followed by auxiliary verbs are also marked as chunk beginners
  1. was dismissed on (correct: B I B predicted: B B B)
  2. was involved in (correct: B I B   predicted: B B B)