CS626 A2 - Shallow Parsing

Shreya Pathak, Neel Aryan Gupta, Mohammad Ali Rehan

180050100, 180050067, 180050061

1 MEMM (without POS tags)

Some wrong classifications include: Mistagging due to punctuations being used within a chunk. The model mistakes them for 'O' tag. Eg,

1. , electronics , (correct: I I I predicted: O B O)
2. , 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,

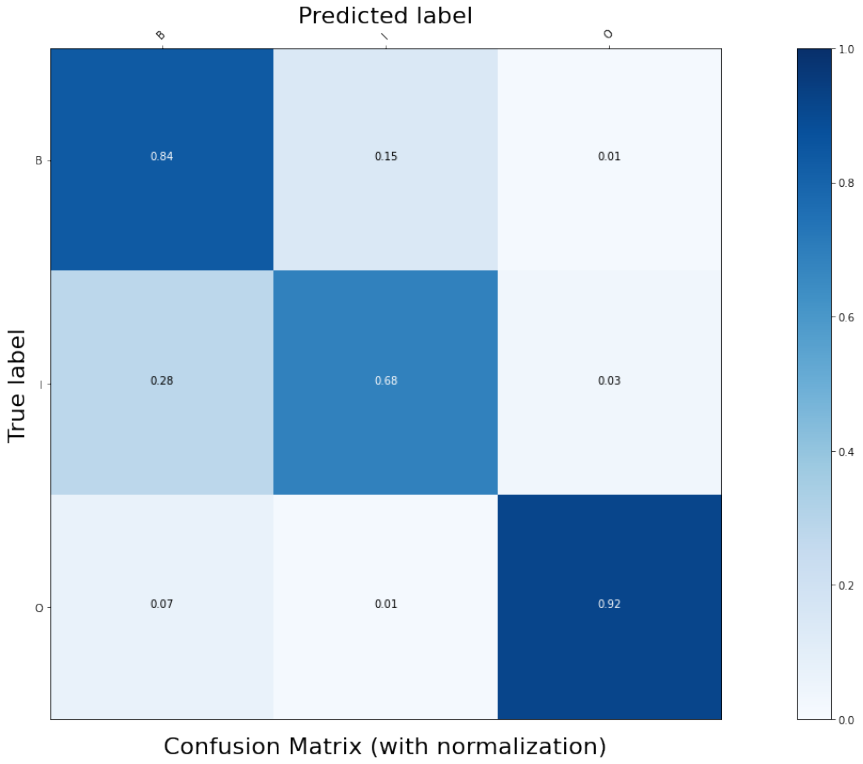
1. about 60%-held by (correct: B I B predicted: B B B)
2. 's 747 jetliners (correct: B I I predicted: B B I)

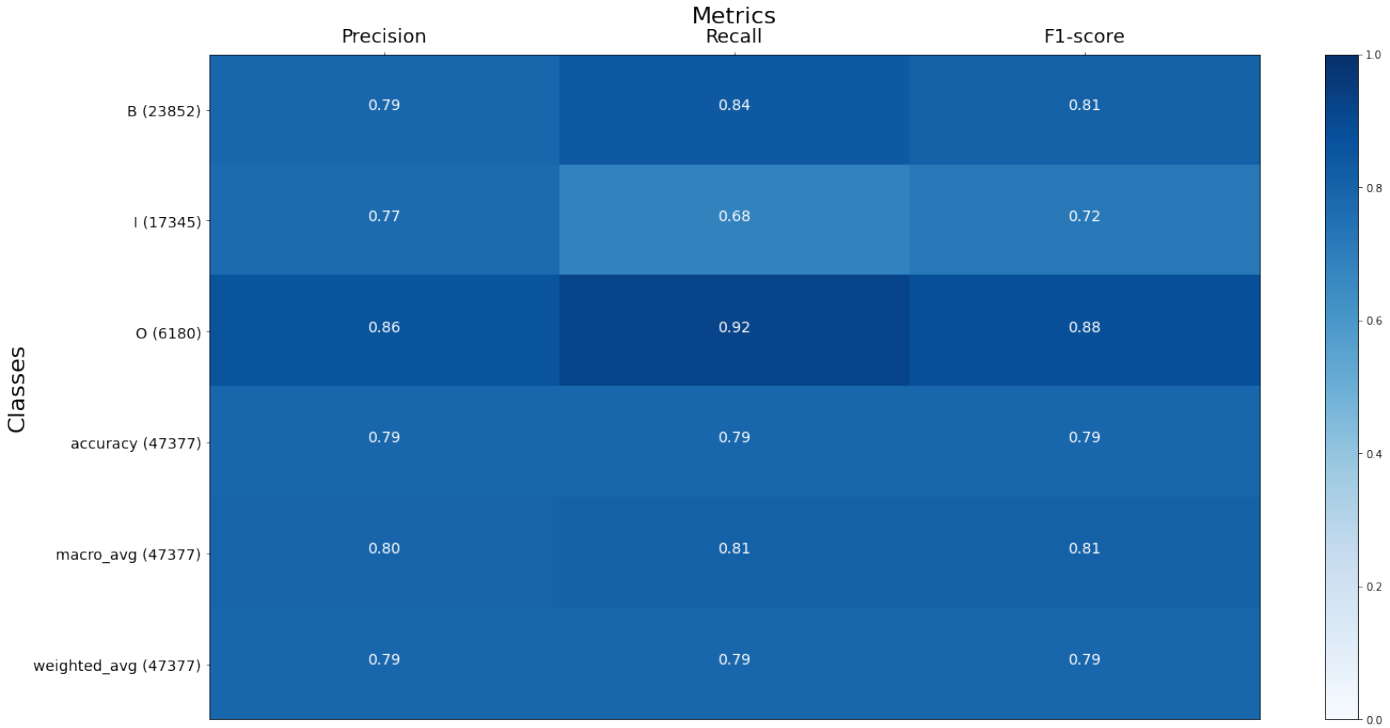
Conjunctions and their following word (within chunks) were often mistagged. Conjunctions were mistagged as 'O' and due to this the following word was considered as a 'B' tag. Eg,

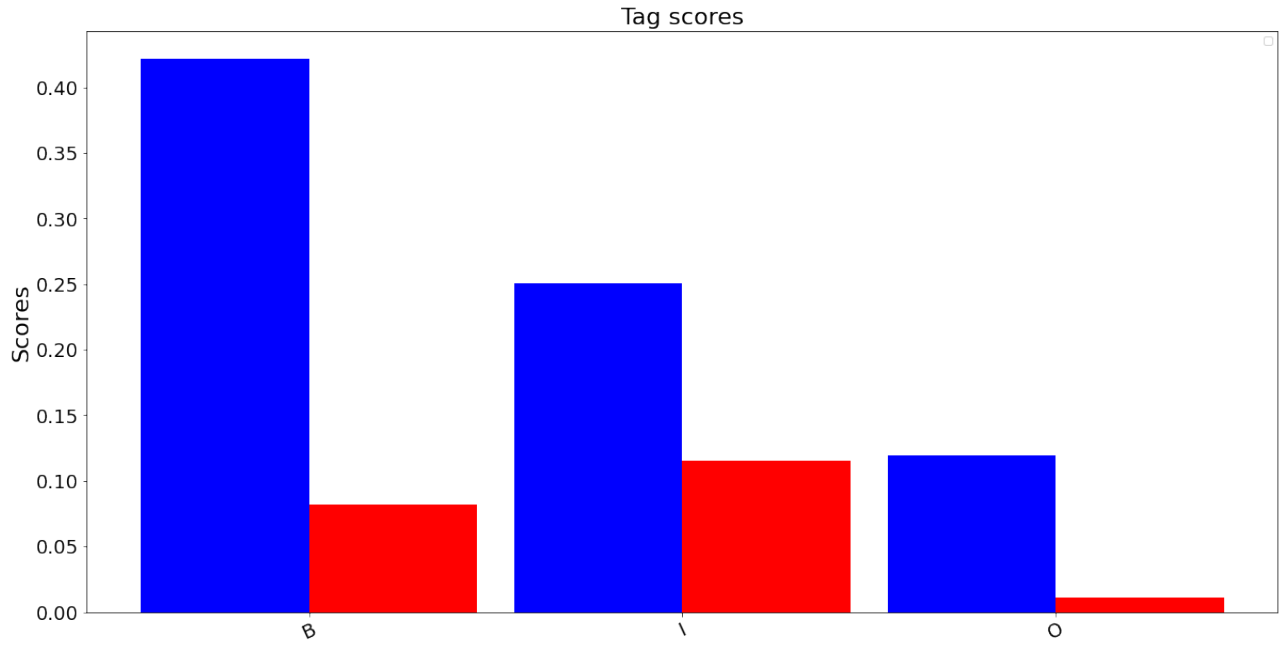
1. investment and asset-management (correct: I I I predicted: I O B)
2. automotive and graphics (correct: I I I predicted: B O B)

Multiple Proper nouns occuring 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,

1. Messrs. Wolf and (correct: B I I predicted: B B O)
2. Van Pell , (correct: B I O predicted: B B O)







2 MEMM (with POS tags)

Some wrong classifications include: Mistagging due to punctuations being used within a chunk. Again the model mistakes them to be 'O' tagged. Eg,

1. , electronics , (correct: I I I predicted: O B O)
2. , Ill. , (correct: I I I predicted: O B O)

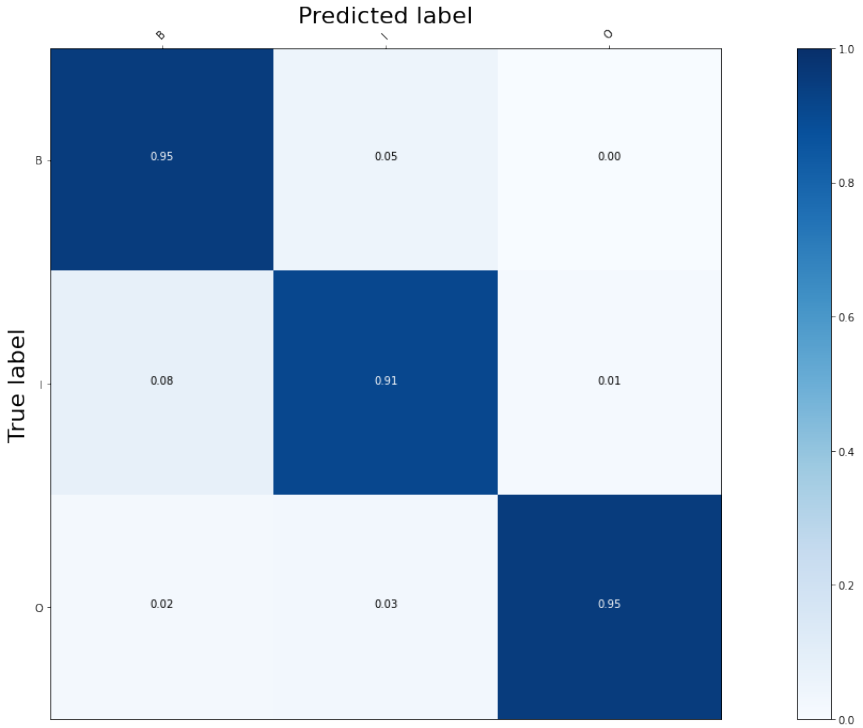
Conjunctions and their following word (within chunks) were often mistagged. Eg,

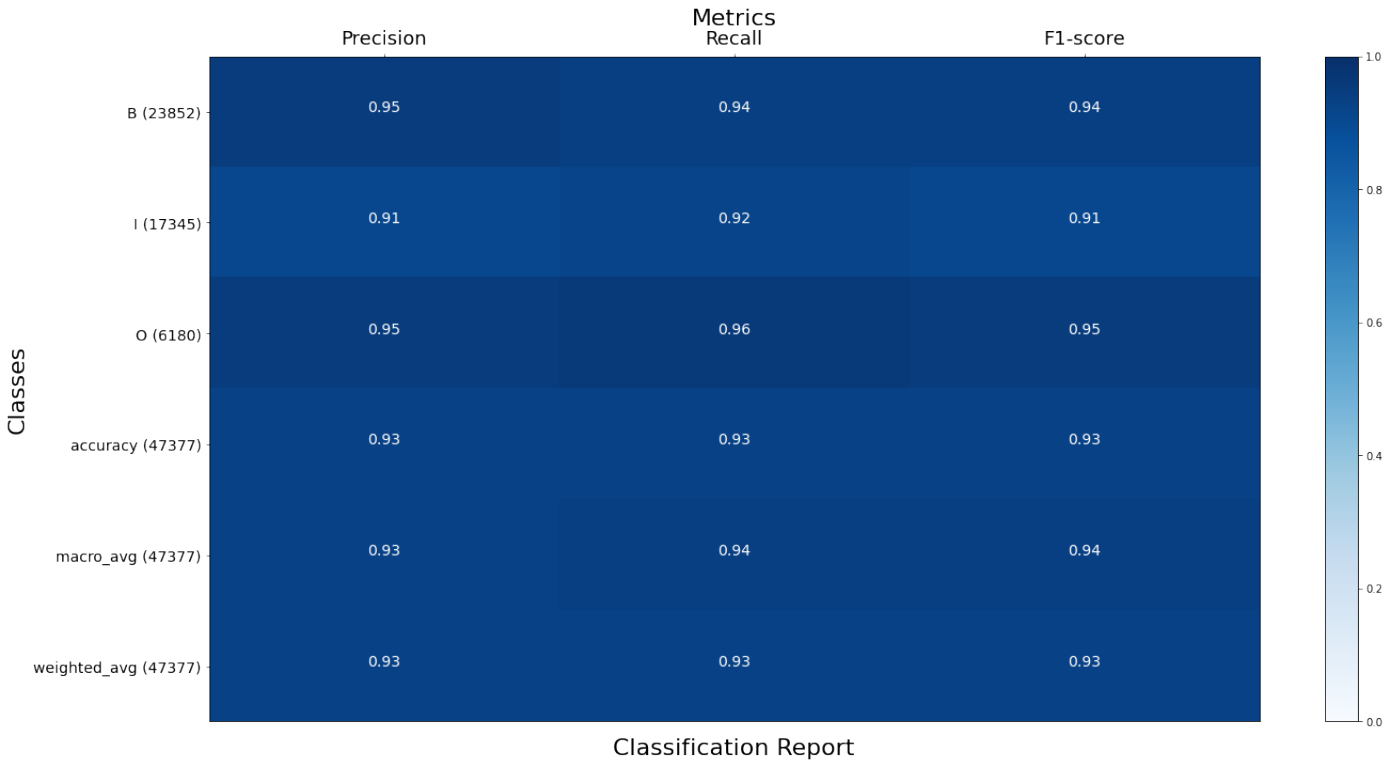
1. loan and real (correct: I I I predicted: I O B)
2. Wolf and Pope (correct: I I I predicted: I O B)

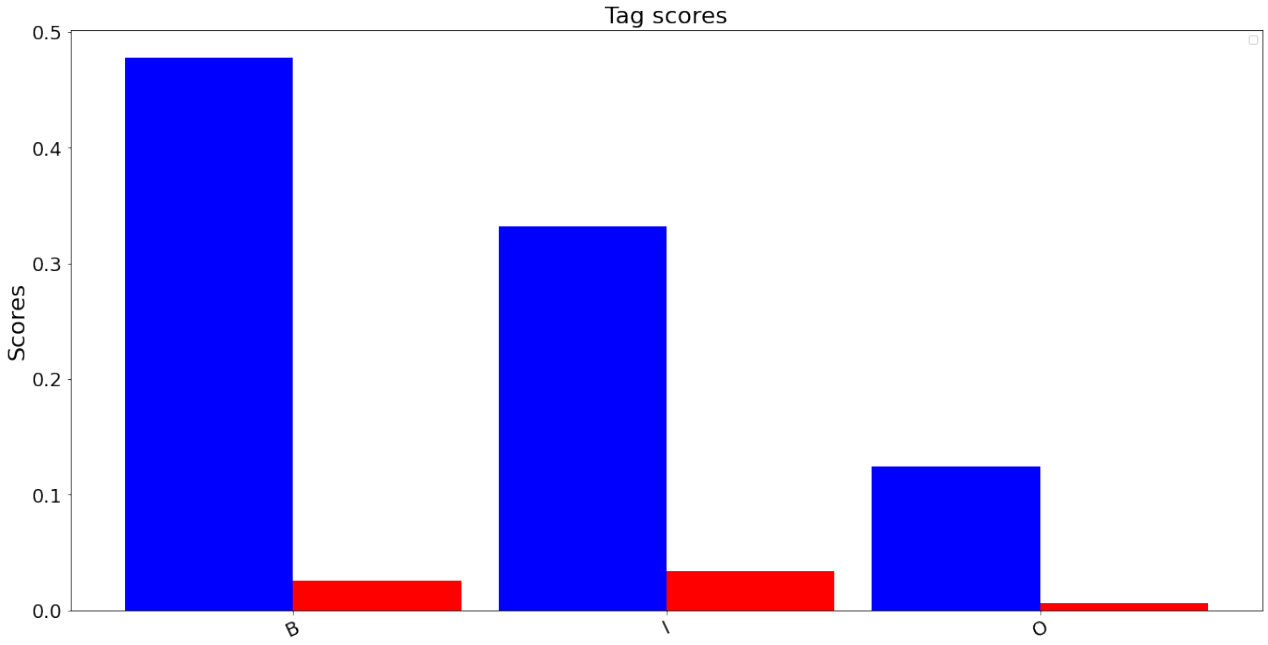
The word 'to' occuring in a chunk is often mistagged when it is present within a chunk. The model mistakes it to be the beginning of a chunk. Eg,

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)







3 CRF (without POS tags)

Some wrong classifications include:

Because of lack of punctuation in data compared to other words, chunk are mistagged near commas and apostrophes. Eg:

1. These include , (correct: B B O predicted: B I O)
2. each jetliner 's (correct: B I B predicted: B B B)
3. aerospace , electronics (correct: I I I predicted: I O B)

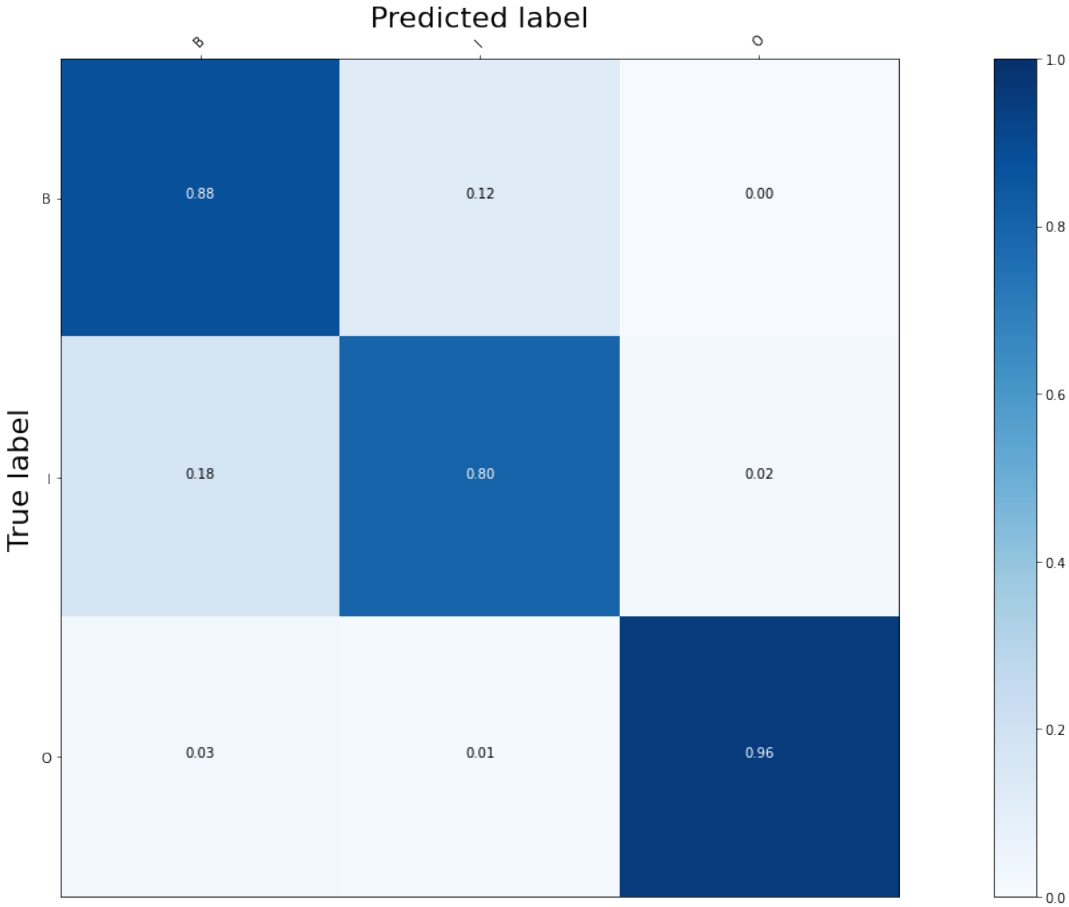
At times the model cannot figure whether a very rarely occurring word is a noun or adjective. Eg: keel beam . (correct: I I O predicted: I B O)
Here 'keel' is a rare word.
Proper nouns are wrongly classified, we expect this to be improved by POS tagging as it would have NNP tag.Eg

1. Frank Carlucci III (correct: B I I predicted: B B I)
2. to Boeing . (correct: B B O predicted: B I O)

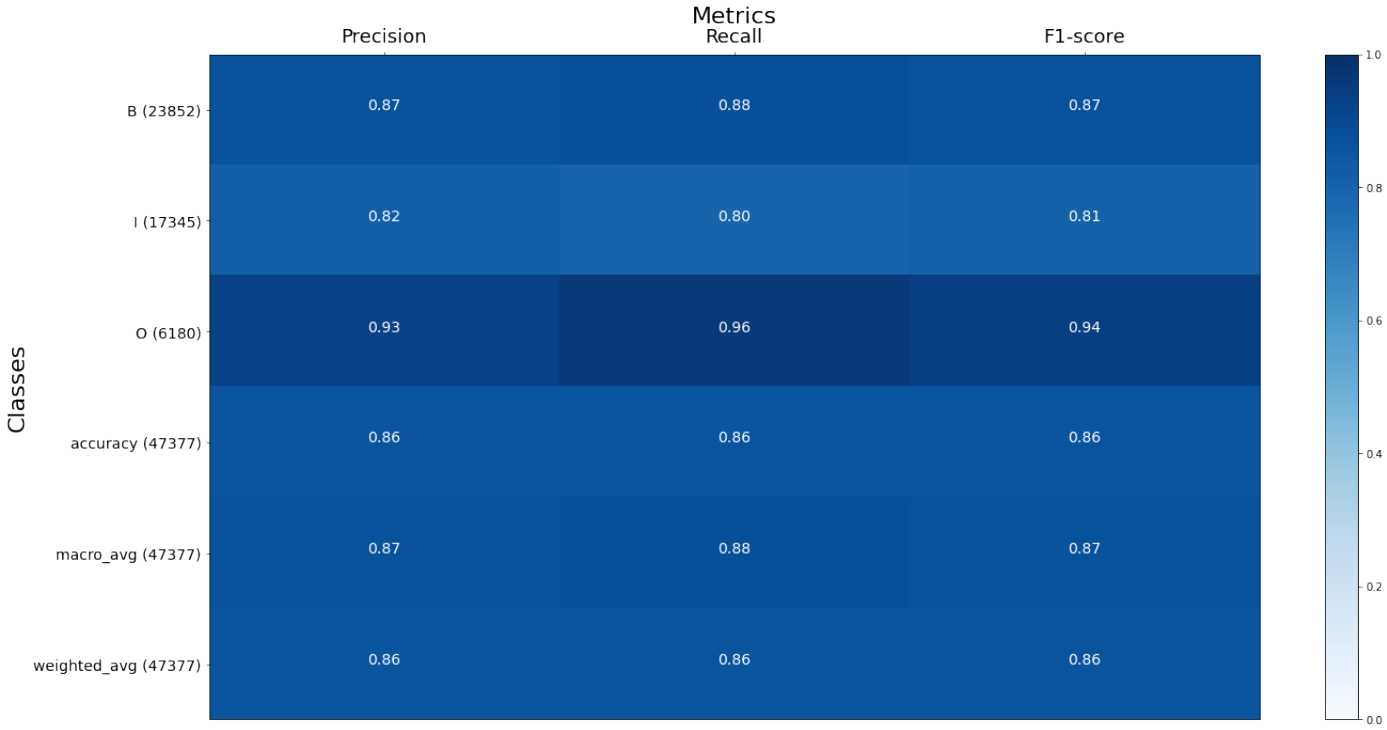
Numbers are also wrongly classified sometimes. Model is not able to figure that numeral followed by year is one chunk, perhaps because numbers occur rarely in corpus. This should improve when POS tags are available.

1. 42 years old (correct: B I B predicted: B B B)
2. 59 years old (correct: B I B predicted: B B B)

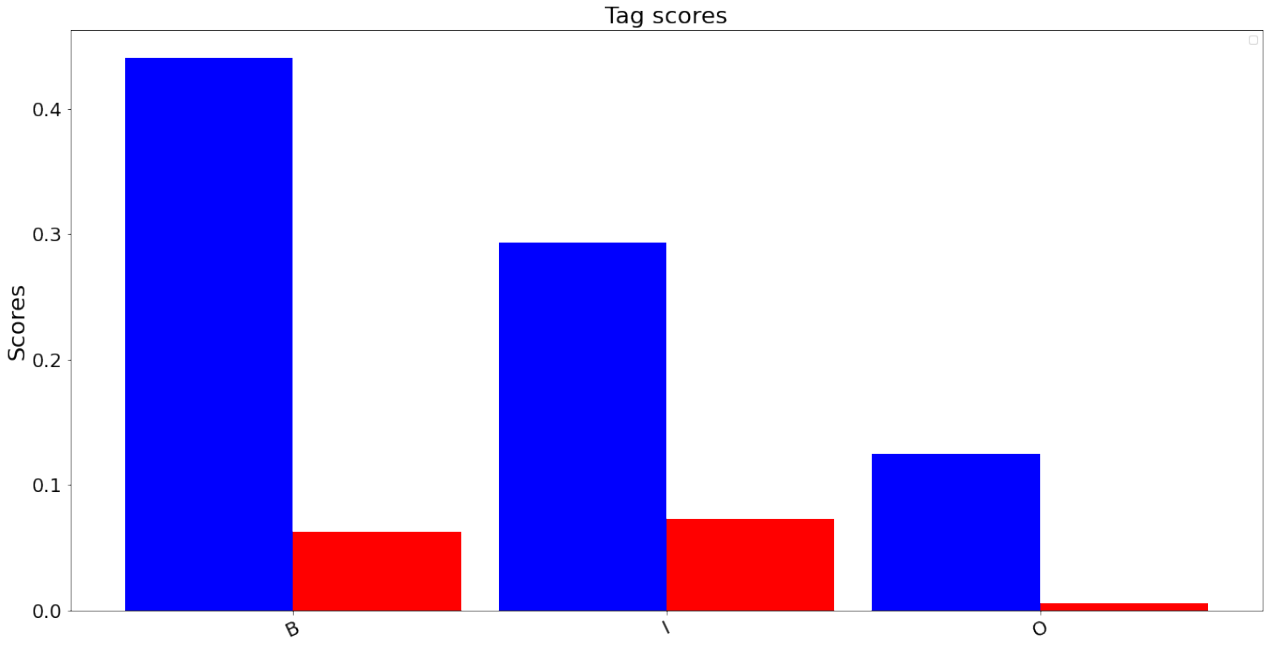
Longer chunks are mishandled at times. Eg: loan and real estate (correct: I I I I predicted: I O B I)



Confusion Matrix (with normalization)

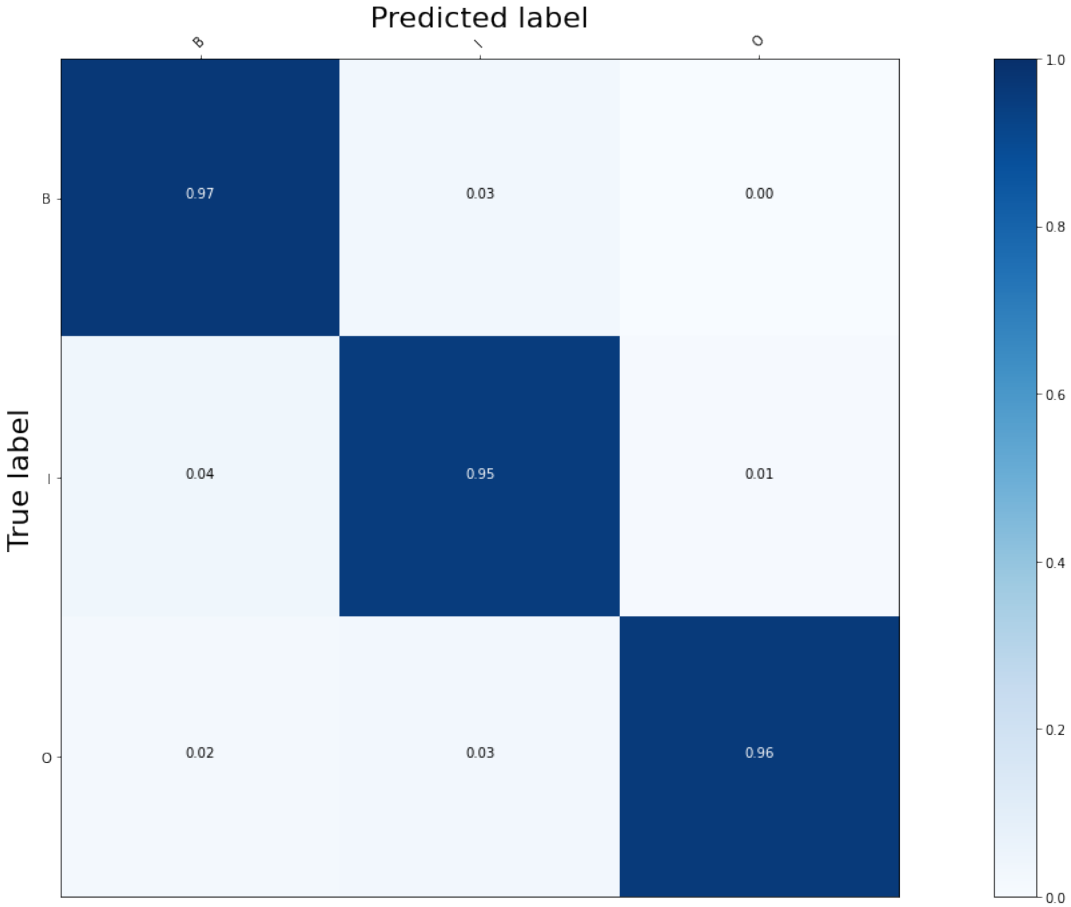


Classification Report

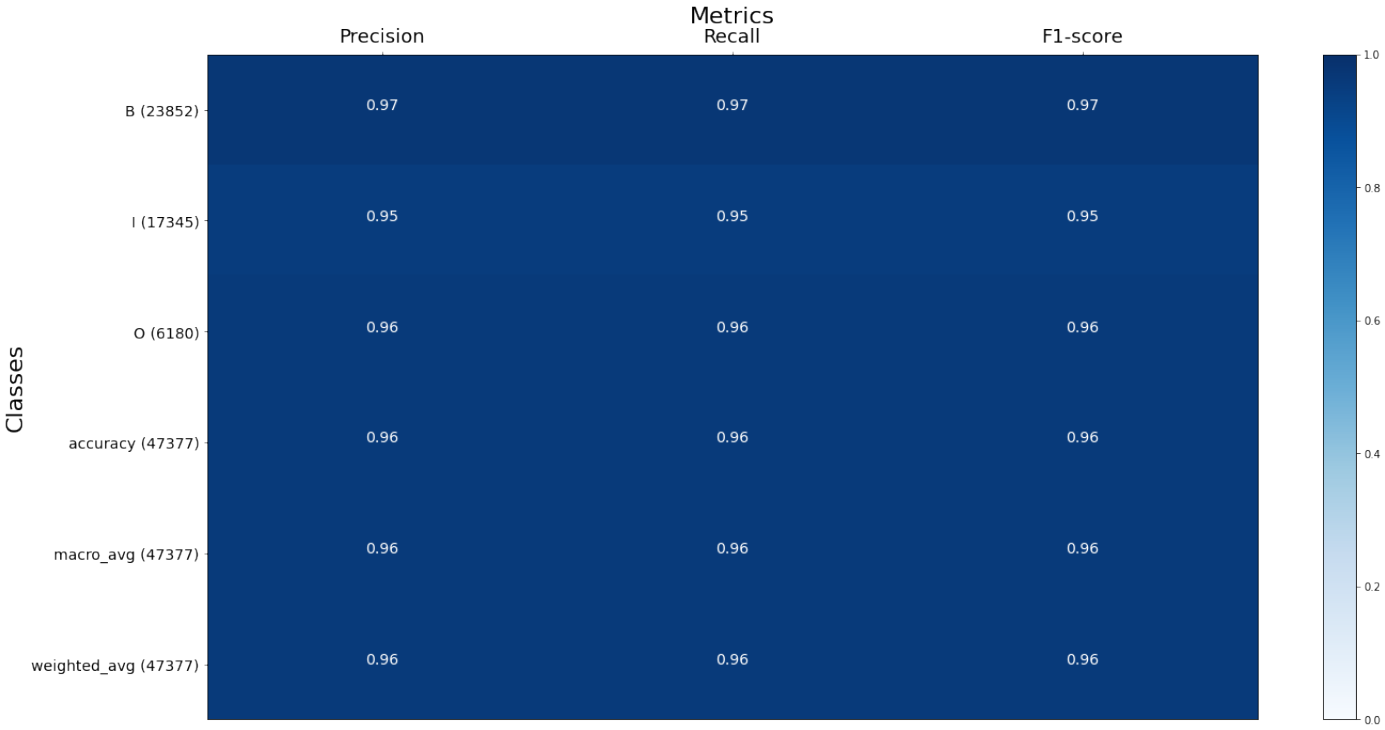


4 CRF (with POS tags)

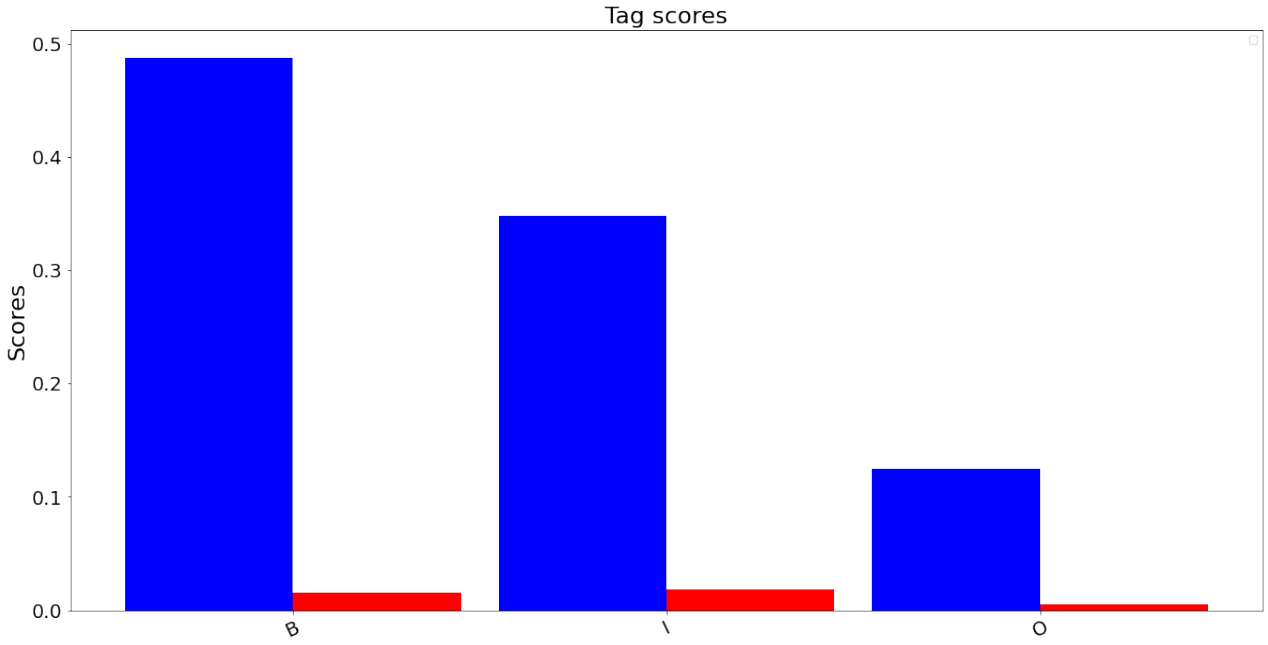
Many of the labels not being identified correctly of the previous sections vanished and number of mistakes reduced by a factor of five. Still some places where we observe mistakes are where there is a long name composed of nouns such as : US Facilities Corp. (correct: B I I predicted: B B I).
Many a times 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: and personal-care businesses (correct: I I I predicted: O B I)
The rest of the wrongly classified instances looked quite random and no concrete inferences could be drawn from them.



Confusion Matrix (with normalization)



Classification Report



5 BiLSTM (without POS tags)

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.

- stood to gain (correct: B I I predicted: B B I)
- Due to health (correct: B I B predicted: B B I)

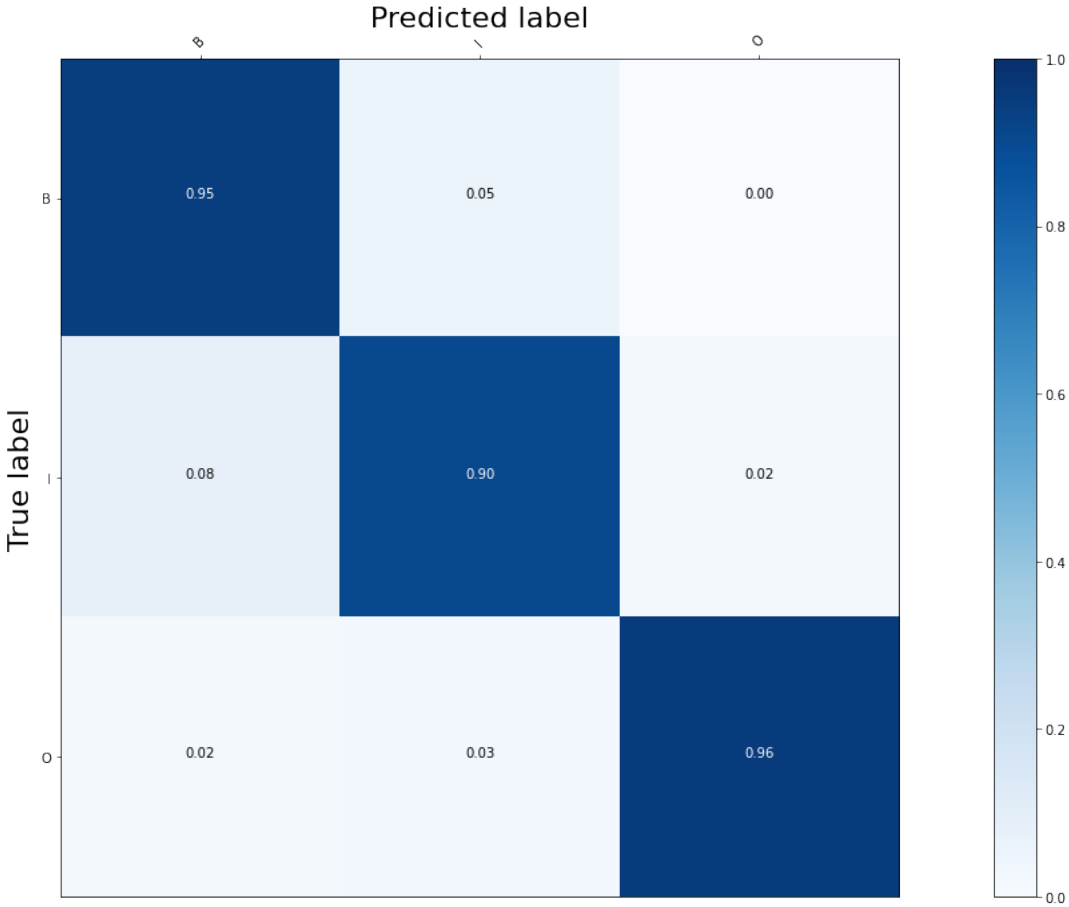
Misclassifies the words which follow O tag in the sentences as a single word chunk, this did not occur in model with POS tags, as NN tag would help the model to identify the type of chunk.

- general and administrative (correct: B I I predicted: B O B)
- Advancing and declining (correct: B I I predicted: B O B)

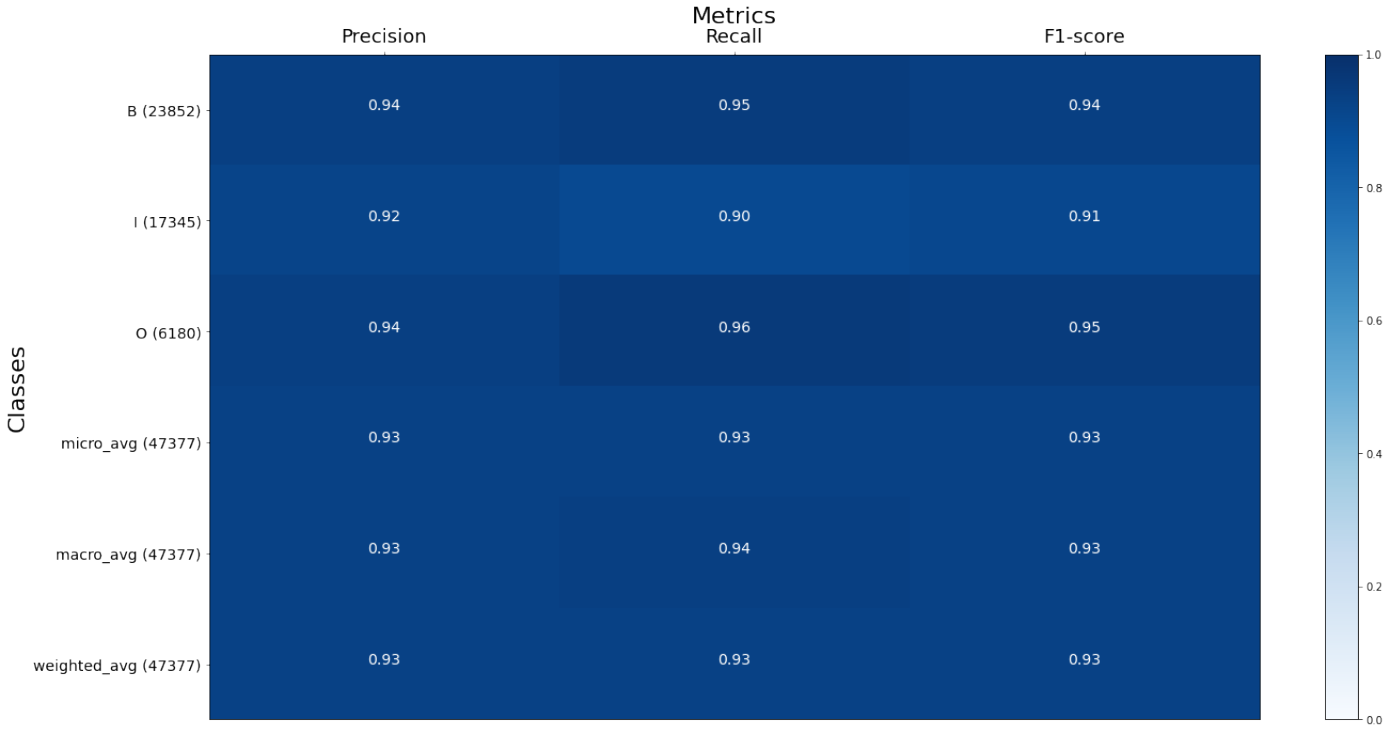
Some verbs which are followed by auxiliary verbs are also marked as chunk beginners.

- was dismissed on (correct: B I B predicted: B B B)
- was involved in (correct: B I B predicted: B B B)

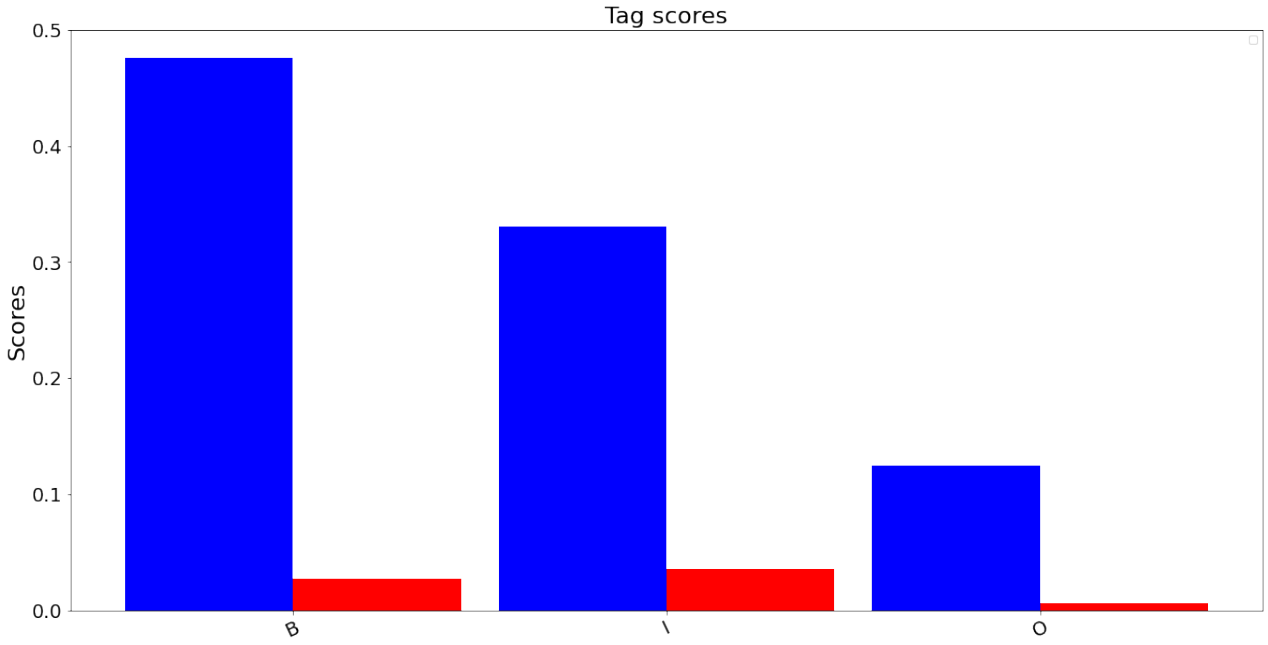
The confusion matrix, classification report and tag scores are shown below :



Confusion Matrix (with normalization)



Classification Report



6 BiLSTM (with POS tags)

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.

- U.S. hesitantly backed (correct: I B I predicted: I I B)
- than shortened his (correct: I B B predicted: I I B)

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.

- aerospace , electronics (correct: I I I predicted: I O B)
- pressed , French-made (correct: I I I predicted: I O B)

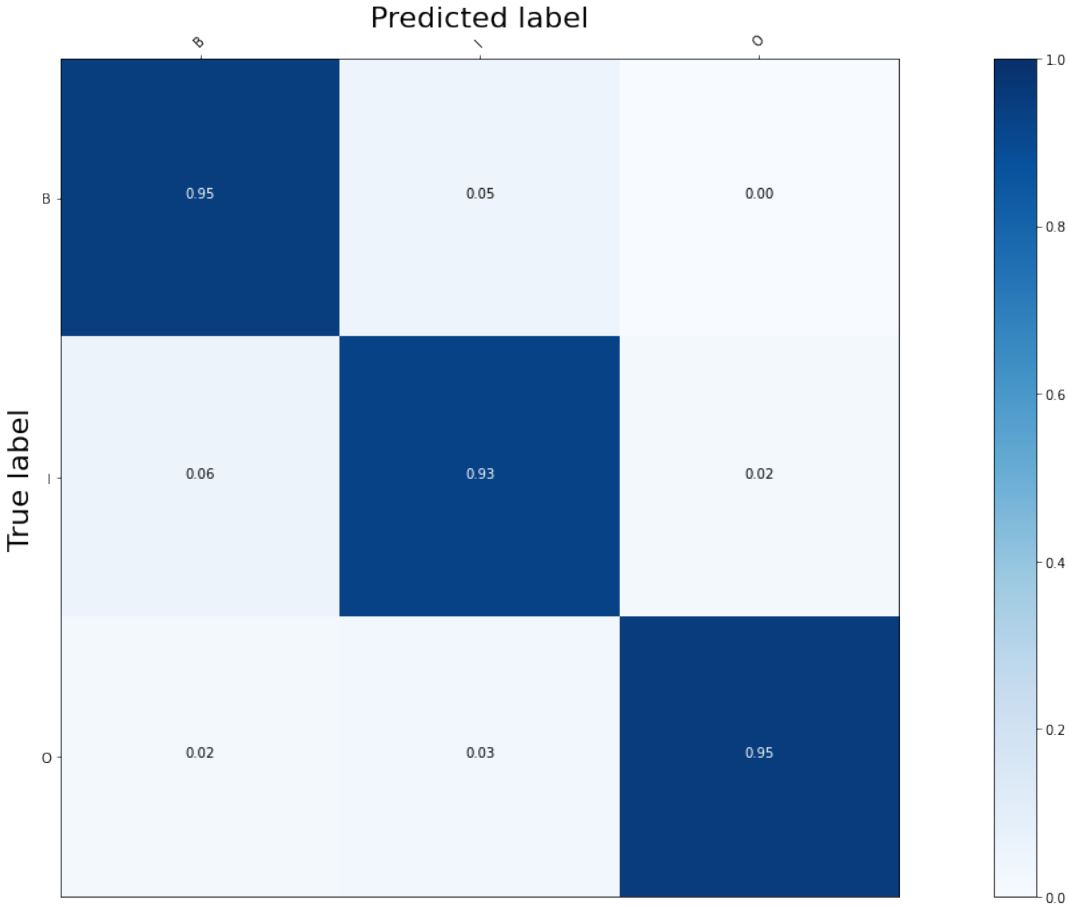
Model sometimes fails to identify conjunctions in the same chunk, similar to comma misclassification.

- chairman and president (correct: I I I predicted: I O B)
- formulate and execute (correct: I I I predicted: I O B)

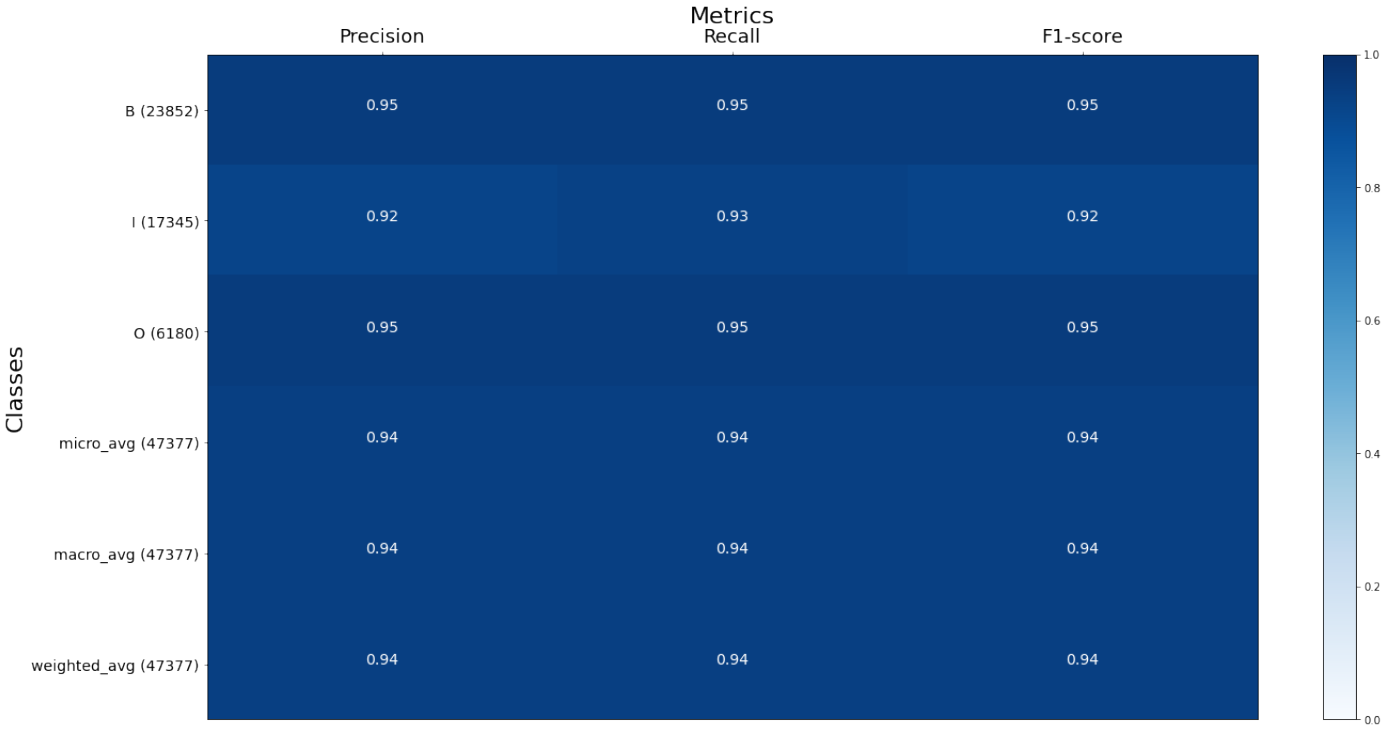
Symbols for currency like '$' are tagged as B when it occurs as a part of compound/longer chunk.

- paltry $ 43.5 (correct: I I I predicted: I B I)
- and $ 76 (correct: I I I predicted: O B I)

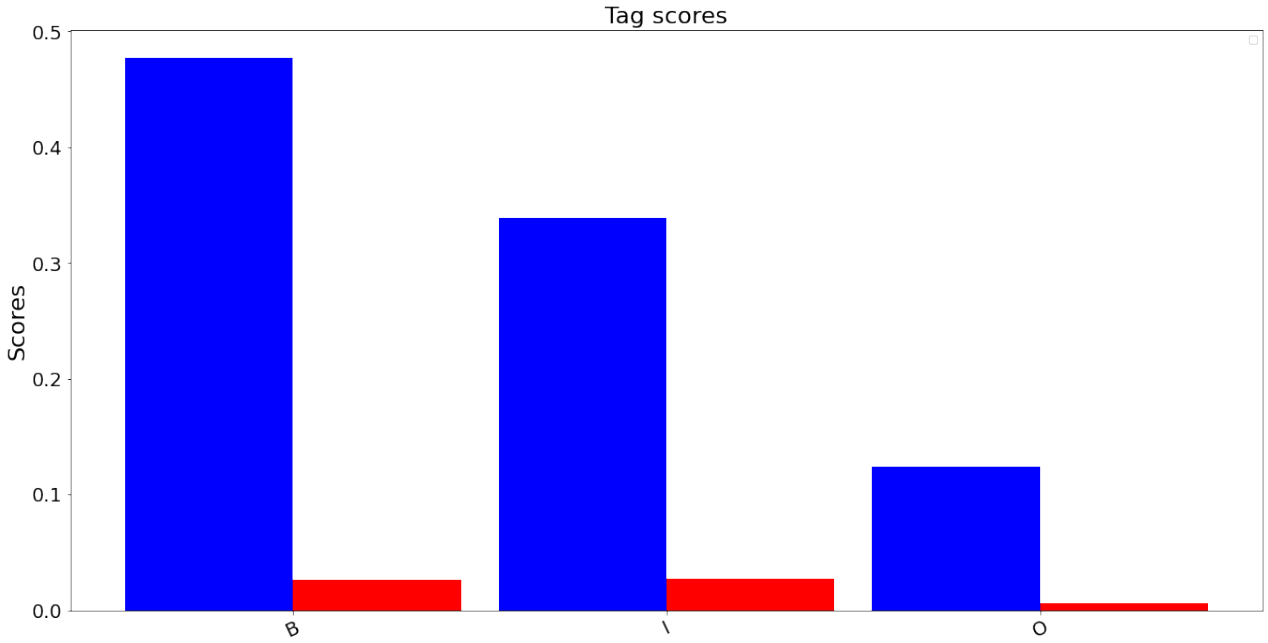
The confusion matrix, classification report and tag scores are shown below :



Confusion Matrix (with normalization)



Classification Report


Tag scores
Scores
0.5
0.4
0.3
0.2
0.1
0.0
B
I
O