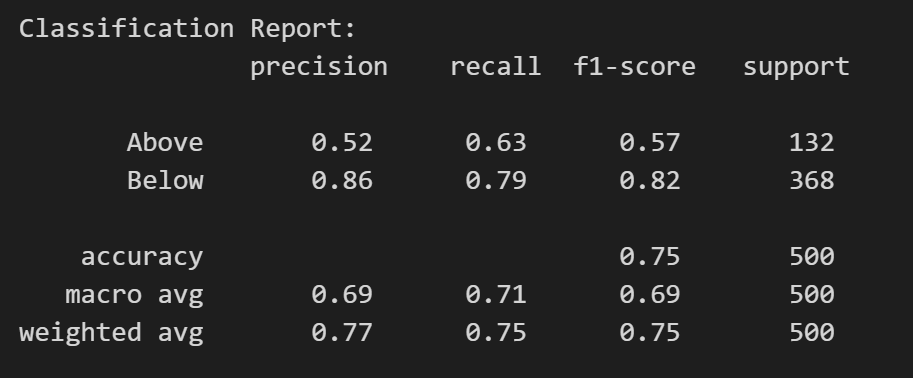
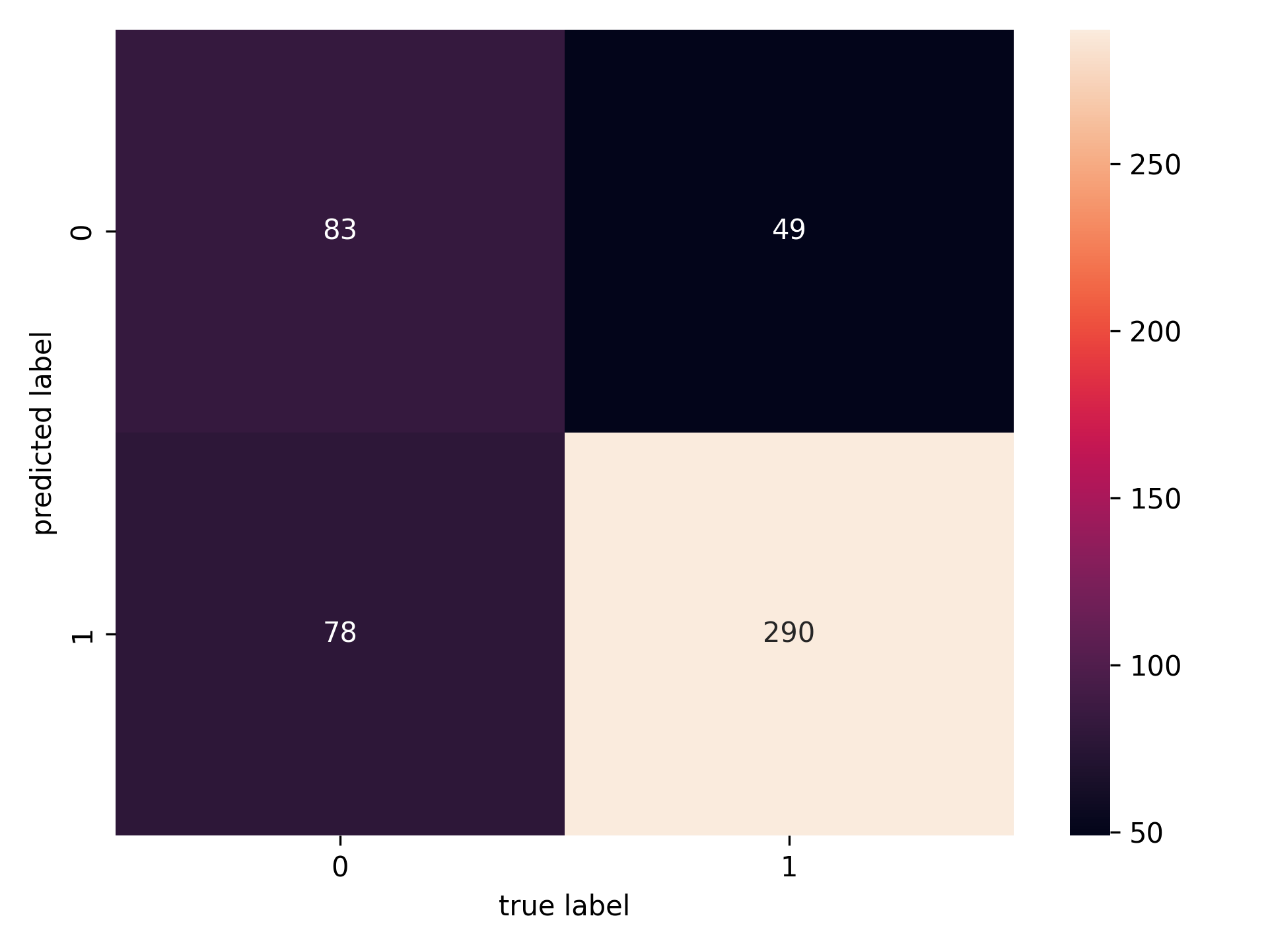
The classification report is:

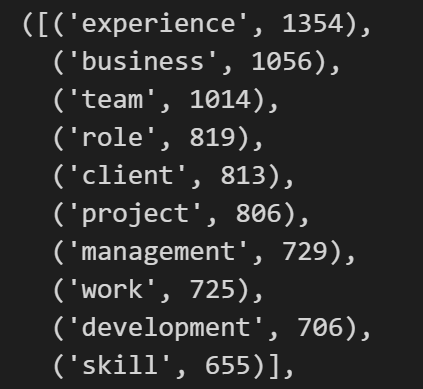


And the confusion matrix is:

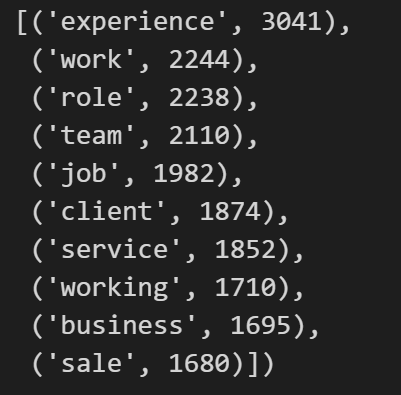


And the model accuracy is : 0.746

Top 10 words of high salary is:



And top 10 words of low salary is:

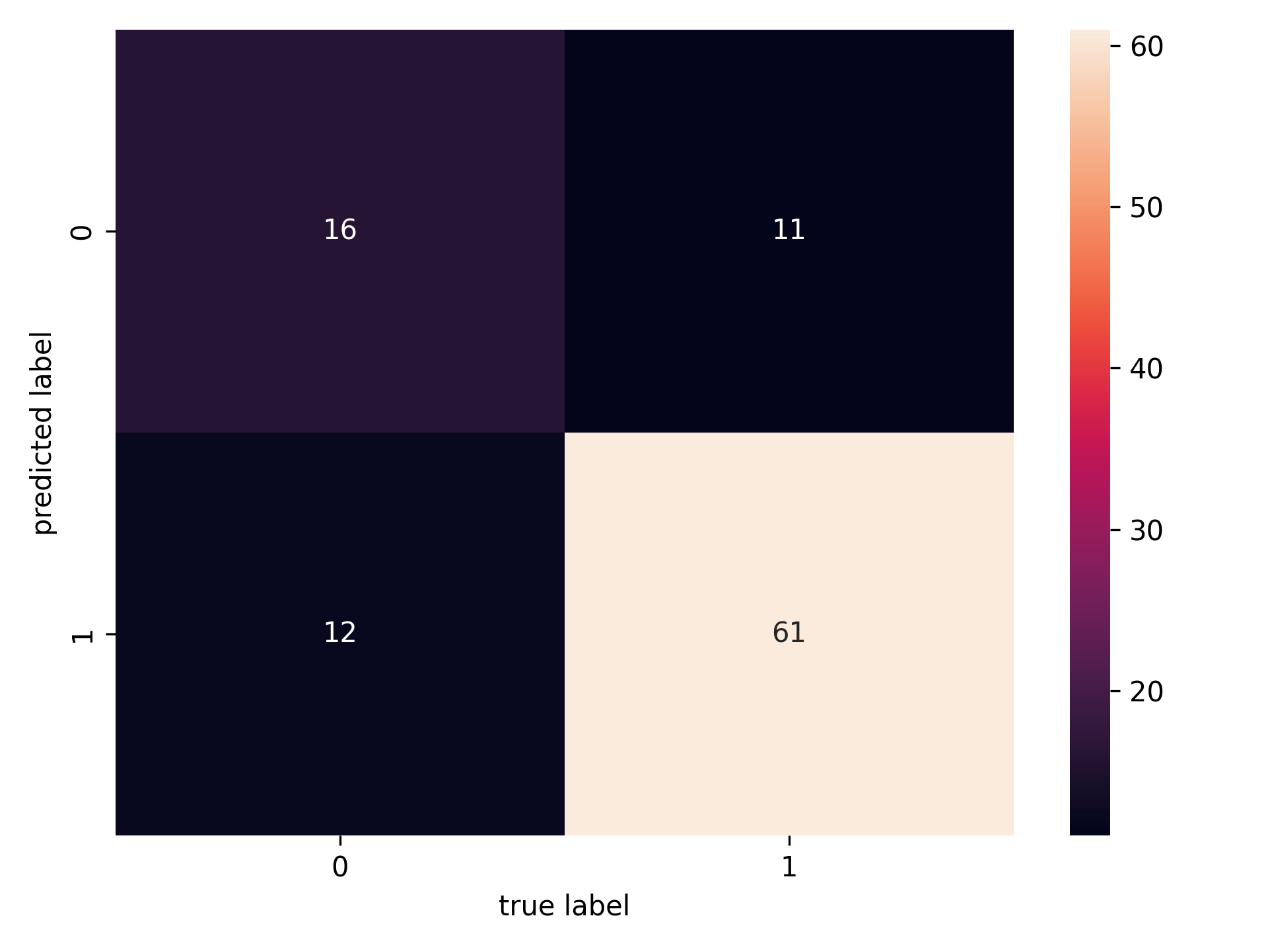


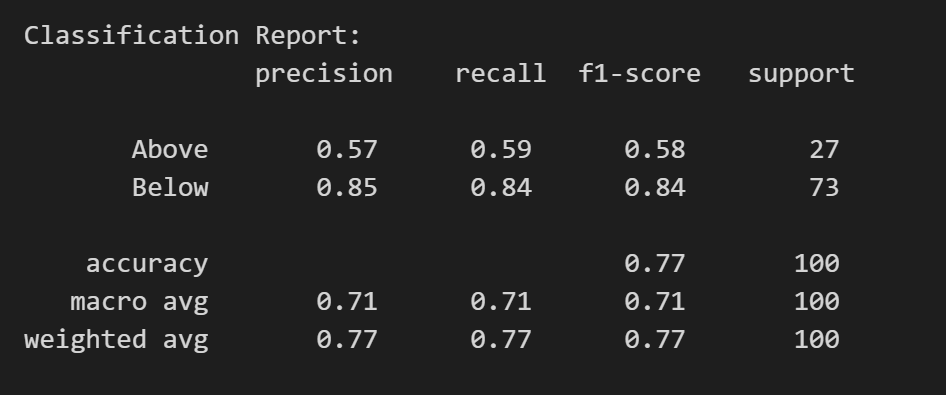
To improve the model accuracy, we could use:

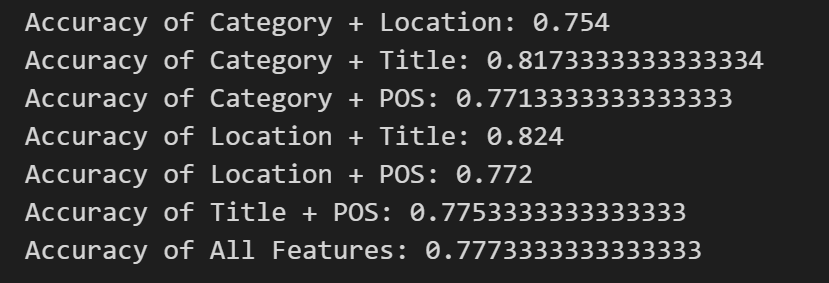
**N-gram**: Trigrams are based on three words, whereas bigrams are tokens of words based on two words each. These can be used in place of our unigram method to test whether accuracy is improved.

**Extra Features**: In addition to our earlier method, features like as Job Title, Location, Contract Type, and Category may be provided for every job posting. Since our salary projections are only dependent on text data, our model's performance may be greatly enhanced by include details like seniority, department, location, and full-time/part-time employment.

Therefore, I used the trigrams and bigrams combination, the accuracy improved from 0.746 to 0.77





Then I incorporate the variables “Title”,”Location” and “Category” with larger dataset (7500 rows) into the classification model for accuracy improvement, the result is like: 

which POS means the FullDescription after tagging.

Then, I tried using POS tagging for those extra variables, but the accuracy wasn’t improved to great extent.

