Selection of Unbalanced SVM Model:

After tuning for parameters, we opted for unbalanced SVM model with C = 0.5 to incorporate into our 3-component analysis. As can be seen from the comparison in Table 1. below, the unbalanced model out-performed the balanced model in accuracy, precision as well as recall by 0.001% to 0.005%.

|  |  |  |  |
| --- | --- | --- | --- |
| **Parameter** | **Accuracy** | **Precision** | **Recall** |
| Unbalanced | 98.54% | 98.50% | 98.54% |
| Balanced | 98.01% | 98.45% | 98.01% |
| Percent Out-performance | 0.005% | 0.001% | 0.005% |

**Table 1.** Comparison of Average Metrics for Unbalanced and Balanced SVM Models

Class-wise Metrics:

The accuracy, precision and recall for each class using unbalanced SVM model are summarized in Table 2. below.

|  |  |  |  |
| --- | --- | --- | --- |
| **Class** | **Accuracy** | **Precision** | **Recall** |
| ADDRESS | 0.18 | 0.16 | 0.18 |
| CARDINAL | 0.74 | 0.87 | 0.74 |
| DATE | 0.96 | 0.87 | 0.96 |
| DECIMAL | 0.79 | 0.95 | 0.79 |
| DIGIT | 0 | 0 | 0 |
| ELECTRONIC | 0.82 | 0.84 | 0.82 |
| FRACTION | 0.47 | 0.63 | 0.47 |
| LETTERS | 0.81 | 0.79 | 0.81 |
| MEASURE | 0.89 | 0.98 | 0.89 |
| MONEY | 0.96 | 0.99 | 0.96 |
| ORDINAL | 0.86 | 0.97 | 0.86 |
| PLAIN | 0.99 | 0.99 | 0.99 |
| PUNCT | 1 | 1 | 1 |
| TELEPHONE | 0.44 | 0.8 | 0.44 |
| TIME | 0.66 | 0.96 | 0.66 |
| VERBATIM | 0.88 | 0.86 | 0.88 |

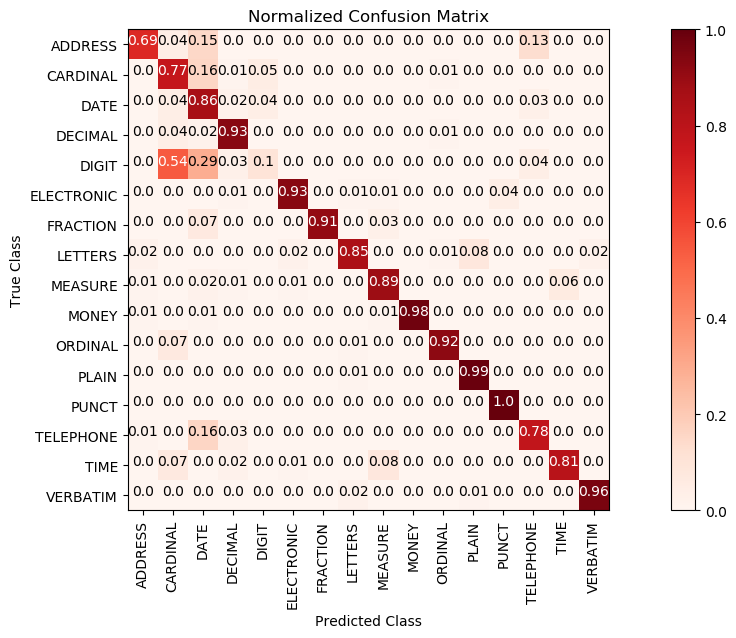
**Table 2.** Class-wise Metrics for Unbalanced Model

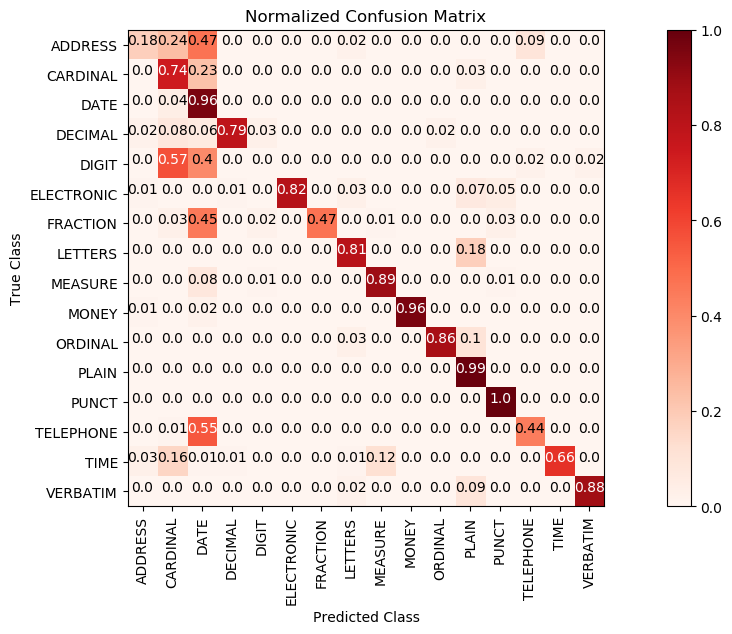
Confusion Matrix:

We plotted normalized confusion matrix of token-to-class transformation for both the unbalanced and balanced model as shown in Figure 1. below.

For both balanced and unbalanced SVM model, the general distribution of the matrices display heavy diagonal values, which denotes that we have satisfying prediction accuracy since most predicted values match their true values.

For the unbalanced SVM model, the classes with most confusion are mostly associated with numbers, such as ‘CARDINAL’, ‘DIGIT’ and ‘DATE’. A major reason is due to the limitation of TF-IDF transformation. Since the different classes of numbers have representations similar to each other, and TF-IDF is based on the frequency of each character, the model cannot easily distinguish between the different classes. As a result, the SVM model will most likely yield predictions based on the classes with dominant number of occurrences. Thus the classes that are rarely seen will be masked by those that are more frequently seen to guarantee overall accuracy. As can be seen from the performance of related previous work, the issue of correctly classifying numbers is still a common concern. [Ref]

For the balanced model, however, the accuracy for each class is higher. This is because the balanced model will try to predict each class accurately by scarifying the accuracy of the more frequently occurred classes. This approach yields a lower overall accuracy, which we showed from the comparison above.

**Figure 1.** Normalized Confusion Matrix for Unbalanced (left) and Balanced (right) SVM Class-wise Prediction