**Summary:** We resolved our data skewness issue with standard machine learning techniques: over-sampling and under-sampling, which makes our datasets balanced. As follows from the results reported in tables below, LSTM performs better when under sampling is used whereas HMM works well in over sampling technique. It was also observed that HMM provides higher results with dataset 1. On the other hand, LSTM shows superior performance when dataset 2 is used.

**Notes:** We calculated **macro averaged** results with the method Tsoumakas et al., 2010. However, **micro averaged** results were computed with Daelemans et al. (2010). Micro averaged results are really a measure of effectiveness on the large classes in a test collection. To get a sense of effectiveness on small classes, you should compute macro averaged results (Manning et al., 2008).

**Macro Averaged 10 Folds Cross Validation Results**

**Date: July 27, 2017**

**Dataset 1: 37 MI transcripts (successful: 880, unsuccessful: 192); # unique codes: 41**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Data Sample** | **Method** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| Under Sampling (successful: 192, unsuccessful: 192) | HMM | 0.6657 | 0.6743 | 0.6663 | 0.6586 |
| LSTM | **0.7564** | **0.7596** | **0.7626** | **0.7528** |
| Over Sampling (successful: 880, unsuccessful: 880) | HMM | **0.7755** | **0.7828** | **0.7755** | **0.7731** |
| LSTM | 0.5153 | 0.5341 | 0.5101 | 0.4112 |

**Dataset 2: 106 MI transcripts (successful: 3427, unsuccessful: 742); # unique codes: 125**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Data Sample** | **Method** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| Under Sampling (successful: 742, unsuccessful: 742) | HMM | 0.5837 | 0.6128 | 0.5852 | 0.5523 |
| LSTM | **0.7577** | **0.7624** | **0.7548** | **0.7541** |
| Over Sampling (successful: 3427, unsuccessful: 3427) | HMM | **0.7564** | **0.8023** | **0.7542** | **0.7455** |
| LSTM | 0.5553 | 0.5619 | 0.5535 | 0.5405 |

**Micro Averaged 10 Folds Cross Validation Results**

**Date: July 27, 2017**

**Dataset 1: 37 MI transcripts (successful: 880, unsuccessful: 192); # unique codes: 41**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Data Sample** | **Method** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| Under Sampling (successful: 192, unsuccessful: 192) | HMM | 0.6657 | 0.6813 | 0.6657 | 0.6619 |
| LSTM | **0.7564** | **0.7743** | **0.7564** | **0.7573** |
| Over Sampling (successful: 880, unsuccessful: 880) | HMM | **0.7755** | **0.7844** | **0.7755** | **0.7739** |
| LSTM | 0.5153 | 0.5381 | 0.5153 | 0.4159 |

**Dataset 2: 106 MI transcripts (successful: 3427, unsuccessful: 742); # unique codes: 125**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Data Sample** | **Method** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| Under Sampling (successful: 742, unsuccessful: 742) | HMM | 0.5837 | 0.6142 | 0.5837 | 0.5522 |
| LSTM | **0.7577** | **0.7628** | **0.7577** | **0.7559** |
| Over Sampling (successful: 3427, unsuccessful: 3427) | HMM | **0.7564** | **0.8013** | **0.7564** | **0.7462** |
| LSTM | 0.5553 | 0.5619 | 0.5553 | 0.5415 |