**Models compared:**

1. TF-IDF vectors, logistics regression
2. Word2Vector embedding, logistics regression
3. BioBert embedding, logistics regression
4. Word2Vector embedding, 200 hidden layers DNN
5. BioBert embedding, 200 hidden layers DNN
6. Ensemble 3+4
7. BertForSequence classification

**Features engineering:**

1. Only embeddings
2. Embeddings + ecig term
3. Embeddings + document template type
4. Embeddings + ecig term + document template type

**Model comparisons for deep learning models:**

Python notebook of training transformers based models and model optimization is “Ecig Transformer-training” in classification\_2022 folder.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | Performance Overview | Categories | Recall | Precision | F1 Score | Support |
| Pyspark LinearSVC  With pretrained W2V embeddings on clinical text from SparkNLP | Accuracy: 0.79 | Active User | 0.78 | 0.90 | 0.84 | 614 |
| Weighted Precision: 0.84 | Usage Unknown | 0.89 | 0.89 | 0.89 | 415 |
| Weighted Recall: 0.79 | Non User | 0.60 | 0.40 | 0.48 | 93 |
| Weighted F1: 0.81 | Irrelevant | 0.52 | 0.33 | 0.40 | 27 |
|  | Former User | 0.25 | 0.06 | 0.10 | 8 |
|  |  |  |  |  |  |  |
| BioBert Sequence Classification trained on  1 sentence  (5-17-epoch-5) | Accuracy: 0.88 | Active User | 0.92 | 0.92 | 0.92 | 408 |
| Weighted Precision: 0.89 | Usage Unknown | 0.94 | 0.91 | **0.93** | 327 |
| Weighted Recall: 0.88 | Non User | 0.68 | **0.73** | **0.70** | 102 |
| Weighted F1: 0.88 | Irrelevant | 0.72 | 0.84 | 0.78 | 32 |
|  | Former User | 0.44 | 0.52 | 0.47 | 32 |
|  |  |  |  |  |  |  |
| BioBert Sequence Classification trained on  3 sentences  (5-19-epoch4) | Accuracy: 0.88 | Active User | 0.93 | 0.92 | **0.93** | 408 |
| Weighted Precision: 0.88 | Usage Unknown | 0.92 | **0.93** | 0.92 | 327 |
| Weighted Recall: 0.88 | Non User | **0.69** | 0.66 | 0.67 | 102 |
| Weighted F1: 0.88 | Irrelevant | **0.79** | 0.79 | 0.79 | 32 |
|  | Former User | 0.56 | 0.64 | 0.60 | 32 |
|  |  |  |  |  |  |  |
| Ensemble of above two models | **Accuracy: 0.89** | Active User | **0.94** | **0.93** | **0.93** | 408 |
| **Weighted Precision: 0.90** | Usage Unknown | **0.94** | 0.92 | **0.93** | 327 |
| **Weighted Recall: 0.89** | Non User | **0.69** | 0.71 | **0.70** | 102 |
| **Weighted F1: 0.89** | Irrelevant | 0.76 | **0.85** | **0.80** | 32 |
|  | Former User | **0.56** | **0.67** | **0.61** | 32 |
|  |  |  |  |  |  |  |
| W2V embdings and trained with simple DNN | Accuracy: 0.83 | Active User | 0.87 | 0.87 | 0.87 | 408 |
| Weighted Precision: 0.84 | Usage Unknown | **0.94** | 0.88 | 0.91 | 327 |
| Weighted Recall: 0.83 | Non User | 0.54 | 0.64 | 0.59 | 102 |
| Weighted F1: 0.83 | Irrelevant | 0.48 | 0.64 | 0.55 | 32 |
|  | Former User | 0.41 | 0.37 | 0.39 | 32 |
| BioBert embdings and trained with simple DNN | Accuracy: 0.83 | Active User | 0.91 | 0.82 | 0.86 | 408 |
| Weighted Precision: 0.85 | Usage Unknown | 0.92 | 0.90 | 0.91 | 327 |
| Weighted Recall: 0.83 | Non User | 0.50 | 0.68 | 0.58 | 102 |
| Weighted F1: 0.84 | Irrelevant | 0.55 | **1.00** | 0.71 | 32 |
|  | Former User | 0.12 | 0.19 | 0.15 | 32 |
|  |  |  |  |  |  |  |
| Ensemble with above 2 models | Accuracy: 0.85 | Active User | **0.94** | 0.83 | 0.88 | 408 |
| Weighted Precision: 0.88 | Usage Unknown | 0.92 | 0.91 | 0.92 | 327 |
| Weighted Recall: 0.85 | Non User | 0.52 | 0.75 | 0.61 | 102 |
| Weighted F1: 0.86 | Irrelevant | 0.59 | **0.89** | 0.71 | 32 |
|  | Former User | 0.16 | 0.31 | 0.21 | 32 |
|  |  |  |  |  |  |  |
| **Final model** | | | | | | |
|  | | | | | | |
| Continue trained from above 3 sentences (5-19-epoch4) on full set 1 and set 2 data, and use test 3 as validation data. | Accuracy: 0.90 | Active User | **0.94** | 0.91 | **0.93** | 408 |
| Weighted Precision: 0.89 | Usage Unknown | **0.94** | **0.93** | **0.93** | 327 |
| Weighted Recall: 0.90 | Non User | **0.73** | **0.79** | **0.76** | 102 |
| Weighted F1: 0.89 | Irrelevant | **0.75** | 0.86 | **0.80** | 32 |
| AUC (one vs rest: 0.962  one vs one: 0.936) | Former User | **0.50** | **0.70** | **0.58** | 32 |