horizontal line

Sprint 7 Results

Sprint Created: 11.24.2019

Sprint End: 12.01.2019

Sprint Captain: Yirou Ge

Sprint Co-captain: Hongying Liu

**─**

# Sprint Goal

## Optimize Classification:

* Implement stratified k-fold cross validation to choose train and validation datasets for neural network model **on track**
* Try to directly use RoBERTA to do the classification **on track**
* Try other methods for dimension reduction **on track**

## Feature Engineering and Label Model Optimization:

* This week, we will try to utilize roBERTa or other python packages to perform Entity Recognition **on track**

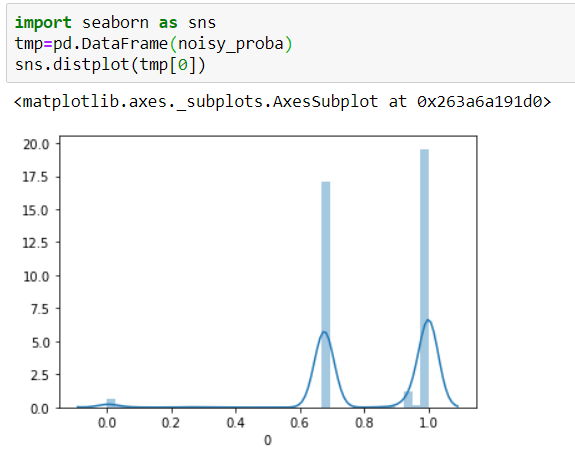
## Babblelabble:

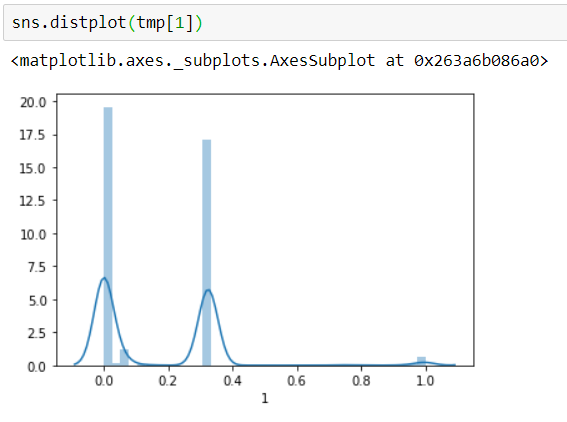
* This week we will mainly focus on writing more explanations and test the results.**on track**
* We will learn more about whether we can automate the whole process and apply them directly to the whole dataset instead of some random examples.**on track**

# Results

1. The distribution of Snorkel results:

Based on the following distribution plot of snorkel results, we found that over 90% of the data are classified into ‘non-positive’ category with a probability around 0.7 or 0.9. Such a distribution resulted in an extremely imbalanced dataset but we are not sure if there are any reasonable solutions to this.





1. Neural Network Results
2. **Feature engineering**

This week, we modified the neural network model to concatenate additional features after the embeddings are processed through the LSTM layer. Besides, we generated more features and tried them out on the model as well.

The features that we’ve used include:

**N-gram**: unigram

**POS tags**: the count of pos tags NN, IN, JJ, NNS, DT, CC, CD, and VB

**Sentiment Score**

**Topic Modeling**: unigram/bigram

**TF-IDF**

**Word Counts**: including char\_count, word\_count, punctuation\_count, title\_word\_count, and upper\_case\_word\_count

By far, the best combination that we’ve found includes N-gram, POS tags, sentiment score, and topic modeling.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Metrics/  Models | Baseline Model | Noisy Model | N-gram+POS+Sentiment  (batch 50) | N-gram+POS+Sentiment  (batch 20) | **N-gram+POS+Sentiment+Topic**  **(batch 20)** | N-gram+POS+Sentiment+Topic+Counts |
| F1 Score | 0.509 | 0.4 | 0.383 | 0.488 | **0.516** | 0.510 |
| Accuracy | 0.58 | 0.63 | 0.62 | 0.644 | **0.665** | 0.651 |
| ROC AUC Score | 0.525 | 0.51 | 0.495 | 0.513 | **0.536** | 0.527 |
| Matthews Correlation Coefficient |  | 0.0165 | -0.049 | 0.036 | **0.102** | 0.073 |

1. **Batch Size**

This week, we also decreased the original batch size of 50 into 20, which helped increase the model’s performance. The detailed evaluation scores can be found in the above chart.

1. **Stratified K-Fold Cross Validation**

In addition, we also added stratified k-fold functions for cross validation (with features of N-gram+POS+Sentiment+Topic). The model’s performances on the validation data (part of the noisy dataset) and the test data (gold standard dataset) are as follows.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | F1 Score | Accuracy | ROC AUC | MCC |
| validation | 0.578 | 0.979 | 0.547 | 0.244 |
| test | 0.486 | 0.656 | 0.517 | 0.053 |

1. Similarity Matrix:

We calculated the cosine similarity among the 22 policy documents in Malawi, and highlighted the documents whose similarity coefficient is greater than 0.5.

