Functional requirements for multi-label model generation using Label Sleuth.

Version: 1

Author: Radu Bengulescu MD, Healthcare Informatics Ph.D.

a. Process description: Generate a deep learning model that would allow the classification of a given chunk of text into one of the categories (labels) from a given set of categories. A model generation is completed when ALL the categories from the workspace have the required number of text chunks labeled
b. Pre-requisites: Datasets containing in a tabular form chunks of text, already labeled by humans (hence considered as correct) that will serve as basis for training / validation (fig. 1)

Workflow steps:

1. The user loads in the system a dataset containing several chunks of text, already labelled Fig. 1
2. The user creates in the system the required labels that would match the one contained in the loaded dataset
3. The user selects the first label in the label selection form
4. The system filters in the background the dataset according to the selected label and starts presenting to the user for manual checking the corresponding text chunks associated with the selected label. Each text chunk should be displayed containing the current label associated in the dataset with the text chunk so that the user can visually check that he is working on the correct text chunk associated with the selected label (Fig. 2).
5. The system continues to offer for checking the text chunks until the threshold needed for a partial model generation is reached (a partial model represents a model that has only the component corresponding to one label completed).
6. In the moment a partial model is generated there should be a visual message to the user informing that the labelling is over for that particular model and the user should proceed to checking the text chunks corresponding to another label. Also, the system should visually display in the workspace the information about the already completed label and the model version that is in work (if any of it has been generated yet). Fig 3
7. Also, in the label selection form (dropbox, etc….) the label that is already completed should be somehow visually marked so that the user can select another label for working on it. Fig. 4
8. The process continues until all the labels in the label selection form are completed. In that moment the system starts generating a model version and informs the user about the success of this operation. It should also prompt the user if he wants to save the generated model in zip format (see below the process description). Also, it will update the information displayed in the workspace about the model version. The information displayed out about the status of completed labels is erased. The label selection form is reset so that all labels are in in initial color / state. The information about the overall model accuracy is displayed.
9. The user starts working on a subsequent model version. The system offers the possibility of chained-labelling by selecting form the dataset only the text chunks corresponding to the selected label and displays the „Label Next“ functionality.

10. The process described in the steps 4-7 until a new model version is generated.
11. At any time when a full model version is available in the system there should be a „Test model“ function that would apply the current model version available in the system to the text chunks from the dataset that have an associated label and have not yet been labeled by the system, calculate the accuracy / recall rate (or any other suitable parameter that describes how the process model performs against a test set) and then display the model information visually. Also, it should allow the exporting of the calculated dataset containing the original label (human assigned), the system predicted label and the accuracy information.

document_id	text	url	tag
118513-Ulcer	mr known lastname is a	localhost	Ulcer
100473-Pneumonia	year old man with histo	localhost	Pneumonia
101847-AbdominalPain	61m w hiv hcv cirrhosis	localhost	Abdominal Pain

Fig. 1

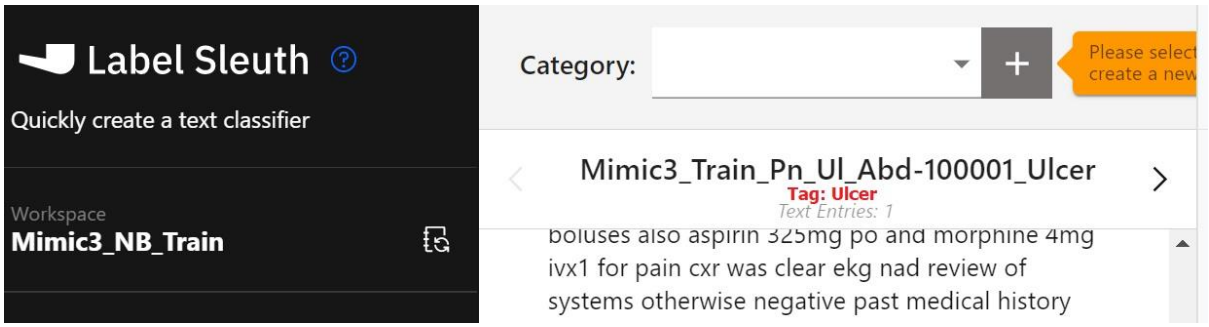

Fig. 2

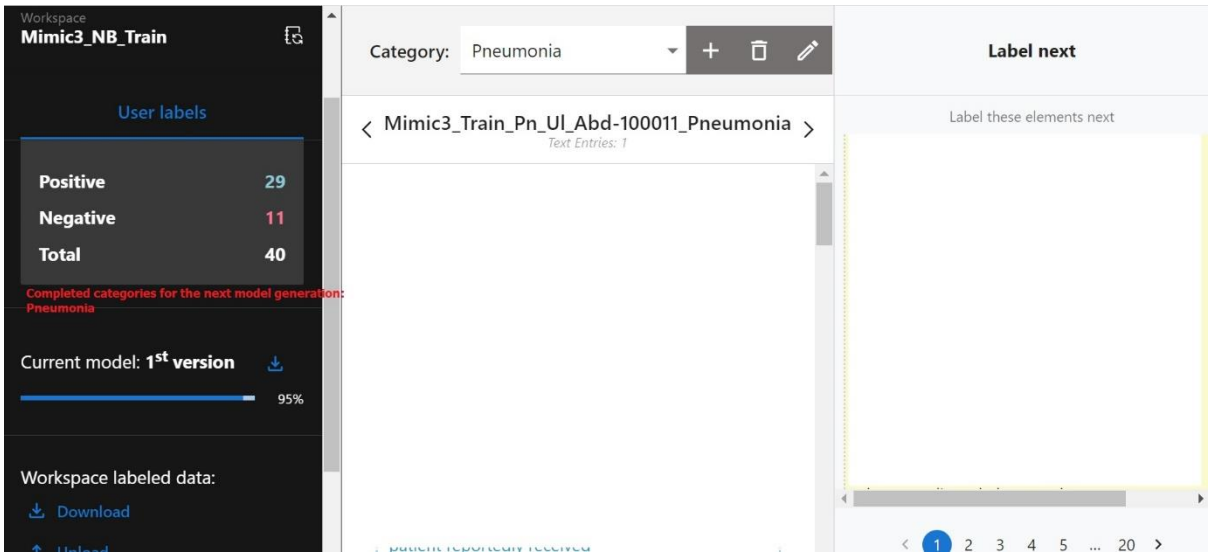

Fig. 3

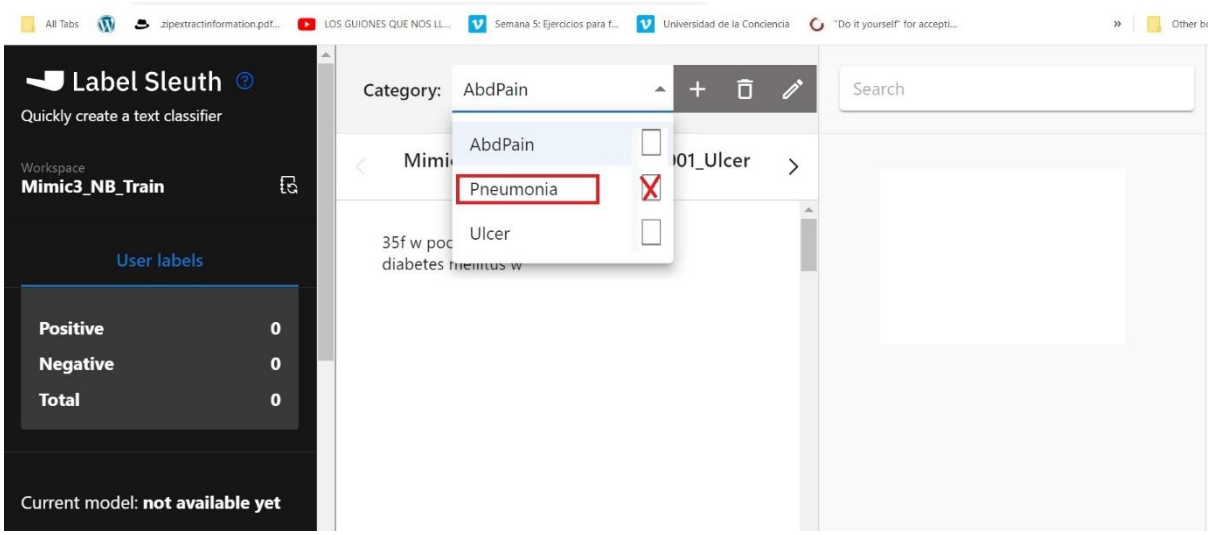
Label Sleuth
Quickly create a text classifier
Workspace
Mimic3_NB_Train
User labels
Positive 0
Negative 0
Total 0
Current model: not available yet
Category:
AbdPain
Search
AbdPain
Pneumonia
Ulcer
35f w poc
diabetes mellitus w

Fig. 4