Multi-Task Learning Strategies for NLI: Insights from Model Performance and Noise Dynamics

Introduction

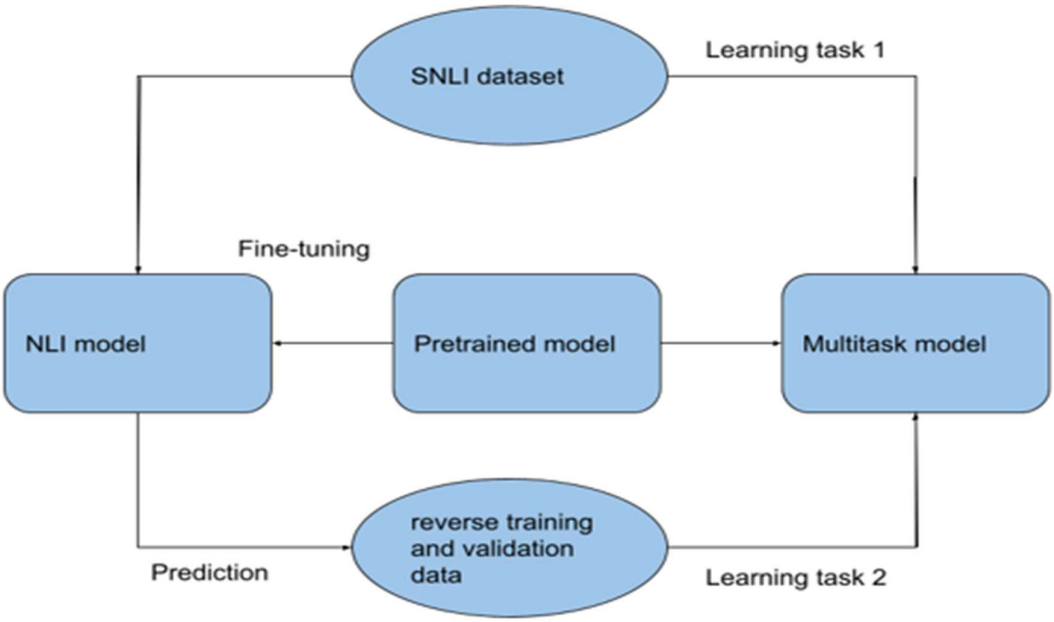
Exploring the potential of multi-task learning in the context of Natural Language Inference (NLI) constitutes the core objective of this study. The central proposition is to investigate whether the implementation of multi-task learning can yield enhancements in the performance of an NLI model. Our approach involves the creation of a multi-task model designed to address two distinct tasks. The first task involves a straightforward NLI assignment utilizing the SNLI dataset. For the second task, we introduce a variation in which the SNLI dataset is manipulated, resulting in an inversion of premise and hypothesis sentences. This innovative alteration enables the simultaneous learning of two senses from a single example. The anticipated outcome is an augmented understanding that contributes to the overall performance of the primary task. Through this research, we aim to uncover the potential benefits of employing multi-task learning techniques to elevate the efficacy of NLI models.

Methods

To address the research objective of enhancing the performance of Natural Language Inference (NLI) models through multi-task learning, we adopt a systematic approach involving diverse models.

The initial phase entails training various NLI models, each utilizing distinct pretrained models. This diversity in pretrained models ensures that we explore the spectrum of multi-task NLI model capabilities comprehensively.

The process starts with training NLI models for two primary reasons: first, to benchmark their individual performances; second, to generate data for the subsequent task. As the second task involves reversing premise and hypothesis sentences, new data rows must be assigned appropriate labels. To accomplish this, we employ the NLI models trained earlier to predict the labels for these new data points. This process is illustrated as follows:



Four distinct NLI models are trained, leading to five different multi-task models. Three models are based on the same codebase, varying only in the pretrained model employed: "bert-base-cased," "roberta-base," and "bert-base-multilingual-cased." These models are trained using a subset of the SNLI dataset, containing the initial 50,000 rows of training data with rows having a label of -1 removed.

The fourth NLI model employs a different strategy, utilizing "roberta-base" but in a novel manner. Instead of using dataset labels, a mechanism is developed to estimate label-specific probabilities. This is achieved by leveraging annotations from multiple annotators (five in total) present in the "label i" columns ($i \in \{1,\ldots,5\}$). The labels predicted by each annotator are used to compute their accuracy for each label, to do that we consider the “gold_label” of the dataset as the true label. This creates a vector of probabilities which represents the accuracy of the annotator for each label. For each row we create a vector which is the sum of annotator probabilities on the labels predicted. By normalizing the resulting vectors, probabilities of belonging to each label are obtained. This nuanced approach

minimizes the impact of inaccuracies from individual annotators and enriches the label set for improved model comprehension.

As an example:

- Annotator A's label accuracy: [0.6, 0.8, 0.7]

- Annotator B's label accuracy: [0.2, 0.9, 0.9]

- For a specific row, A predicted 2 and B predicted 1. The normalized label vector: [0, 0.5625, 0.4375]

For multitask model, two datasets are required: one for the primary task (SNLI dataset, possibly with label modifications) and the other for the reversed premise-hypothesis task. In the latter, hypothesis and premise columns are swapped, and labels are replaced with predictions generated by the previously trained NLI models. This setup results in three similar models, each utilizing "roberta," "bert," and "multilingual bert."

Two distinctive "roberta"-based models are also introduced:

1. "full_probas": This model focuses on learning from probabilities, wherein predictions are based on probabilities rather than raw labels for both tasks. The probabilities of the first task are the same than the fourth NLI model.

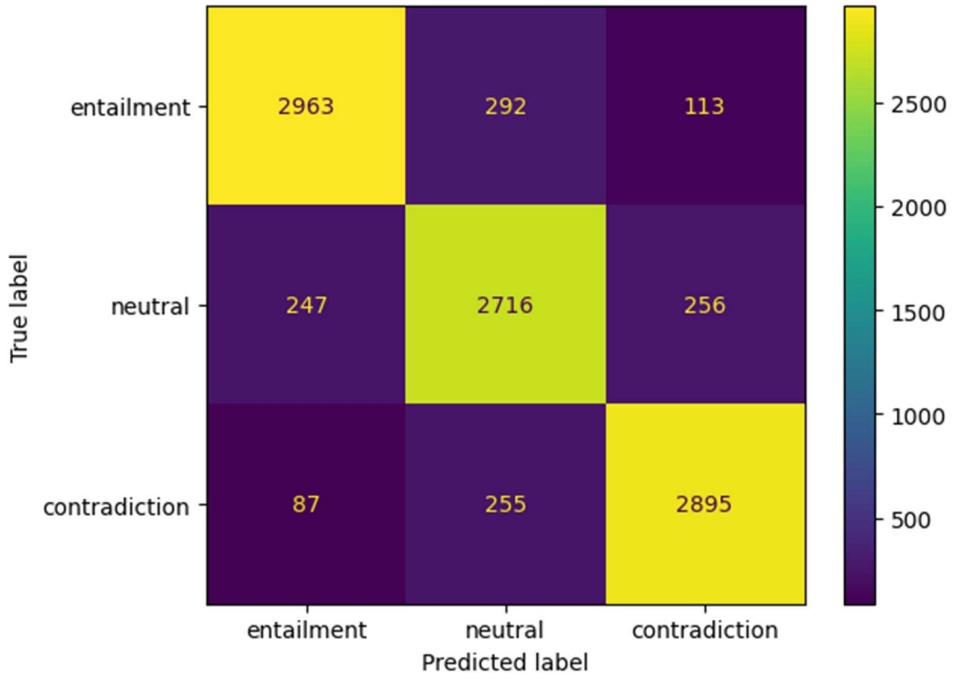
2. "general_probas": Based on the principle that entailed sentences cannot contradict if you inverse the sense of comparison (premise, hypothesis) and vice versa, this model employs a unique approach. For the second task, a "truncated probabilities" dataset is constructed, where predictions are adjusted based on the SNLI labels. This ensures the new label vector distribution conforms to the SNLI label. For example, if in a row of the SNLI dataset the label is 0 (entailed), and our predictions for the “reverse” first row is [0.2, 0.6, 0.2] we will put the probability of this example is 2 (contradiction) to 0 and distribute the value. The new label vector is [0.3, 0.7, 0].

Through these comprehensive strategies, we aim to harness multi-task learning to enhance the performance and understanding of NLI models.

Results

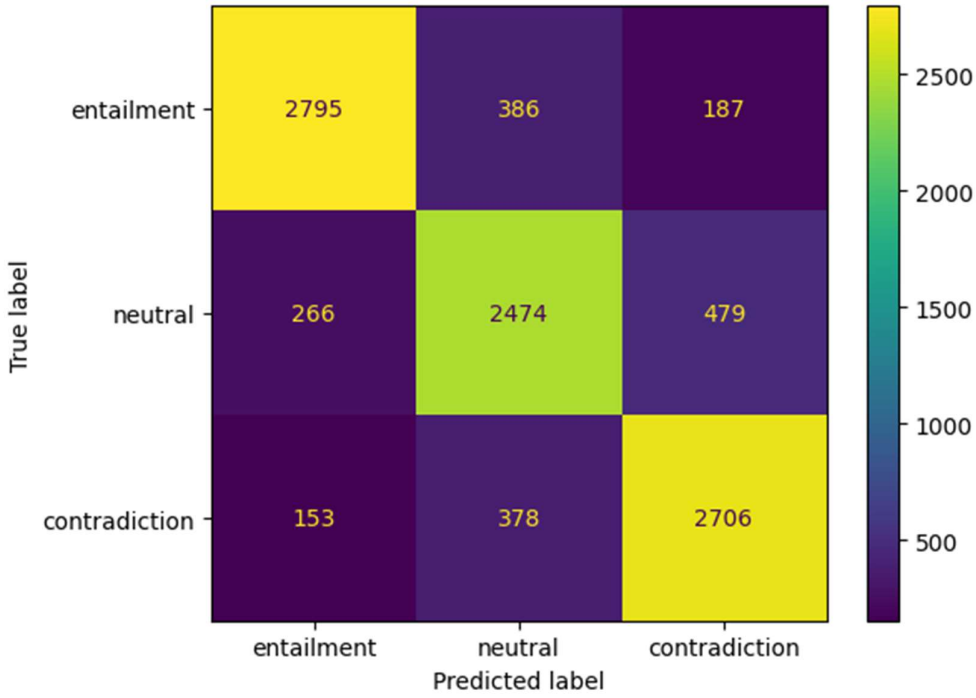
To facilitate model comparison, we adopted a dual evaluation approach involving accuracy metrics and analysis of confusion matrices. To evaluate those models, we predict labels on SNLI test set and compare it to the labels of the same dataset. Our evaluation starts with an examination of the NLI models.

Probabilistic model:



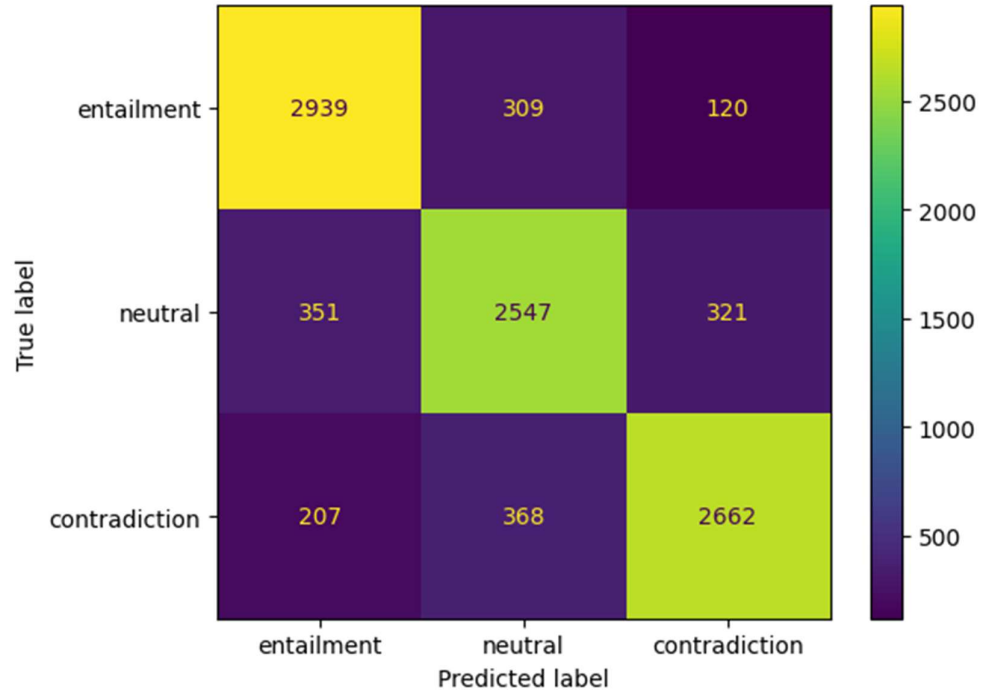
accuracy : 0.8727605863192183

Multilingual bert model:



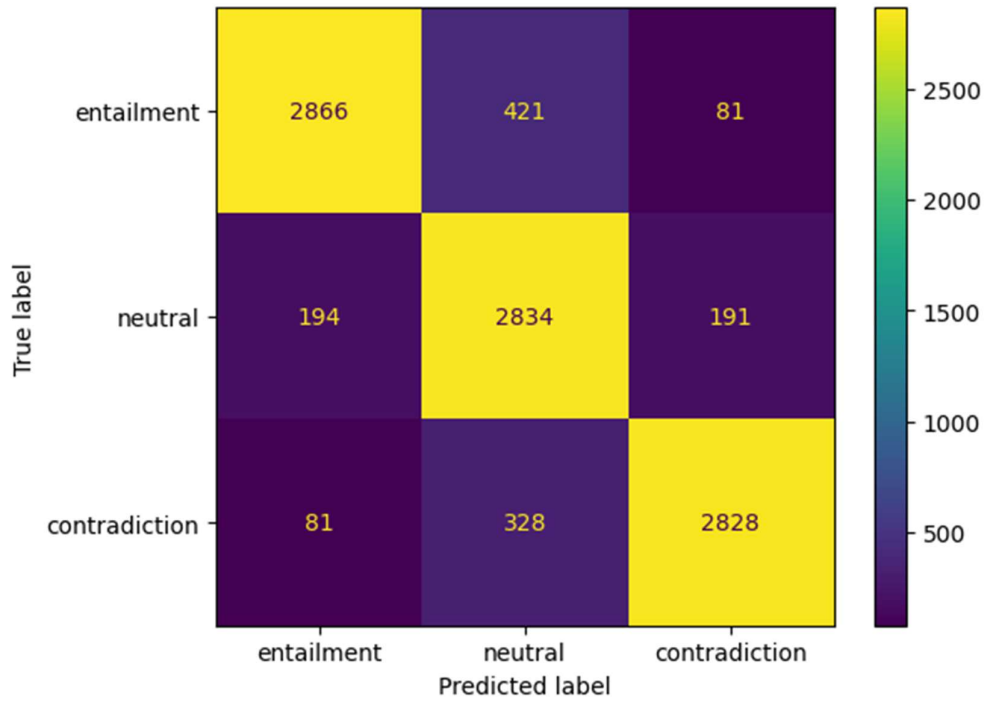
accuracy : 0.8117874592833876

Bert model:




```
0.8293973941368078
```

Roberta model:




```
accuracy : 0.8680781758957655
```

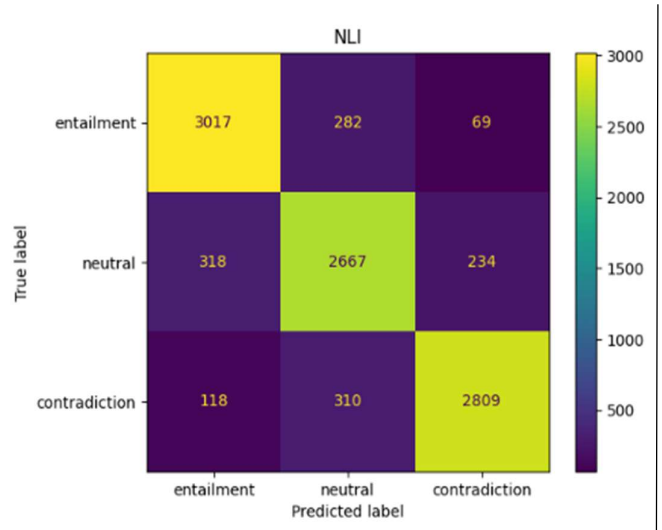
As expected, the "roberta-base" pretrained model exhibited superior performance compared to its counterparts, "bert-base-cased" and "multilingual-bert-base-cased." The advantage of "roberta-base" can be attributed to its larger training dataset and more effective training procedure, which contributed to its enhanced performance in comparison.

However, direct comparison between the probabilistic model and the "roberta" model is complex due to unoptimized hyperparameters. Moreover,, Roberta model could be trained on more than 50 000 rows which could improve its performance but the probabilistic can only be trained on 32 000 rows because only this rows has 5 annotations in the dataset.

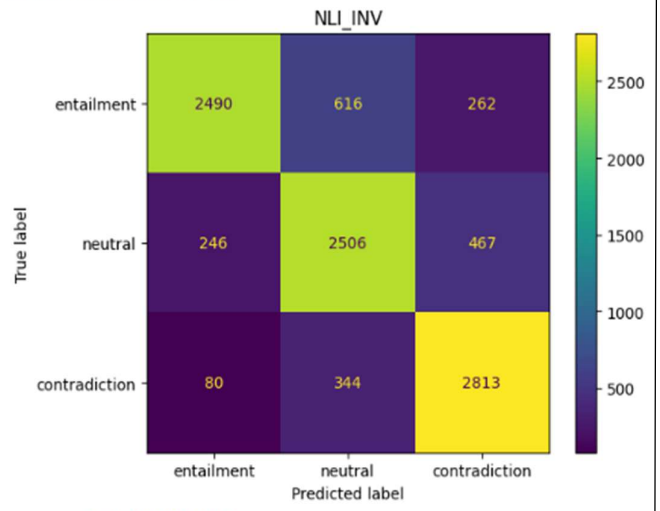
More generally challenges in predicting neutral labels were observed, potentially stemming from nuanced homonymic interpretations and intricate relationships between sentences. We can assure that it is not due to the number of examples because the training data was balanced between labels.

Now we study the results for multitask models :

Full probas model :

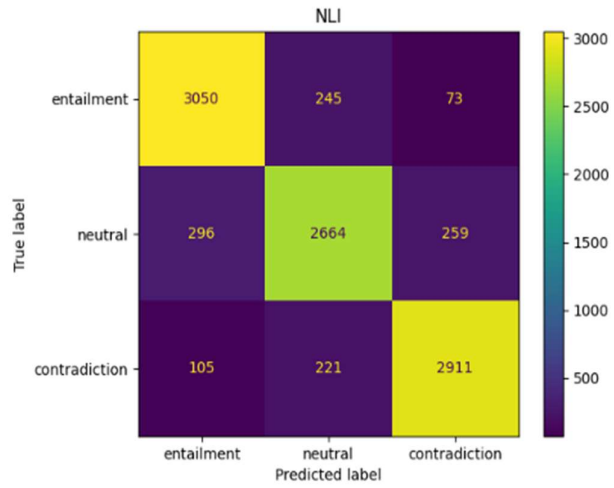


accuracy: 0.8645154723127035

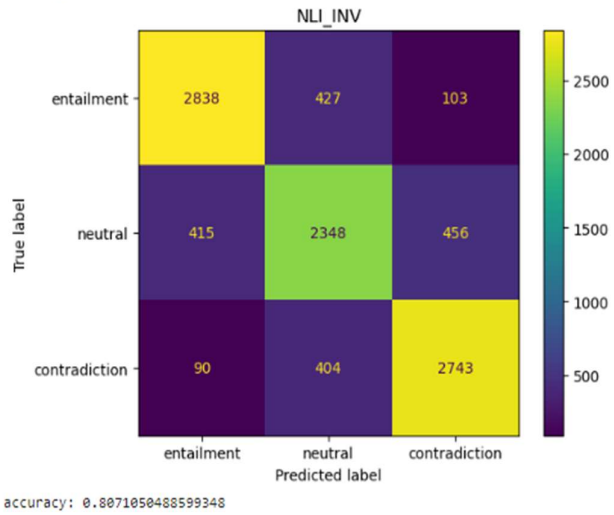


accuracy: 0.7948900651465798

G probas model :

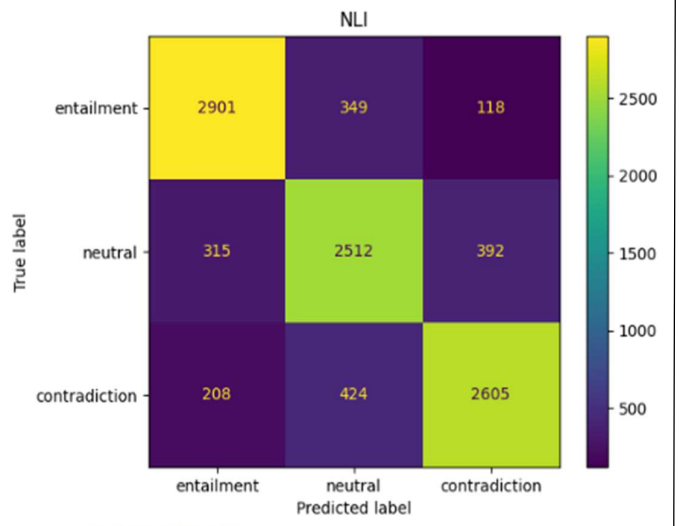


accuracy: 0.8779519543973942

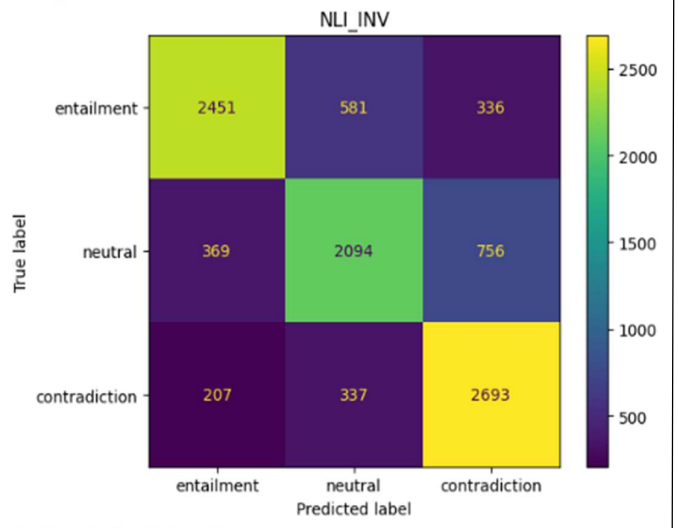


accuracy: 0.8071050488599348

Bert model :

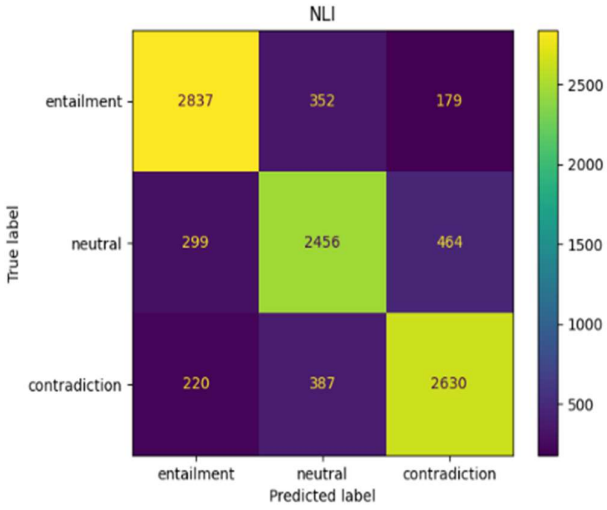


accuracy: 0.8161644951140065

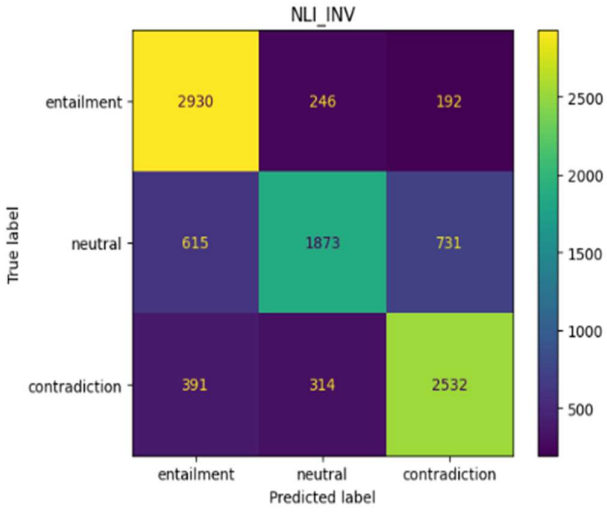


accuracy: 0.7367671009771987

Multilingual BERT model :

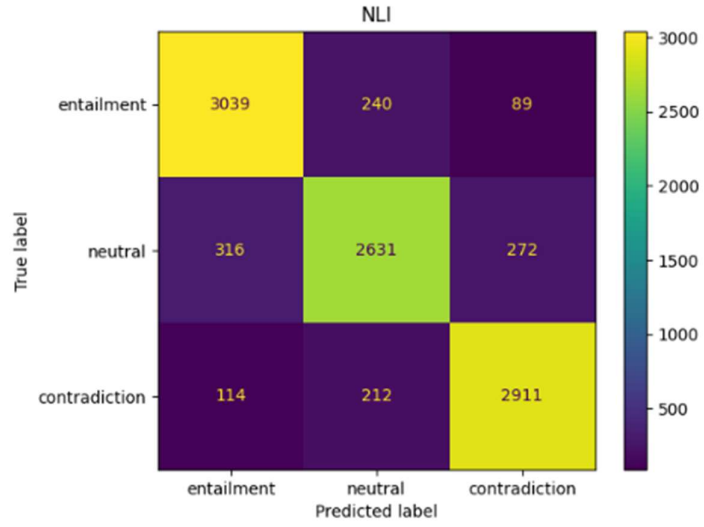


accuracy: 0.8064942996742671

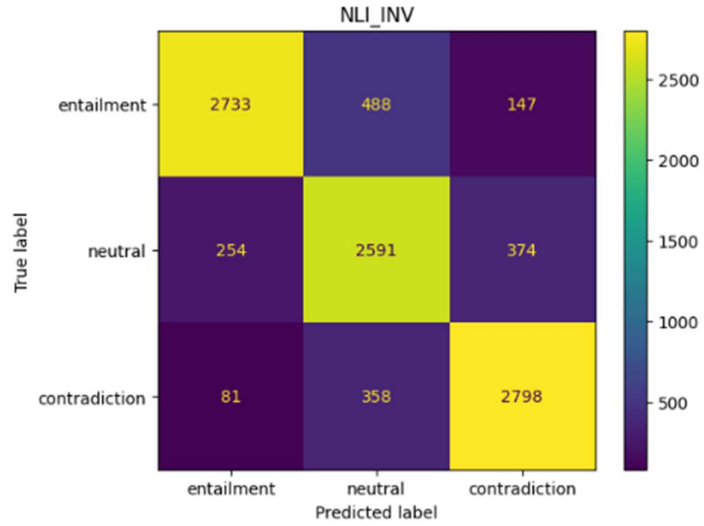


accuracy: 0.7466408794788274

RoBERTa model :



accuracy: 0.8734731270358306



accuracy: 0.8267508143322475

A clear pattern emerged where the first task's (NLI) performances consistently outperformed those of the second task (NLI_INV). This discrepancy is rationalized by the second task's reliance on predictions, possibly introducing noise to the training process. Strikingly, the "bert" and "multilingual-bert" models exhibited notably poor performance on the second task, indicating that knowledge sharing in these cases amplified noise rather than benefiting the model.

In contrast, other models demonstrated performance patterns closely aligned with those of individual NLI models. An intriguing trend surfaced wherein models exhibited a bias toward predicting fewer neutral labels compared to other labels. That is due to a bias in prediction of neutral label in NLI models. Furthermore, certain models, notably "g_probas" and the "roberta" model, displayed marginal improvements compared to the baseline NLI "roberta" model. While these enhancements lack statistical significance due to the absence of hyperparameter fine-tuning, they provide a promising direction for further exploration.

Conclusion

In this study, we focused on an exploration of multi-task learning’s potential to enhance NLI models. During this exploration we go through different type of models and training methodologies.

As anticipated, the supremacy of the "roberta-base" pretrained model underscored the importance of training procedures and dataset sizes. This finding emphasizes the significance of leveraging advancements in pretraining techniques to achieve superior model performance. However, the nuances introduced by the probabilistic model underscore the intricate balance between model optimization and data enrichment.

Multi-task models exhibited intriguing dynamics, with the first task consistently outperforming the second. The challenge posed by noise propagation in certain models emphasized the critical role of well-defined task relationships and noise management mechanisms. The observed reluctance of models to predict neutral labels warrants further investigation into potential model biases.

The modest performance enhancements showcased by select models, such as "g_probas" and the "roberta" model, offer a glimpse into the potential benefits of model refinement. While these improvements remain preliminary, they beckon further exploration through hyperparameter tuning and expanded datasets.

In conclusion, this research serves as a stepping stone toward an enhanced understanding of multi-task learning's impact on NLI models. The complex landscape of model performance, task interdependence, and data nuances unveiled here provides a solid foundation for future endeavors. By delving deeper into model design, dataset intricacies, and optimization strategies, we envision unlocking the true potential of multi-task learning to drive transformative advancements in the field of Natural Language Inference.