Homework 2 MNLP

Anonymous ACL submission

1 Introduction of the Datasets

In this homework assignment, I tackled a Natural Language Inference (NLI) task. Initially, I had to test various transformer-based models on the dataset I had, and then evaluate my results on two different test sets. The first was the original dataset's test set, while the second was an adversarial test, designed to be more challenging and rigorously test the models I implemented.

As mentioned earlier, I was provided with two datasets for my task. The first dataset consists of 55,000 samples divided into: approximately 51,000 samples for training, around 2,000 for the validation set, and the remaining 2,000 for the test set. Each sample had several columns available: ID, Premise, Hypothesis, Label, WSD, and SRL (which I will discuss in more detail later, including their specific uses and how they were optimally leveraged). The purpose of the other columns is straightforward. Is important to specify that the Label column contains three different values: Entailment, which is assigned when the hypothesis is correct given the premise; Contradiction, which is assigned when the hypothesis is incorrect given the premise; and Neutral, which is assigned when there is insufficient information to determine the validity of the hypothesis.

I approached my task by first performing checks on the provided dataset and immediately noticed, as shown in [Figure 1], that the training set is imbalanced. It contains 60.9% samples with the label "Entailment," 24.1% with "Contradiction," and 14.9% with "Neutral." In contrast, the validation and test sets are balanced.

The second dataset, on the other hand, consists simply of a test set and 337 samples.

2 Models and Evaluations

I will now present my implementations for this task. I tested several transformer models and concluded that the two most interesting and performant were RoBERTa and DistilBERT.

RoBERTa (Robustly optimized BERT approach) is an advanced variant of BERT (Bidirectional Encoder Representations from Transformers) that improves on BERT's performance by training with more data and longer sequences. RoBERTa is known for its strong performance on various NLP tasks due to its robust pre-training and fine-tuning processes.

DistilBERT is a smaller, faster, and more efficient version of BERT. It achieves this by applying knowledge distillation, a technique that compresses the original BERT model while retaining much of its performance. DistilBERT is designed to be less resource-intensive and faster, making it suitable for scenarios where computational efficiency is crucial.

Regarding the evaluation, each model I trained was assessed using: Accuracy, F1 score, Precision and Recall.

3 Results Default Dataset

Training the two models with the Default dataset initially yielded good results. Specifically, training the DistilBERT model first, I obtained an F1 score of 70.85% on the initial test set and 52.26% on the adversarial set. For RoBERTa, I achieved an F1 score of 74.79% on the default test set and 58.24% on the adversarial set.

These results align with my expectations, as mentioned earlier. RoBERTa is a more robust model compared to DistilBERT and generally has better performance.

However, what is concerning is the difference in execution time between the two models. RoBERTa takes more than twice as long as DistilBERT. Consequently, I can confidently say that while the first model is more performant, it comes with significantly longer training times. Therefore, choosing the better model depends on the specific needs: if long training times are not an issue, RoBERTa is certainly the more powerful model.

4 Data augmentation

For the second part of my task, I had to create an adversarial training dataset starting from the default dataset, thus generating different types of samples. To achieve this, I utilized the aforementioned WSD and SRL columns to analyze the various samples as thoroughly as possible and create new ones using various techniques. Below, I will analyze all the techniques I employed for creating the samples.

4.1 WSD

Regarding Word Sense Disambiguation (WSD), I chose to work using WordNet. Among the various methods I experimented with to create new samples, we have: substitution of synonyms for some adjectives within the premise [Figure 2], substitution of synonyms for some adjectives within the hypothesis [Figure 3], substitution with antonyms within the hypothesis [Figure 4], substitution with hypernyms within the hypothesis [Figure 5], and finally, substitution of verbs from positive to negative within the hypothesis [Figure 6].

Of course, some of these modifications required me to change the label outcomes for each sample I was creating. Specifically, I had to reverse the labels from ENTAILMENT to CONTRADICTION and vice versa when I created new samples using antonyms and negative verbs instead of positive ones.

For each method I used, I chose to modify either one word or one verb at a time to avoid conflicts. For instance, modifying two verbs could generate a double negation, which would conflict with the label.

4.2 SRL

As for Semantic Role Labeling (SRL), I chose four different approaches: the first involves the inversion of AGENT-PATIENT [Figure 7], the second inverts THEME-ATTRIBUTE, while in the third approach I decided to try something new. I noticed that in a significant portion of the samples from the default dataset, the age or birth year of some famous individuals was present. Consequently, by using the SRL column, I extracted this information from each sample and built new hypotheses based on it. The first calculates how many years ago a famous person was born (obviously referring to 2024, the current year) and returns the sample shown in [Figure 9]. The second, given the birth year, returns how old that person will be in 2050.

Although I found these methods interesting, I chose not to use the first two, as I noticed, as shown in Figure 7, that the sentences often didn't make much sense. Moreover, in some cases, the labels had to be changed, while in others they did not. Therefore, I decided to use only the last two SRL methods in my augmented dataset, while still leaving the implementations of the first two methods.

4.3 Dataset Augmentation

After experimenting with these data augmentation methods on my dataset, I created a new dataset with 25 thousand new samples, all generated using the previously presented methods. I could have made the dataset even larger, but I chose to take only a few samples from each method to ensure the dataset was as clean as possible.

5 Results Augment Dataset

Having created the new dataset, I chose to test it not only individually on my two previously described models, but I also decided to mix it with the default dataset to see how my created dataset would actually impact the standard one.

First, I tested the augmented dataset on both RoBERTa and DistilBERT, using both the default test and the adversarial test. The results were quite interesting: DistilBERT achieved performances of 62.85% and 47.97% on the default and adversarial tests, respectively, while RoBERTa reached 68.44% and 55.66%. All the results were slightly below those obtained with the default dataset.

What surprised me, however, were the results from the complete dataset, consisting of 75 thousand samples. I obtained 69.41% and 54.77% with DistilBERT, and more notably, 73.64% on the default test and 62.22% on the adversarial test with RoBERTa.

I thus achieved my best results on the adversarial test by combining the initial dataset with the one I generated, improving performance by 4%. On the default test, the best performance came from RoBERTa, with 74.79% on the default dataset

6 Run the code

To run the code in Colab, you just need to run all the cells in order.

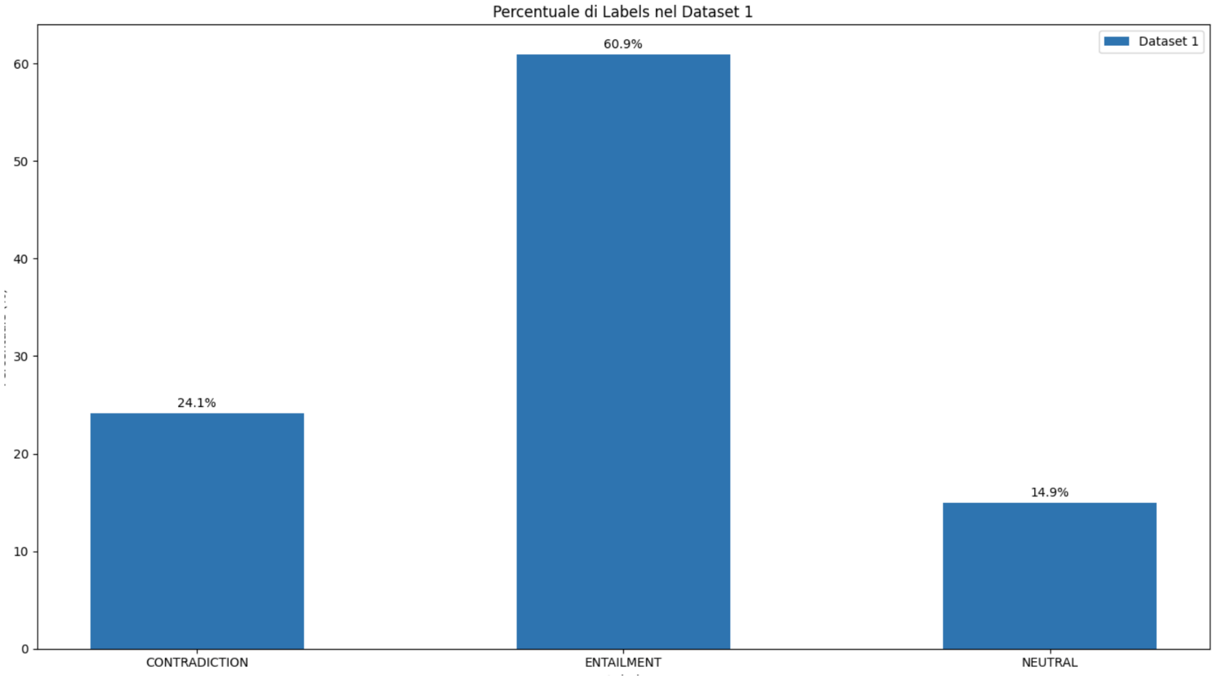
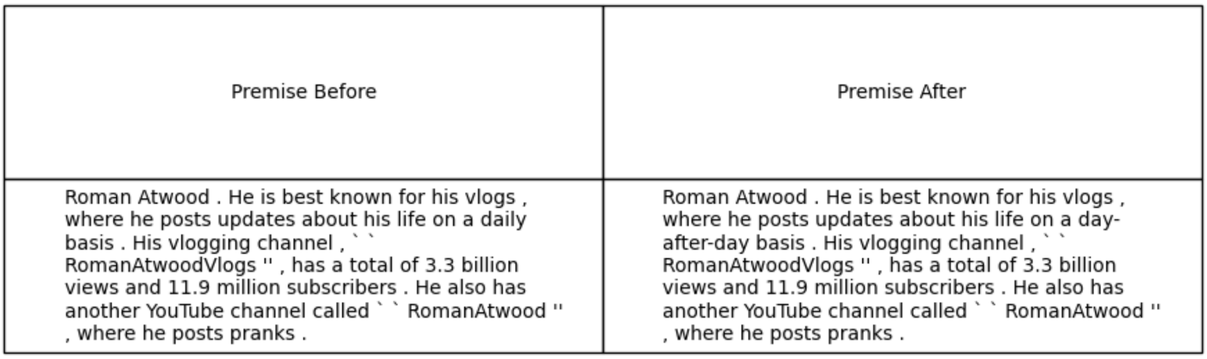
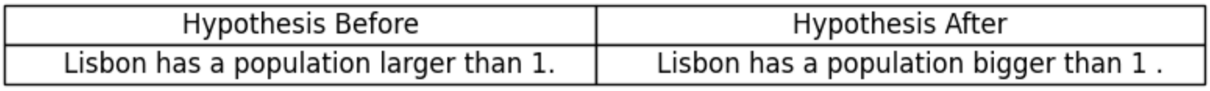


Figure 1: Percentage of labels in the default dataset



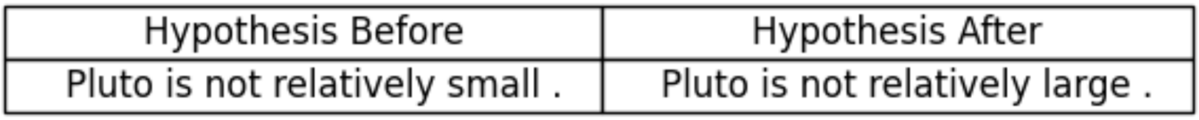
Premise Before	Premise After
Roman Atwood . He is best known for his vlogs , where he posts updates about his life on a daily basis . His vlogging channel , ` ` RomanAtwoodVlogs '' , has a total of 3.3 billion views and 11.9 million subscribers . He also has another YouTube channel called ` ` RomanAtwood '' , where he posts pranks .	Roman Atwood . He is best known for his vlogs , where he posts updates about his life on a day-after-day basis . His vlogging channel , ` ` RomanAtwoodVlogs '' , has a total of 3.3 billion views and 11.9 million subscribers . He also has another YouTube channel called ` ` RomanAtwood '' , where he posts pranks .

Figure 2: First WSD method



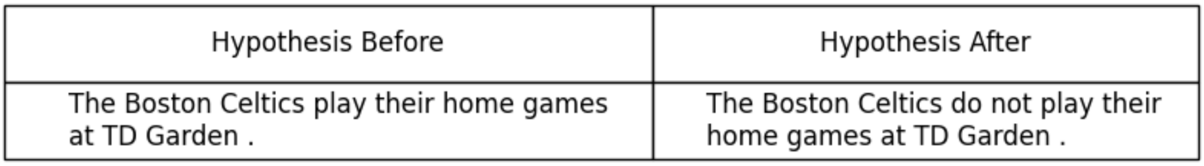
Hypothesis Before	Hypothesis After
Lisbon has a population larger than 1.	Lisbon has a population bigger than 1 .

Figure 3: Second WSD method



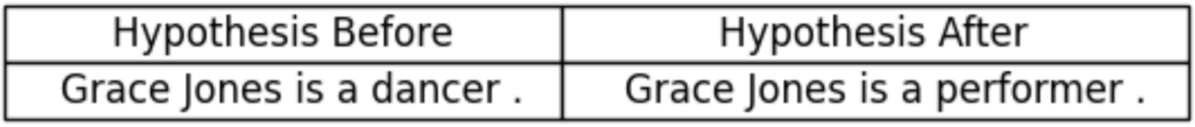
Hypothesis Before	Hypothesis After
Pluto is not relatively small .	Pluto is not relatively large .

Figure 4: Third WSD method



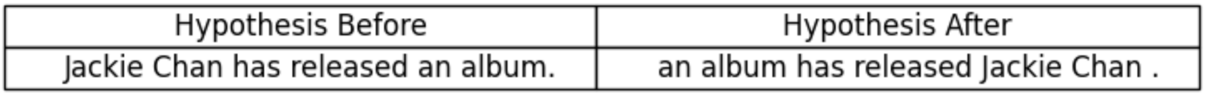
Hypothesis Before	Hypothesis After
The Boston Celtics play their home games at TD Garden .	The Boston Celtics do not play their home games at TD Garden .

Figure 5: Fourth WSD method



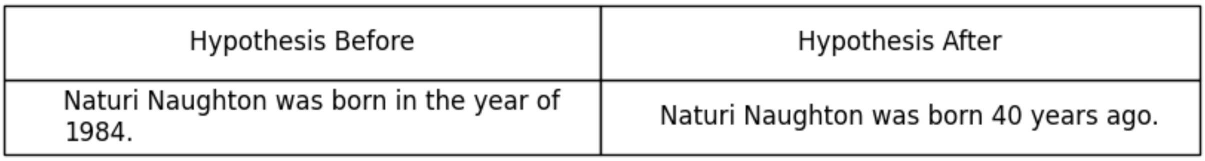
Hypothesis Before	Hypothesis After
Grace Jones is a dancer .	Grace Jones is a performer .

Figure 6: Fifth WSD method



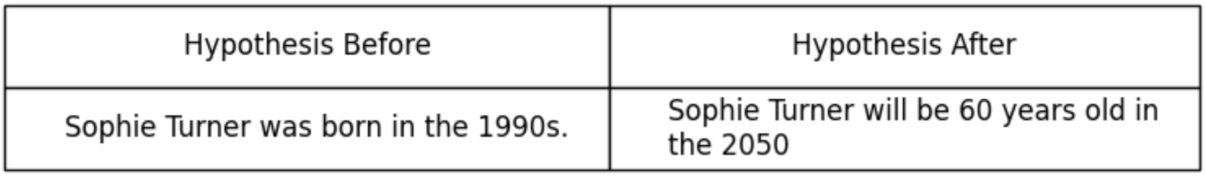
Hypothesis Before	Hypothesis After
Jackie Chan has released an album.	an album has released Jackie Chan .

Figure 7: First SRL method



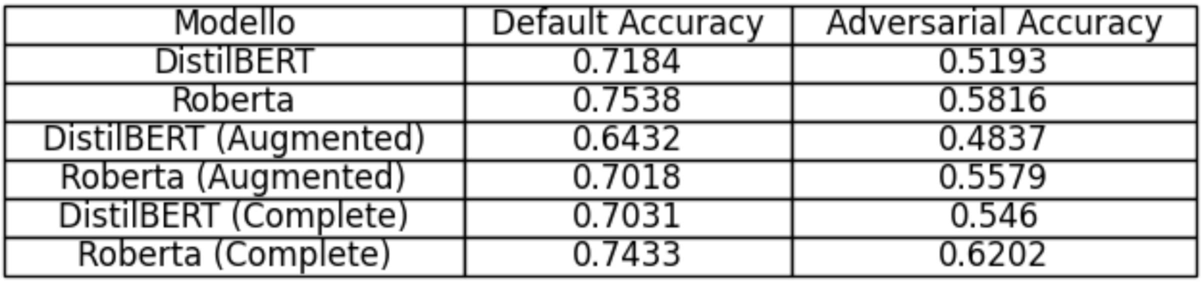
Hypothesis Before	Hypothesis After
Naturi Naughton was born in the year of 1984.	Naturi Naughton was born 40 years ago.

Figure 8: Second SRL method



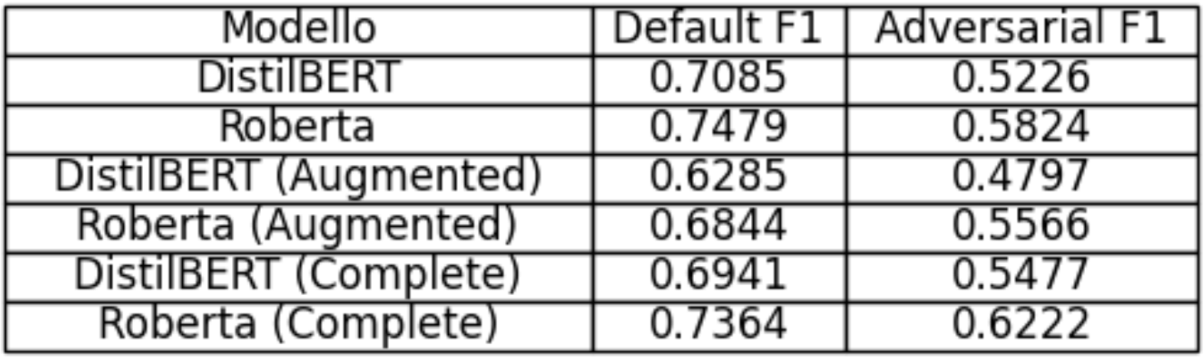
Hypothesis Before	Hypothesis After
Sophie Turner was born in the 1990s.	Sophie Turner will be 60 years old in the 2050

Figure 9: Third SRL method



Modello	Default Accuracy	Adversarial Accuracy
DistilBERT	0.7184	0.5193
Roberta	0.7538	0.5816
DistilBERT (Augmented)	0.6432	0.4837
Roberta (Augmented)	0.7018	0.5579
DistilBERT (Complete)	0.7031	0.546
Roberta (Complete)	0.7433	0.6202

Figure 10: Accuracy of the models



Modello	Default F1	Adversarial F1
DistilBERT	0.7085	0.5226
Roberta	0.7479	0.5824
DistilBERT (Augmented)	0.6285	0.4797
Roberta (Augmented)	0.6844	0.5566
DistilBERT (Complete)	0.6941	0.5477
Roberta (Complete)	0.7364	0.6222

Figure 11: F1 score of the models

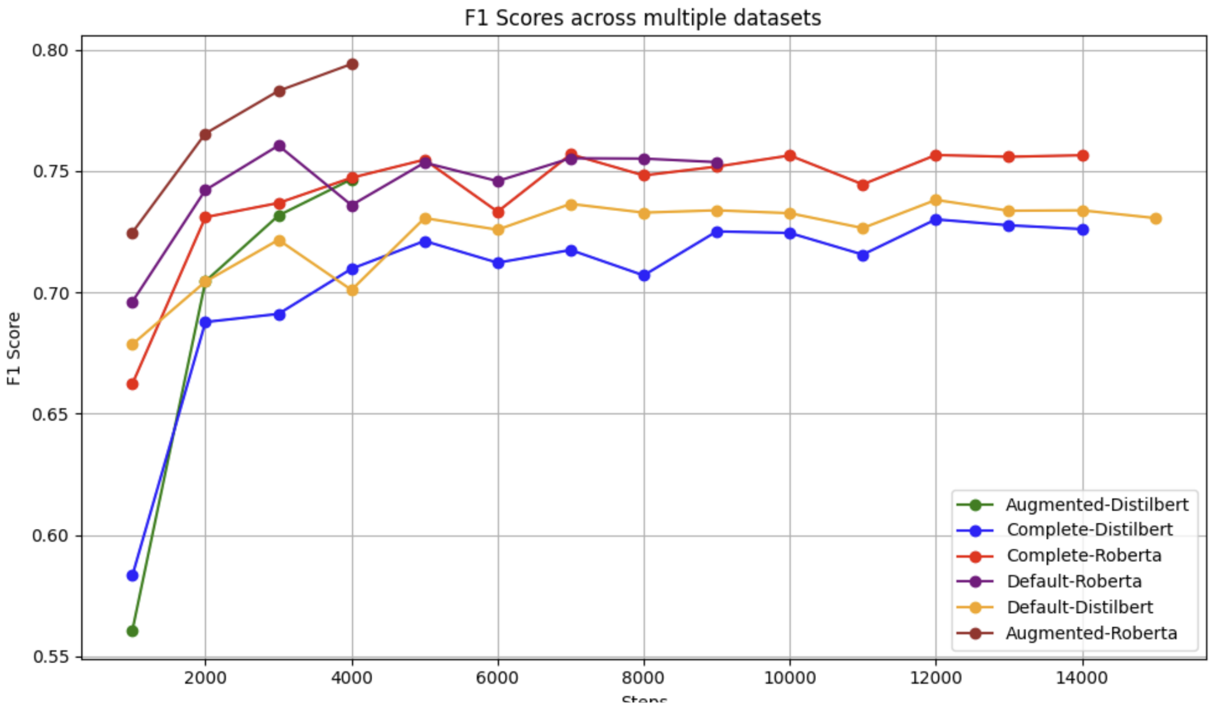


Figure 12: Performance during the training

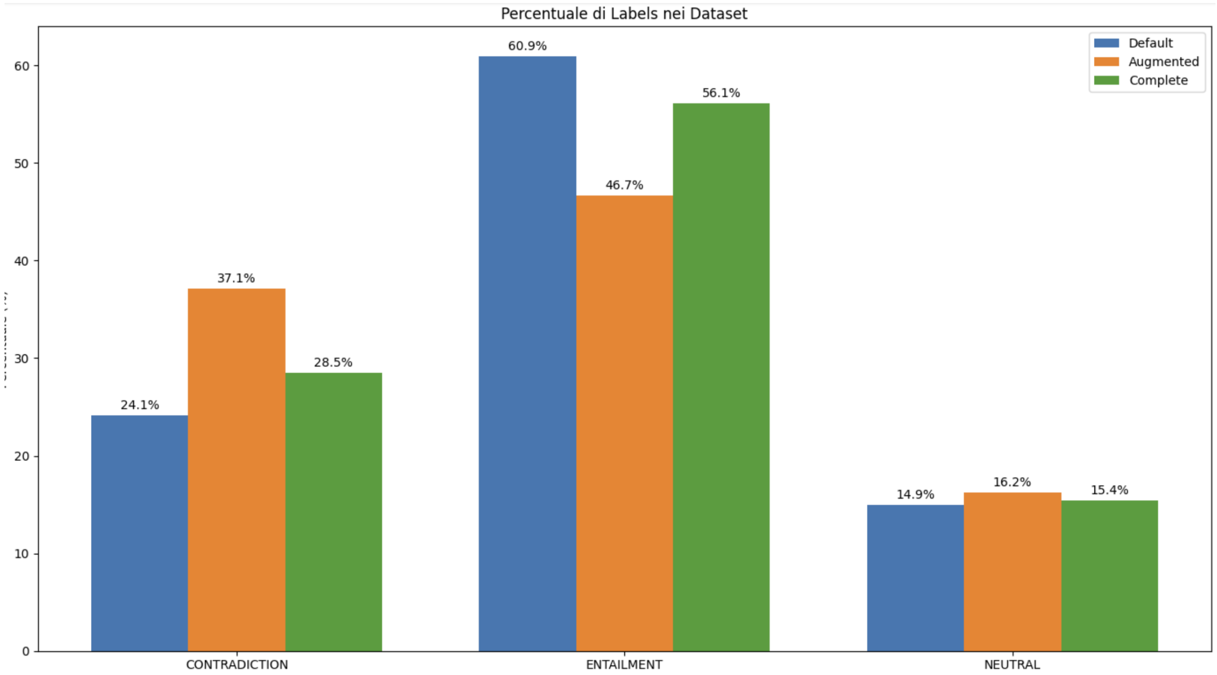


Figure 13: Percentage of labels in the Datasets