# A Novel Cartography-Based Curriculum Learning Method Applied on RoNLI: The First Romanian Natural Language Inference Corpus

**Eduard Poesina**
University of Bucharest
Bucharest, Romania
eduardgabriel.poe@gmail.com

**Cornelia Caragea**
University of Illinois Chicago
Chicago, IL, USA
cornelia@uic.edu

**Radu Tudor Ionescu***
University of Bucharest
Bucharest, Romania
raducu.ionescu@gmail.com

## Abstract

Natural language inference (NLI), the task of recognizing the entailment relationship in sentence pairs, is an actively studied topic serving as a proxy for natural language understanding. Despite the relevance of the task in building conversational agents and improving text classification, machine translation and other NLP tasks, to the best of our knowledge, there is no publicly available NLI corpus for the Romanian language. To this end, we introduce the first Romanian NLI corpus (RoNLI) comprising 58K training sentence pairs, which are obtained via distant supervision, and 6K validation and test sentence pairs, which are manually annotated with the correct labels. We conduct experiments with multiple machine learning methods based on distant learning, ranging from shallow models based on word embeddings to transformer-based neural networks, to establish a set of competitive baselines. Furthermore, we improve on the best model by employing a new curriculum learning strategy based on data cartography. Our dataset and code to reproduce the baselines are available at https://github.com/Eduard6421/RONLI.

## 1 Introduction

Given a sentence pair composed of a premise and a hypothesis, natural language inference (NLI) (Korman et al., 2018), a.k.a. textual entailment recognition, is the task of determining if the premise entails, contradicts, or is neutral to the hypothesis. NLI is an actively studied problem (Bigoulaeva et al., 2022; Chakrabarty et al., 2021; Chen et al., 2021; Luo et al., 2022; Mathur et al., 2022; Sadat and Caragea, 2022b,a; Snijders et al., 2023; Varshney et al., 2022; Wang et al., 2022; Wijnholds, 2023), being an essential task to be solved before addressing natural language understanding (NLU). Its complexity stems from the fact that NLU is generally considered an AI-hard problem (Yampolskiy, 2013). Notably, NLI forms the foundation of numerous advanced natural language processing systems (Korman et al., 2018), providing a backbone for multiple study areas such as language modeling (Merrill et al., 2022; Mitchell et al., 2022), conversational agents (Raghu et al., 2022), zero-shot text classification (Pàmies et al., 2023), image captioning (Shi et al., 2021), text summarization (Falke et al., 2019), discourse parsing (Sluyter-Gäthje et al., 2020), and many others (Korman et al., 2018). The importance of NLI is well recognized, being included as a reference task in benchmarks such as GLUE (Wang et al., 2019b) and SuperGLUE (Wang et al., 2019a).

To date, the NLI task has been intensively studied in the English language (Bowman et al., 2015; Koreeda and Manning, 2021; Romanov and Shivade, 2018; Sadat and Caragea, 2022b; Williams et al., 2018) and other languages, such as Chinese (Hu et al., 2020), Turkish (Budur et al., 2020), Portugese (Real et al., 2020) and Indonesian (Mahendra et al., 2021), as well as in multi-lingual scenarios (Conneau et al., 2018). However, the sparsity of resources led researchers to overlook the development of NLI models for low-resource languages, such as Romanian (Rotaru et al., 2024). In this context, we introduce a novel **Ro**manian **N**atural **L**anguage **I**nference corpus (RoNLI) composed of 64K sentence pairs. The samples are divided into 58K for training, 3K for validation, and 3K for testing. We collected the sentence pairs from the Romanian version of Wikipedia, while searching for certain linking phrases between contiguous sentences to automatically label the sentence pairs. The training pairs are obtained via an automatic rule-based annotation process known as *distant supervision* (Mintz et al., 2009), while the validation and test sentence pairs are manually annotated with the correct labels. To the best of our knowledge, RoNLI is the first publicly available corpus for Romanian natural lan-

* Corresponding author.



*Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 236–253

guage inference. RoNLI is meant to attract the attention of researchers in studying and developing NLI models for under-studied languages, where NLU systems often produce subpar results, due to data scarcity and cross-lingual settings. To this end, we believe that RoNLI is a valuable resource. We publish our data and source code under the open-source CC BY-NC-SA 4.0 license at: https://github.com/Eduard6421/RONLI.

We carry out experiments with multiple machine learning methods, ranging from shallow models based on word embeddings to transformer-based neural networks, to establish a set of competitive baselines. Moreover, we employ data cartography (Swayamdipta et al., 2020) to characterize our training samples as easy-to-learn (E2L), ambiguous (A), and hard-to-learn (H2L), from the perspective of the best baseline model. Further, we study several approaches that harness the resulting data characterization to mitigate the inherent labeling noise caused by the automatic labeling process. We manage to improve the top model via a novel curriculum learning method based on data cartography.

## 2 Related Work

To date, there are multiple NLI datasets for the English language that are derived from *different data sources* and capture *a different label space*. For example, the RTE dataset (Dagan et al., 2006), which was crucial for the emergence of the NLI task, consists of premises extracted from news articles, with each <*premise, hypothesis*> pair being labeled as either entailment or non-entailment. SICK (Marelli et al., 2014) contains sentence pairs derived from image captions and video descriptions classified into three classes: entailment, contradiction, or neutral. Similar to SICK, SNLI (Bowman et al., 2015) consists of premises extracted from image captions (with hypotheses being written by human annotators), covering the same label space as SICK. Despite being very large in size, one limitation of SNLI is that it lacks data source diversity. Williams et al. (2018) introduced MNLI to specifically increase the diversity of the data. That is, in MNLI, the premises are extracted from diverse sources such as face-to-face conversations, travel guides, and the 9/11 event, while the hypotheses are derived exactly as in SNLI. While both SNLI and MNLI are widely used to track progress in natural language understanding, they have been shown to contain spurious correlations (Gururangan et al., 2018). In an effort to reduce spurious correlations and create a harder dataset, Nie et al. (2020) developed ANLI, where human annotators wrote each hypothesis (conditioned on the premise and the label—entailment, contradiction, neutral) in an adversarial fashion, until the model failed to correctly predict the example.

| Category | Romanian | English Translation |
|---|---|---|
| Contrastive | În contrast | In contrast |
| | În contradicţie | In contradiction |
| | În opoziţie | In opposition |
| Entailment | Cu alte cuvinte | In other words |
| | În alţi termeni | In other terms |
| | Pe larg | In broader terms |
| Reasoning | Astfel | Thus |
| | Prin urmare | Therefore |
| | În concluzie | In conclusion |
| Neutral | N/A | N/A |

Table 1: Examples of original (Romanian) and translated (English) linking phrases and transition words that suggest certain relations between two sentences. Upon finding a premise and a hypothesis linked by one such phrase, we remove the respective phrase to force NLI models trained on the extracted sentences to focus on other clues.

There are many other relevant NLI datasets for English, which are briefly mentioned next. QNLI (Wang et al., 2019b) is derived from the SQuAD question answering dataset (Rajpurkar et al., 2016). SciTaiL (Khot et al., 2018) is derived from a school level science question-answer corpus with the sentence pairs being classified into two classes: entailment or not-entailment. MedNLI (Romanov and Shivade, 2018) is obtained from medical records of patients with the <*premise, hypothesis*> pairs being annotated by doctors as entailment, contradiction, or neutral. ContractNLI (Koreeda and Manning, 2021) is derived from document-level contracts and annotated with a set of hypotheses (or claims) per contract with each <*contract, hypothesis*> pair belonging to entailment, contradiction, or neutral. SciNLI (Sadat and Caragea, 2022b) is obtained from research articles published in the ACL Anthology, but unlike the above NLI datasets where the hypotheses are written by humans specifically with the NLI task in mind, in SciNLI, both sentences in a pair are extracted from research articles. These sentences are written by researchers without the NLI task in mind, and thus, are expected to contain fewer spurious correlations. The labels in SciNLI are entailment, contrasting, reasoning, and neutral (extending the usual set with the reasoning relation that often occurs in formal writing).

| Relation | #Samples | | | Average #Words | | | | | | Overlap Ratio | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Premise | | | Hypothesis | | | | | |
| | Train | Val | Test | Train | Val | Test | Train | Val | Test | Train | Val | Test |
| Contrastive | 2,592 | 74 | 74 | 26.7 | 25.7 | 24.6 | 25.4 | 27.1 | 23.5 | 0.07 | 0.07 | 0.08 |
| Entailment | 1,300 | 72 | 96 | 26.6 | 25.6 | 24.5 | 23.4 | 20.9 | 23.0 | 0.07 | 0.08 | 0.07 |
| Causal | 25,722 | 1,134 | 952 | 25.0 | 25.3 | 25.4 | 23.7 | 23.8 | 23.3 | 0.06 | 0.05 | 0.05 |
| Neutral | 28,500 | 1,778 | 1,878 | 23.8 | 23.5 | 23.8 | 23.8 | 23.9 | 24.3 | 0.03 | 0.03 | 0.04 |
| Overall | 58,114 | 3,059 | 3,000 | 24.5 | 24.4 | 24.3 | 23.7 | 23.9 | 23.8 | 0.04 | 0.04 | 0.04 |

Table 2: Statistics for each of the training, validation and test splits of the RoNLI dataset, including the class distribution (2nd to 4th columns), the average number of words in a premise (5th to 7th columns), the average number of words in a hypothesis (8th to 10th columns), and the average ratio of overlapping words between a premise and a hypothesis (11th to 13th columns).

Owing to the importance of the NLI task, several efforts have been directed to the development of datasets in other languages. For example, Conneau et al. (2018) introduced XNLI, a cross-lingual evaluation dataset obtained by translating examples from MNLI. Hu et al. (2020) created OCNLI, the first large Chinese dataset for NLI, whereas Kovatchev and Taulé (2022) explored NLI in the Spanish language. Other works focused on the creation of NLI datasets from low-resource languages such as Creole (Armstrong et al., 2022), Indonesian (Mahendra et al., 2021), and Turkish (Budur et al., 2020), either by human annotations or automatic translations from English. Similar to these latter works, we focus on a low-resource language, Romanian, to create a large NLI dataset, the first of its kind. In contrast to these studies and inspired by the creation of SciNLI (Sadat and Caragea, 2022b), instead of resorting to English translations, we use the linking phrases between contiguous sentences in text to automatically label a large training set for the Romanian language, while the validation and test sets are manually annotated.

## 3 Corpus

**Data gathering and automatic labeling.** The data source for building our corpus is the Romanian Wikipedia, chosen for its sizable, diverse and expansive content generated by a wide user base. Wikipedia presents the particularity of allowing users to define special pages: help pages, discussion pages, disambiguation pages, and pages dedicated to hosting meta-information about uploaded files. We consider that such pages are less relevant to the NLI task, and hence, we decided to filter them out. Next, we develop and apply a custom text parser to systematically deconstruct the corpus into individual articles, sub-articles, and chapters. The resulting text samples are further split using the sentence tokenizer available in NLTK (Bird et al., 2009). However, not all sections of Wikipedia articles are equally valid sources to extract sentence pairs for natural language inference. To ensure the quality and relevance of our dataset, we exclude segments such as the list of references, the image XML markers, the external links to other articles, and the metadata articles, which usually contain unstructured or less meaningful text.

Due to the collaborative nature of Wikipedia's content, there is a high variety in the quality of sentences. We observed that shorter sentences tend to introduce content that is often deemed unrepresentative for the NLI task, lowering the overall quality of our dataset. To extract high-quality sentences from Wikipedia, we thus set a minimum length of 50 characters for each sentence. We select this limit empirically, based on visually inspecting the sentences extracted with our parsing methodology.

Building upon the methodology presented by Sadat and Caragea (2022b), we divide the sentence pairs into four separate relationship types: contrastive, entailment, reasoning and neutral. To automatically label the sentence pairs using distant supervision, we establish a set of language-specific linking phrases and transition words that are very likely to predetermine the relationship between two sentences. Some examples of linking phrases for the related sentence pairs are shown in Table 1. Note that, for neutral sentence pairs, we require the absence of linking phrases. The full set of linking phrases for each relationship type (contrastive, entailment, reasoning) are listed in Appendix A. For instance, starting a sentence with the phrase "Îîn concluzie" (translated to "In conclusion") could be easily identified as a predictive cue, suggesting a reasoning relationship with the preceding sentence.

While recognizing that this approach is not infallible, it is undoubtedly an effective strategy for automatically collecting representative data for NLI in an unsupervised manner. The method allows us to harness the inherent relational information

embedded in the Romanian vocabulary without requiring specific knowledge about the subject matter. By employing this methodology, we successfully extract a total of 64K sentence pairs. In all the extracted examples, the linking phrases are removed to make sure that models do not learn superficial patterns (based on linking phrases) from the data.

To create our training, validation, and test sets, we employ stratified sampling to maintain the inherent distribution of the original data. In Table 2, we report several statistics for each of the three subsets, such as the class distribution, the average number of words in a premise, the average number of words in a hypothesis, and the average ratio of overlapping words between a premise and a hypothesis. Notably, we observe that the chosen data source, Wikipedia, leads to an uneven sample distribution, with significantly fewer sentences being attributed to the contrastive and entailment classes.

**Manual labeling.** To ensure that the validation and test labels are reliable, and models do not reach good performance levels due to biases in our automatic annotation process, we instruct three human annotators who are native Romanian speakers (with bachelor degrees) to manually relabel our validation and test sentences. To avoid biasing the annotators towards preferring the automatic labels, we preclude them from viewing the automatically generated labels (and the linking phrases). Our annotators received detailed guidelines and examples to ensure a consistent annotation process. The inter-rater Fleiss Kappa coefficient is 0.71, which denotes a substantial agreement between annotators. Notably, the agreement level among our annotators is larger than the Kappa agreement of 0.62 reported for SciNLI (Sadat and Caragea, 2022b), and the Kappa agreement of 0.70 reported for SNLI (Bowman et al., 2015). We originally allocated more data samples to the validation and test sets, but discarded validation and test samples that had no majority label provided by the annotators. The reported numbers of data samples for validation and test do not include the discarded ones. We compared the automatic annotations with the aggregated manual labels, obtaining a Cohen's Kappa coefficient between the automatic and manual labels of 0.62, which indicates a substantial agreement. Based on this high agreement, we have strong reasons to believe that the automatic labeling process is sufficiently accurate.

**Data cartography.** To further investigate the correctness and utility of the collected samples, we employ data cartography (Swayamdipta et al., 2020) on top of our best model, a version of the Romanian BERT (Ro-BERT) (Dumitrescu et al., 2020) fine-tuned on the RoNLI training data. Swayamdipta et al. (2020) proposed *data cartography* as a visualization technique based on interpreting the behavior of a machine learning model during the training stage. The method is built upon two insightful metrics recorded during the training process, specifically the level of *confidence* when accurately categorizing a data point into the correct class, and the *variability* (fluctuation) of the confidence during training. They are complemented by *correctness*, a metric that tells how often a sample is correctly classified during training.

As illustrated in Figure 1, data cartography provides a characterization of the training examples as easy-to-learn (E2L), ambiguous (A), and hard-to-learn (H2L), for the chosen Ro-BERT model. Easy-to-learn examples reside in the top-left quadrant of the data map, describing items which exhibit low variability and high confidence. The model's strong certainty in its decisions indicates that it faces minimal difficulty in comprehending these examples. In contrast, hard-to-learn examples, occupying the bottom-left quadrant, demonstrate low variability but differentiate themselves from the previous category in terms of confidence, which stays low in this case. The low confidence of the model in the ground-truth label suggests that the examples from this region are likely to be mislabeled. This is because most examples from the underrepresented classes (contrastive, entailment) are labeled as H2L (see Table 3). In contrast, ambiguous examples are characterized by high variability and mid-level confidence, and are thus positioned on the middle-right side of the map. Frequent changes in the predicted classes from one epoch to another prove that these examples are very challenging for the model.

For most training examples, our data map shows that Ro-BERT presents predominantly high confidence (over 30K training samples have the average confidence above 0.9) and high correctness (over 40K samples are correctly classified), suggesting that the model is particularly effective in recognizing and learning consistent patterns. A secondary cluster can be observed pertaining to the hard-to-learn scenarios reflecting that the model starts and remains strongly biased across training epochs for a relatively small portion of samples. In summary, the data map suggests that most samples are consistently labeled. Corroborating this observation with

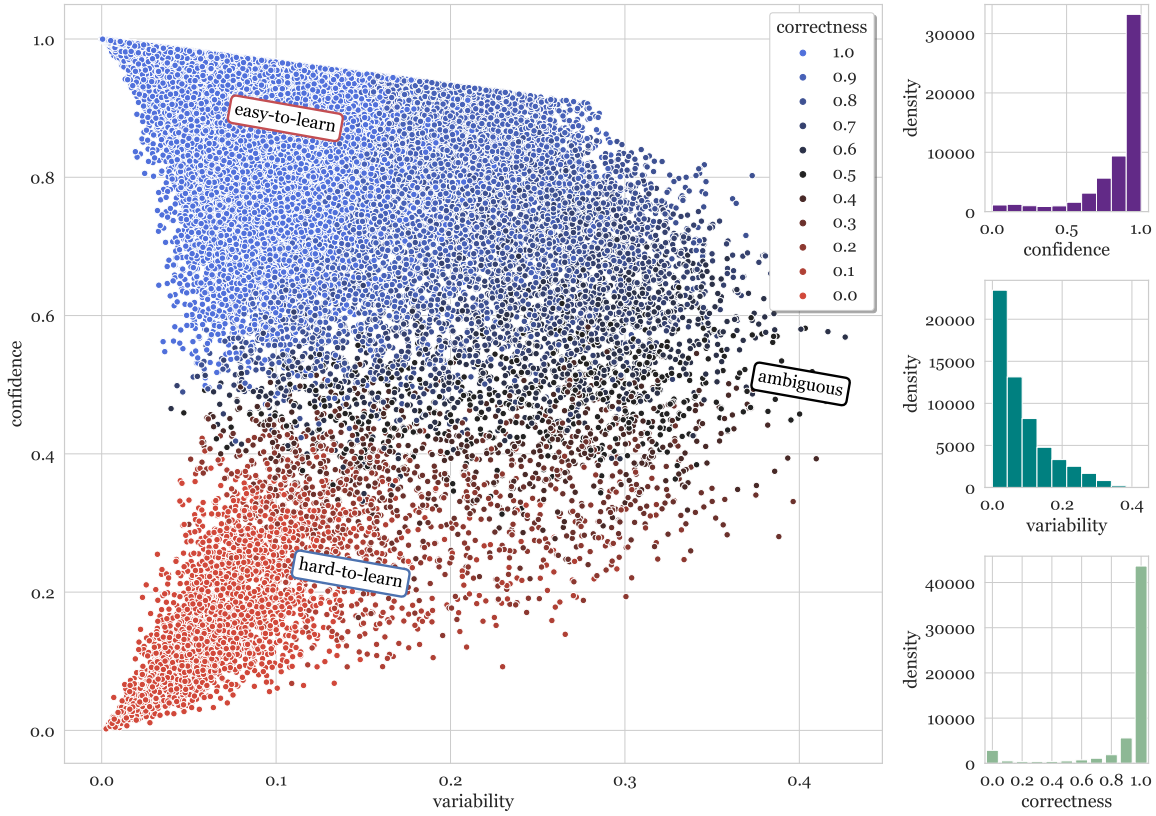


Figure 1: Data cartography visualization of the RoNLI dataset based on fine-tuning the Ro-BERT model (Dumitrescu et al., 2020). In the left-hand side plot, the $y$-axis corresponds to the level of confidence exhibited by the model during training, while the $x$-axis represents the variability of the confidence level. Adjacent to the primary plot, three histograms are displayed on the right-hand side, each representing a different metric: the confidence, the variability of confidence, and the correctness. The visualization offers a comprehensive overview of our dataset characteristics and the behavior of Ro-BERT during training. Best viewed in color.

the substantial Cohen's Kappa agreement between the automatic and manual labels suggests that the automatic labels are mostly correct. Nonetheless, we further investigate the implications of the observed data distribution in Section 5 by considering training scenarios where we deliberately choose the type of data to incorporate, encompassing varying combinations of E2L, A, and H2L instances.

## 4 Models

We employ grid-search to tune the hyperparameters of all models (see details in Appendix A).

**Classifiers based on word embeddings.** Our first two baselines are based on word embeddings, leveraging the 300-dimensional word vectors returned by the fastText framework (Bojanowski et al., 2017), which includes support for the Romanian language. We select fastText primarily due to its inherent capability to manage out-of-vocabulary words, a crucial aspect considering the diverse nature of Wikipedia content. We represent each sentence as a continuous bag-of-words (CBOW), i.e., we compute the average of the word embeddings in each sentence. Next, the CBOW representations generated for a sentence pair are concatenated, resulting in a joint representation of the premise and the hypothesis. We consider two alternative classifiers to learn from the extracted CBOW embeddings, namely a linear Support Vector Machines (SVM) model and a Multinomial Logistic Regression (Softmax) model.

**RoGPT2.** The next baselines draw upon the capabilities of transformer models (Vaswani et al., 2017), which have shown impressive performance on various NLP tasks, including NLI. Our third baseline is based on the generative capabilities of the GPT family of models. By harnessing the inherent generative knowledge embedded within GPT, we can establish meaningful correlations between two sentences, a premise and a hypothesis, and accurately infer their relationship. For this approach, we employ a GPT2 model pre-trained on Romanian corpora (Niculescu et al., 2021). We extract the last hidden states of the end-of-sentence (EOS)

token and feed it to a softmax classification head.

**ChatGPT 3.5 Turbo.** Another way to utilize transformer-based large language models (LLMs) for various tasks is via in-context (prompt-based) learning. This approach leverages the meta-learning capabilities of LLMs. To this end, we employ a powerful multilingual LLM, namely ChatGPT 3.5 Turbo (Ouyang et al., 2022), with in-context learning (two examples per class), to directly predict the test labels. This model does not require any fine-tuning.

**Multilingual BERT.** We employ a multilingual BERT model which is pre-trained on the XNLI corpus (Conneau et al., 2018). We then consider two options, one based on zero-shot learning, and one based on fine-tuning. The zero-shot multilingual BERT follows the Definition-Restrictive approach described by Yin et al. (2019). This is because our set of classes (defined in Appendix A) do not coincide exactly with those of XNLI.

**Ro-BERT.** Our sixth baseline is a variant of BERT (Devlin et al., 2019) pre-trained on Romanian corpora, namely Ro-BERT (Dumitrescu et al., 2020). We hypothesize that Ro-BERT discerns inherent characteristics within and between sentences, allowing it to accurately predict the relationship type. We extract the [CLS] token from the final hidden state of each sentence in order to obtain an internal representation. The extracted embeddings are then concatenated and fed into a softmax classification head, enabling the system to predict the relationship between pairs.

**Ro-BERT (spurious).** Natural language inference datasets commonly face the challenge of spurious correlations (Gururangan et al., 2018), which are statistical associations that machine learning models learn to interpret and rely upon when making predictions. Such correlations do not necessarily reflect a robust understanding of NLI, but rather illustrate common language patterns and expressions. In order to assess how much RoNLI is affected by this phenomenon, we fine-tune Ro-BERT only on the hypotheses, with the same hyperparameters as above. The performance of this model is proportional to the amount of spurious correlations.

**Ro-BERT + cartography subsampling.** We explore additional baselines based on the categorization produced by data cartography. Following the work of Swayamdipta et al. (2020), we fine-tune the Ro-BERT model on subsets of the training set, determined by the difficulty categories: easy-to-learn, ambiguous and hard-to-learn. We train one Ro-BERT model for each of the three categories of samples. In addition, we train a Ro-BERT model on easy-to-learn and ambiguous (E2L+A) samples, based on the intuition that hard-to-learn samples suffer from high rates of labeling noise, and should thus be discarded.

**Ro-BERT + Length-CL.** We employ a Ro-BERT model based on curriculum learning, a training paradigm proposed by Bengio et al. (2009). The technique emulates the human learning process, which typically progresses from simpler to more complex concepts, and applies it to the training of machine learning models. The idea behind the technique is to expose the model to easier examples first, thus building a foundational understanding of the task, which will in turn allow it to accommodate more complex samples faster. Curriculum learning has received significant attention from researchers due to its vast applicability, as shown by Soviany et al. (2022). Following Nagatsuka et al. (2023), the considered baseline applies the curriculum with respect to text length.

For a fair comparison with the conventional training paradigm, we ensure that the total number of training iterations is equal with those of the base model. We perform the curriculum training for half the number of total iterations. Subsequently, we continue training under the standard paradigm, using all available data for the remaining iterations. As for the baseline Ro-BERT, we introduce early stopping to prevent overfitting. We apply the same training procedure to all curriculum models described below.

**Ro-BERT + STS-CL.** An intuitive way to introduce curriculum learning with respect to the NLI task is to use the semantic text similarity (STS) between sentences to measure difficulty. We thus add a baseline Ro-BERT model that uses curriculum learning based on STS. To calculate STS, we utilize the cosine similarity based on the fine-tuned Ro-BERT embeddings of the sentences. The training starts with the most similar sentence pairs, and progressively adds less similar sentence pairs.

**Ro-BERT + Cart-CL.** We propose a Ro-BERT model based on curriculum learning with respect to the difficulty groups established via data cartography. We start the training on the E2L samples for a number of iterations. We gradually incorporate more challenging batches, adding ambiguous data instances, then H2L samples. If the total number of iterations is $N$, we perform $N/4$ iterations with E2L samples, $N/4$ iterations with A samples, and

| | Contrastive | Entailment | Causal | Neutral |
|---|---|---|---|---|
| E2L | 0 | 0 | 2,207 | 17,082 |
| A | 967 | 375 | 9,614 | 8,334 |
| H2L | 2,592 | 1,300 | 11,632 | 3,765 |

Table 3: Distribution of classes in each difficulty group defined via data cartography.

$N/2$ iterations with H2L samples.

**Ro-BERT + Cart-CL++.** Our next model confronts certain caveats introduced by data cartography (Swayamdipta et al., 2020) in terms of data difficulty characterization. Given that the grouping of data instances relies on one metric (either confidence or variability) instead of both, we observed that the E2L, A, and H2L groups defined by Swayamdipta et al. (2020) do not form a partition of the training set, leaving samples outside these groups, while assigning other samples to two different groups. Leaving data outside or duplicating some of the examples can bias the model and lead to suboptimal performance. We conjecture that a better curriculum can be created by designing a difficulty scoring function that jointly takes confidence and variability into account. Let $c_i$ and $v_i$ be the confidence and variability of the $i$-th training sample. We introduce a new difficulty scoring function $s : [0, 1] \times [0, 1] \rightarrow [0, 2]$ defined as follows:

$$s(c_i, v_i) = \begin{cases} 1 - c_i + v_i, \text{if } c_i < 0.5 \\ 2 - c_i - v_i, \text{otherwise} \end{cases}, \quad (1)$$

where $i \in \{1, ..., n\}$, and $n$ is the number of training samples. By design, our novel scoring function $s$ assigns low scores to items characterized by low variability and high correctness, medium scores for items perceived as ambiguous, and high scores to difficult examples.

**Ro-BERT + Cart-Stra-CL++.** By analyzing the distribution of classes in each data cartography group (see Table 3), we observe that the E2L group does not contain any contrastive or entailment instances. Hence, we propose yet another model based on curriculum learning, which constructs stratified easy-to-hard batches. This ensures the diversity of class labels right from the beginning of the training process, avoiding to bias the model towards certain classes. This novel curriculum learning approach is more suitable for imbalanced datasets, which are more prone to be affected by introducing further class biases.

## 5 Results and Observations

**Experimental setup.** We perform NLI experiments with the set of baselines described above. We also carry out experiments with the data cartography groups, considering scenarios when our top performing model is trained with subsets of the training set, namely using only E2L examples, A examples, H2L examples, and E2L+A examples. We add the models based on curriculum learning to our line-up of models, specifically Length-CL, STS-CL, Cart-CL, Cart-CL++, and Cart-Stra-CL++. As evaluation metrics, we report the precision, recall and $F_1$ score for each of the four classes, on the manually labeled test set. We also report the micro and macro $F_1$ scores to determine how models are affected by the imbalanced nature of RoNLI. Note that the micro $F_1$ is equivalent to the classification accuracy, while the macro $F_1$ is also known as the group-averaged $F_1$.

Our experimental methodology is based on conducting multiple trials of each experiment, specifically five runs, to ensure the robustness and reliability of our findings. The model that demonstrates the best performance on the validation set across the five trials is the one selected for evaluation on the test set. This procedure is fairly performed for all models, including the baselines.

**Results.** The results obtained by the baselines on the individual classes are presented in Table 4. The zero-shot and fine-tuned versions of multilingual BERT obtain subpar results, suggesting that cross-lingual systems are not suited for Romanian NLI. Among the various language models, Ro-BERT achieves the highest performance level. Still, we observe that models based on word embeddings are more able to learn the underrepresented classes than the vanilla transformers (despite the strong performance of transformer models on NLI for the English language). To overcome this limitation, we retrain the most promising transformer (Ro-BERT) by oversampling the underrepresented classes such that the class distribution becomes balanced. Since the results indicate that oversampling brings significant performance gains, we integrate oversampling in the training process of all the subsequent versions of Ro-BERT.

The overall results in terms of the micro and macro $F_1$ scores obtained by the baselines are shown in Table 5. We observe that no baseline is capable of surpassing a micro-averaged $F_1$ score of 80% and a macro-averaged $F_1$ score of 60%. This observation highlights the complexity of the RoNLI dataset, pointing to a new challenging benchmark for natural language inference.

The various training scenarios applied on Ro-

| Method | Over sampling | Contrastive | | | Entailment | | | Reasoning | | | Neutral | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | P | R | $F_1$ | P | R | $F_1$ | P | R | $F_1$ | P | R | $F_1$ |
| CBOW + SVM | ✗ | 0.16 | 0.32 | 0.22 | 0.19 | 0.51 | 0.28 | 0.51 | 0.67 | 0.58 | 0.88 | 0.63 | 0.74 |
| CBOW + Softmax | ✗ | **0.50** | 0.01 | 0.03 | **0.57** | 0.04 | 0.08 | 0.49 | 0.81 | 0.61 | 0.85 | 0.65 | 0.74 |
| RoGPT2 | ✗ | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.52 | 0.83 | 0.63 | 0.87 | 0.69 | 0.77 |
| ChatGPT 3.5 Turbo | - | 0.13 | 0.20 | 0.16 | 0.14 | 0.55 | 0.22 | 0.45 | 0.72 | 0.56 | 0.96 | 0.51 | 0.66 |
| Zero-shot multilingual BERT | ✗ | 0.02 | 0.77 | 0.05 | 0.03 | 0.23 | 0.05 | 0.37 | 0.15 | 0.21 | 0.49 | 0.01 | 0.02 |
| Fine-tuned multilingual BERT | ✗ | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.38 | 0.89 | 0.53 | 0.85 | 0.35 | 0.49 |
| Ro-BERT | ✗ | 0.04 | 0.01 | 0.01 | 0.00 | 0.00 | 0.00 | 0.55 | 0.86 | 0.67 | 0.89 | 0.72 | 0.79 |
| Ro-BERT (spurious) | ✗ | 0.07 | 0.01 | 0.03 | 0.00 | 0.00 | 0.00 | 0.53 | 0.69 | 0.60 | 0.80 | 0.75 | 0.77 |
| Ro-BERT | ✓ | 0.19 | 0.80 | 0.31 | 0.35 | 0.54 | 0.41 | 0.62 | 0.72 | 0.67 | 0.96 | 0.73 | 0.83 |
| Ro-BERT + E2L | ✓ | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.52 | 0.49 | 0.50 | 0.71 | **0.79** | 0.74 |
| Ro-BERT + A | ✓ | 0.17 | 0.07 | 0.10 | 0.31 | 0.06 | 0.10 | 0.55 | 0.95 | **0.70** | **0.97** | 0.68 | 0.80 |
| Ro-BERT + H2L | ✓ | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.31 | **0.99** | 0.48 | 0.46 | 0.00 | 0.00 |
| Ro-BERT + E2L + A | ✓ | 0.21 | 0.10 | 0.13 | 0.35 | 0.07 | 0.11 | 0.55 | 0.93 | 0.69 | 0.96 | 0.68 | 0.80 |
| Ro-BERT + Length-CL | ✓ | 0.17 | 0.72 | 0.28 | 0.34 | 0.43 | 0.36 | 0.62 | 0.71 | 0.66 | 0.95 | 0.75 | **0.84** |
| Ro-BERT + STS-CL | ✓ | 0.20 | **0.86** | 0.32 | 0.37 | 0.40 | 0.37 | 0.63 | 0.76 | 0.69 | 0.96 | 0.74 | **0.84** |
| Ro-BERT + Cart-CL | ✓ | 0.13 | 0.63 | 0.21 | 0.22 | **0.62** | 0.32 | **0.64** | 0.63 | 0.63 | 0.96 | 0.74 | **0.84** |
| Ro-BERT + Cart-CL++ | ✓ | 0.19 | 0.79 | 0.31 | 0.36 | 0.58 | **0.44** | 0.63 | 0.74 | 0.68 | 0.96 | 0.74 | **0.84** |
| Ro-BERT + Cart-Stra-CL++ | ✓ | 0.26 | 0.77 | **0.38** | 0.45 | 0.44 | **0.44** | 0.62 | 0.79 | **0.70** | 0.96 | 0.74 | **0.84** |

Table 4: Per class precision, recall and $F_1$ scores of the proposed baseline models on the manually labeled RoNLI test set. The scores are independently reported for each class to enable a detailed class-based assessment of the classification performance. The best results are shown in bold blue and the second best are underlined and in blue.

| Method | Over sampling | $F_1$ | |
|---|---|---|---|
| | | Micro | Macro |
| CBOW + SVM | ✗ | 0.63 | 0.45 |
| CBOW + Softmax | ✗ | 0.66 | 0.36 |
| RoGPT2 | ✗ | 0.70 | 0.30 |
| ChatGPT 3.5 Turbo | - | 0.57 | 0.40 |
| Zero-shot multilingual BERT | ✗ | 0.08 | 0.08 |
| Fine-tuned multilingual BERT | ✗ | 0.50 | 0.30 |
| Ro-BERT | ✗ | 0.72 | 0.37 |
| Ro-BERT (spurious) | ✗ | 0.69 | 0.35 |
| Ro-BERT | ✓ | 0.73 | 0.56 |
| Ro-BERT + E2L | ✓ | 0.65 | 0.31 |
| Ro-BERT + A | ✓ | 0.73 | 0.44 |
| Ro-BERT + H2L | ✓ | 0.31 | 0.12 |
| Ro-BERT + E2L + A | ✓ | 0.73 | 0.44 |
| Ro-BERT + Length-CL | ✓ | 0.73 | 0.54 |
| Ro-BERT + STS-CL | ✓ | 0.74 | 0.57 |
| Ro-BERT + Cart-CL | ✓ | 0.70 | 0.51 |
| Ro-BERT + Cart-CL++ | ✓ | 0.73 | 0.57 |
| Ro-BERT + Cart-Stra-CL++ | ✓ | **0.75** | **0.59** |

Table 5: Overall micro and macro $F_1$ scores of the proposed baseline models on the manually labeled RoNLI test set. The micro and macro $F_1$ scores are both reported to acknowledge the behavior of models on our imbalanced dataset. The best results are shown in bold blue and the second best are underlined and in blue.

BERT lead to some interesting observations. When Ro-BERT is trained on hypotheses, its performance drops by 3% in terms of micro $F_1$, and 2% in terms of macro $F_1$, indicating that spurious correlations are not enough to reach high performance levels.

The Ro-BERT trained on E2L examples performs worse than the vanilla Ro-BERT, suggesting that the former model does not benefit much from the homogeneity of the easy examples. The Ro-BERT trained on H2L examples reaches an even lower performance, suggesting that focusing too much on complex examples can hurt overall performance, possibly due to overfitting on these hard cases (and their potentially erroneous labels). In contrast, the Ro-BERT trained on ambiguous examples and the one trained on easy and ambiguous examples (E2L + A) are fairly close to vanilla Ro-BERT, indicating that the ambiguous samples represent the most useful group to train the model on.

The state-of-the-art curriculum method, Length-CL (Nagatsuka et al., 2023), as well as our first curriculum learning approach, Cart-CL, exhibit lower performance than the vanilla Ro-BERT. Cart-CL++ outperforms Cart-CL, confirming that our difficulty scoring function $s$ is useful. Furthermore, our function boosts the macro $F_1$ score of Ro-BERT + Cart-CL++ over the vanilla Ro-BERT. However, it is likely that Cart-CL++ is still suboptimal because, in the early stages of the curriculum, when using only the E2L group, there is a complete lack of contrastive and entailment pairs (see Table 3). This issue should be fixed by the Cart-Stra-CL++ strategy, which makes sure to add examples from each class, in each difficulty batch. Cart-Stra-CL++ surpasses the Length-CL (Nagatsuka et al., 2023), STS-CL, Cart-CL and Cart-CL++ strategies, as well as the vanilla Ro-BERT, confirming our intuition. Still, the *overall performance* (Table 5) on RoNLI does not exceed 80% in terms of the micro $F_1$ (accuracy), which highlights both the need for the development of more robust and accurate models, and the research opportunity offered by our RoNLI

| Method | Micro $F_1$ | Macro $F_1$ |
|---|---|---|
| BERT | 0.748 | 0.747 |
| BERT + Length-CL | 0.709 | 0.705 |
| BERT + Cart-Stra-CL++ | **0.756** | **0.755** |

Table 6: Overall micro and macro $F_1$ scores of BERT, BERT+Length-CL and BERT+Cart-Stra-CL++ on the SciNLI (Sadat and Caragea, 2022b) test set. The best results are shown in bold blue.

dataset.

**Statistical testing.** We performed a Cochran's Q statistical test to compare the Ro-BERT based on oversampling (micro $F_1 = 0.73$, macro $F_1 = 0.56$) with the proposed Ro-BERT + Cart-Stra-CL++ (micro $F_1 = 0.75$, macro $F_1 = 0.59$). The test indicates that the proposed model is significantly better than the Ro-BERT based on oversampling, with a p-value of $0.001$. Note that the Cochran's Q test is applied on the contingency tables. Hence, the test rather indicates that the predictions are significantly different, without particularly using the reported micro / macro $F_1$ scores. We also performed the Mann-Whitney U test between Ro-BERT with oversampling and Ro-BERT+Cart-Stra-CL++. The Mann-Whitney U test further confirms that our method is significantly better than the baseline, with a p-value lower than $0.001$.

**Generalization results.** To assess the generalization capacity of our novel learning method (Cart-Stra-CL++), we extend the evaluation to an additional dataset, namely SciNLI (Sadat and Caragea, 2022b). We compare three models on SciNLI: BERT (Devlin et al., 2019), BERT+Length-CL (Nagatsuka et al., 2023) and BERT+Cart-Stra-CL++ (ours). All BERT models use the English base cased version. Note that SciNLI is balanced, so oversampling is not needed for any of the methods. The results on SciNLI are shown in Table 6. Our curriculum learning method achieves the best performance among the three methods, showing that Cart-Stra-CL++ generalizes to other datasets.

## 6 Discussion

A unique characteristic of RoNLI, by contrasting it with other existing NLI datasets, is that the hypotheses in RoNLI are sentences from Wikipedia which are not written by crowd-workers with the NLI task in mind, and these hypotheses are more diverse. In contrast, hypotheses in SNLI (Bowman et al., 2015) and MNLI (Williams et al., 2018) are written by paid crowd-workers and lack diversity. For example, in SNLI, we observed many pairs such as those presented in Table 10 from Appendix A, which are less challenging for a model and are predicted as *contradiction* by models without even looking at the first sentence (the premise). This issue was also observed by Gururangan et al. (2018).

Furthermore, we emphasize that Romanian is a unique Eastern Romance language, being part of a linguistic group that evolved from several dialects of Vulgar Latin which separated from the Western Romance languages between the 5th and the 8th centuries AD. Being surrounded by Slavic neighbors, it has a strong Slavic influence, which makes Romanian a unique language (among the Latin languages). Therefore, the results obtained by multilingual BERT in the zero-shot and fine-tuned settings (presented in Tables 4 and 5) suggest that there are linguistic differences between our dataset for NLI in Romanian and other NLI-supported languages. This confirms the need for a language-specific dataset such as RoNLI, despite the presence of other Latin languages, such as French and Spanish, and even Slavic languages, such as Bulgarian and Russian, in XNLI.

## 7 Conclusion

In this work, we introduced RoNLI, the first public corpus for Romanian natural language inference. We trained and evaluated several models to establish a set of competitive baselines for future research on our corpus. We also performed experiments to analyze the effect of spurious correlations, as well as the effect of harnessing data cartography to establish groups of useful samples or to develop curriculum learning strategies. Notably, we introduced a novel curriculum learning strategy based on data cartography and stratified sampling, boosting the overall micro and macro $F_1$ scores of Ro-BERT by $2\%$ and $3\%$, respectively. These improvements are statistically significant. Nevertheless, our empirical results indicate that there is sufficient room for future research on the Romanian NLI task. We make our dataset and code available to the community to lay the grounds for future research on Romanian NLI.

## 8 Acknowledgments

We thank reviewers for their constructive feedback, leading to significant improvements of our work.

## 9 Limitations

Our work is slightly limited by the number of samples with manual labels included in our corpus. This limitation is caused by the laborious annotation process involving human effort. However, we underline that RoNLI is not the only NLI corpus with training data based on automatic labels. For example, SciNLI (Sadat and Caragea, 2022b) is created using a very similar process, based on automatically labeling English sentence pairs via linking phrases.

## 10 Ethics Statement

The manual labeling was carried out by volunteering students, who agreed to annotate the news articles in exchange for bonus points. Prior to the annotation, they also agreed to let us publish the labels along with the dataset. We would like to emphasize that the students understood that the annotation task is optional, and they could also get the extra bonus points from alternative tasks. Moreover, all students had the opportunity to get a full grade without the optional tasks. Hence, there was no obligation for any of the students to perform the annotations.

Our data is collected from Wikipedia, which resides in the public web domain. We note that the European regulations[1] allow researchers to use data in the public web domain for non-commercial research purposes. Thus, we release our data and code under the CC BY-NC-SA 4.0 license[2].

We acknowledge that some sentences could refer to certain public figures. To ensure the right to be forgotten, we will remove all references to a person, upon receiving removal requests via an email to any of the authors.

## References


Ruth-Ann Armstrong, John Hewitt, and Christopher Manning. 2022. JamPatoisNLI: A Jamaican Patois Natural Language Inference Dataset. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5307–5320, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 41–48, New York, NY, USA. ACM.

Irina Bigoulaeva, Rachneet Singh Sachdeva, Harish Tayyar Madabushi, Aline Villavicencio, and Iryna Gurevych. 2022. Effective Cross-Task Transfer Learning for Explainable Natural Language Inference with T5. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 54–60, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly, Beijing.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Emrah Budur, Rıza Özçelik, Tunga Gungor, and Christopher Potts. 2020. Data and Representation for Turkish Natural Language Inference. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8253–8267, Online. Association for Computational Linguistics.

Tuhin Chakrabarty, Debanjan Ghosh, Adam Poliak, and Smaranda Muresan. 2021. Figurative language in recognizing textual entailment. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3354–3361, Online. Association for Computational Linguistics.

Zeming Chen, Qiyue Gao, and Lawrence S. Moss. 2021. NeuralLog: Natural Language Inference with Joint Neural and Logical Reasoning. In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 78–88, Online. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating Cross-lingual Sentence Representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL Recognising Textual Entailment


[1] https://eur-lex.europa.eu/eli/dir/2019/790/oj

[2] https://creativecommons.org/licenses/by-nc-sa/4.0/

Challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, pages 177–190.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 4171–4186.

Carmen Dobrovie-Sorin. 1994. *The Syntax of Romanian. Comparative Studies in Romance*. De Gruyter Mouton, Berlin, Boston.

Ştefan Daniel Dumitrescu, Andrei-Marius Avram, and Sampo Pyysalo. 2020. The birth of Romanian BERT. In *Findings of the Association for Computational Linguistics (EMNLP)*, pages 4324–4328.

Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. Ranking Generated Summaries by Correctness: An Interesting but Challenging Application for Natural Language Inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2214–2220.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 107–112.

Hai Hu, Kyle Richardson, Liang Xu, Lu Li, Sandra Kübler, and Lawrence Moss. 2020. OCNLI: Original Chinese Natural Language Inference. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3512–3526, Online. Association for Computational Linguistics.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. SciTaiL: A Textual Entailment Dataset from Science Question Answering. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI)*, volume 17, pages 41–42.

Yuta Koreeda and Christopher Manning. 2021. ContractNLI: A Dataset for Document-level Natural Language Inference for Contracts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1907–1919, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Daniel Z. Korman, Eric Mack, Jacob Jett, and Allen H. Renear. 2018. Defining textual entailment. *Journal of the Association for Information Science and Technology*, 69(6):763–772.

Venelin Kovatchev and Mariona Taulé. 2022. InferES: A natural language inference corpus for Spanish featuring negation-based contrastive and adversarial examples. In *Proceedings of the 29th International Conference on Computational Linguistics (COLING)*, pages 3873–3884.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Cheng Luo, Wei Liu, Jieyu Lin, Jiajie Zou, Ming Xiang, and Nai Ding. 2022. Simple but Challenging: Natural Language Inference Models Fail on Simple Sentences. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3449–3462, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Rahmad Mahendra, Alham Fikri Aji, Samuel Louvan, Fahrurrozi Rahman, and Clara Vania. 2021. IndoNLI: A Natural Language Inference Dataset for Indonesian. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 10511–10527, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*, pages 216–223.

Puneet Mathur, Gautam Kunapuli, Riyaz Bhat, Manish Shrivastava, Dinesh Manocha, and Maneesh Singh. 2022. DocInfer: Document-level Natural Language Inference using Optimal Evidence Selection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 809–824, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

William Merrill, Alex Warstadt, and Tal Linzen. 2022. Entailment Semantics Can Be Extracted from an Ideal Language Model. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 176–193, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011.

Eric Mitchell, Joseph Noh, Siyan Li, Will Armstrong, Ananth Agarwal, Patrick Liu, Chelsea Finn, and Christopher Manning. 2022. Enhancing Self-Consistency and Performance of Pre-Trained Language Models through Natural Language Inference. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1754–1768, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Koichi Nagatsuka, Clifford Broni-Bediako, and Masayasu Atsumi. 2023. Length-Based Curriculum Learning for Efficient Pre-training of Language Models. *New Generation Computing*, 41(1):109–134.

Mihai Alexandru Niculescu, Stefan Ruseti, and Mihai Dascalu. 2021. RoGPT2: Romanian GPT2 for Text Generation. In *2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 1154–1161. IEEE.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4885–4901. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 27730–27744. Curran Associates, Inc.

Marc Pàmies, Joan Llop, Francesco Multari, Nicolau Duran-Silva, César Parra-Rojas, Aitor Gonzalez-Agirre, Francesco Alessandro Massucci, and Marta Villegas. 2023. A weakly supervised textual entailment approach to zero-shot text classification. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 286–296, Dubrovnik, Croatia. Association for Computational Linguistics.

Dinesh Raghu, Suraj Joshi, Sachindra Joshi, and Mausam. 2022. Structural Constraints and Natural Language Inference for End-to-End Flowchart Grounded Dialog Response Generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 10763–10774, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2383–2392.

Livy Real, Erick Fonseca, and Hugo Gonçalo Oliveira. 2020. The ASSIN 2 Shared Task: A Quick Overview. In *Proceedings of the 14th International Conference on Computational Processing of the Portuguese Language (PROPOR 2020)*, pages 406–412.

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1586–1596, Brussels, Belgium. Association for Computational Linguistics.

Codruţ Rotaru, Nicolae-Cătălin Ristea, and Radu Tudor Ionescu. 2024. RoDia: A New Dataset for Romanian Dialect Identification from Speech. In *Findings of the Association for Computational Linguistics: NAACL 2024*.

Mobashir Sadat and Cornelia Caragea. 2022a. Learning to Infer from Unlabeled Data: A Semi-supervised Learning Approach for Robust Natural Language Inference. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4763–4776, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mobashir Sadat and Cornelia Caragea. 2022b. SciNLI: A corpus for natural language inference on scientific text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7399–7409, Dublin, Ireland. Association for Computational Linguistics.

Zhan Shi, Hui Liu, and Xiaodan Zhu. 2021. Enhancing descriptive image captioning with natural language inference. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 269–277, Online. Association for Computational Linguistics.

Henny Sluyter-Gäthje, Peter Bourgonje, and Manfred Stede. 2020. Shallow Discourse Parsing for Under-Resourced Languages: Combining Machine Translation and Annotation Projection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC)*, pages 1044–1050.

Ard Snijders, Douwe Kiela, and Katerina Margatina. 2023. Investigating multi-source active learning for natural language inference. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 2187–2209, Dubrovnik, Croatia. Association for Computational Linguistics.

Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and Nicu Sebe. 2022. Curriculum learning: A survey. *International Journal of Computer Vision*, 130:1526–1565.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset Cartography: Mapping and Diagnosing Datasets with Training Dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online. Association for Computational Linguistics.

Neeraj Varshney, Pratyay Banerjee, Tejas Gokhale, and Chitta Baral. 2022. Unsupervised Natural Language

Inference Using PHL Triplet Generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2003–2016, Dublin, Ireland. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NIPS)*, volume 30, pages 6000–6010.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019a. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32. Curran Associates, Inc.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019b. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *International Conference on Learning Representations (ICLR)*.

Yuxia Wang, Minghan Wang, Yimeng Chen, Shimin Tao, Jiaxin Guo, Chang Su, Min Zhang, and Hao Yang. 2022. Capture human disagreement distributions by calibrated networks for natural language inference. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1524–1535, Dublin, Ireland. Association for Computational Linguistics.

Gijs Wijnholds. 2023. Assessing Monotonicity Reasoning in Dutch through Natural Language Inference. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1494–1500, Dubrovnik, Croatia. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 1112–1122.

Roman V. Yampolskiy. 2013. *Turing Test as a Defining Feature of AI-Completeness*, pages 3–17. Springer Berlin Heidelberg, Berlin, Heidelberg.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3914–3923, Hong Kong, China. Association for Computational Linguistics.

## A Data Collection and Annotation

In this appendix, we present additional details about the data collection and annotation process.

**Definition of sentence relationships.** We use the same four relationship categories as Sadat and Caragea (2022b). These categories are defined as follows:

- **Contrastive:** captures pairs of sentences where one statement either opposes, contrasts with, criticizes, or points out a limitation in relation to the other.
- **Entailment:** sentence pairs are structured such that the first sentence provides a foundation, cause, or precondition for the outcome articulated in the second sentence.
- **Reasoning:** represents pairs of sentences where one statement explains, refines, generalizes, or conveys a meaning akin to the other.
- **Neutral:** This group consists of sentence pairs that are semantically unrelated. For example, sentences that are targeting completely different subjects.

**Linking phrases used for automatic labeling.** We next provide the specific linking phrases and transition words that are used to retrieve the sentence pairs from Wikipedia. These are listed in Table 7 for the contrastive pairs, Table 8 for the entailment pairs, and Table 9 for the causal pairs. There are no linking phrases to be listed for the neutral sentences. For each linking phrase, we report the number of extracted sentences.

We underline that after automatically assigning the labels, the linking phrases are removed from the sentences, making the NLI task much harder. Since we use the linking phrases to automatically label the training samples, leaving these phrases in place would make the task too simple for automated models. We argue that models should not only rely on obvious cues (such as the linking phrases) to reach human-level capabilities in NLI. We thus believe that the designed task is more suitable for the desired end goal. Moreover, the performance levels of the tested models are well above random chance, indicating that there are sufficient clues left for models to learn. Nevertheless, we perform an experiment where the linking phrases are included and the performance of Ro-BERT increases by considerable margins, reaching a micro $F_1$ of $0.81$ and macro $F_1$ of $0.68$ on the test set. The error rate of this model can be explained by the fact that the test

| Category | Romanian | English | Occurrences |
|---|---|---|---|
| Contrastive | Pe de altă parte | On the other hand | 1981 |
| | În contrast | In contrast | 616 |
| | În ciuda acestui fapt | In spite of this fact | 267 |
| | În opoziţie | In opposition | 49 |
| | În contradicţie | In contradiction | 40 |
| | În ciuda acestui lucru | In spite of this thing | 33 |
| | În ciuda acestor fapte | In spite of these facts | 23 |
| | În ciuda acestor lucruri | In spite of these things | 11 |
| | În mod contrar | In an opposite way | 11 |
| | Pe de cealaltă parte | On the other hand | 11 |
| | Cu toate acestea însă | Nevertheless | 9 |
| | Contrastând | In contrast | 5 |
| | În dezacord | In disagreement | 4 |
| | În sens opus | In the opposite sense | 3 |
| | În antiteza | In antithesis | 3 |
| | În contradictoriu | Contradictory | 2 |
| | Într-un contrast | In a contrast | 2 |
| | Contrar convingerilor | Contrary to the beliefs | 1 |
| | În pofida acestor lucruri | Contrary to these things | 1 |

Table 7: Original (Romanian) and translated (English) linking phrases and transition words used to retrieve contrastive sentence pairs, along with the number of retrieved sentences.

| Category | Romanian | English | Occurrences |
|---|---|---|---|
| Entailment | Cu alte cuvinte | In other words | 553 |
| | Adică | Specifically | 296 |
| | În esenţă | In essence | 155 |
| | Altfel spus | In other words | 149 |
| | Asta înseamnă că | This means that | 92 |
| | În fond | In essence | 53 |
| | Sintetizând | Synthesizing | 13 |
| | Rezumând | Summarizing | 12 |
| | În rezumat | In summary | 10 |
| | În termeni simpli | In simpler terms | 7 |
| | În traducere liberă | In translation | 6 |
| | Mai pe scurt | In short | 6 |
| | În alţi termeni | In other terms | 5 |
| | Simplificând | Simplifying | 5 |
| | Simplu spus | Simply said | 3 |
| | Mai concis | More concisely | 2 |
| | Pe larg | In broader terms | 5 |
| | În termeni populari | In more popular terms | 1 |
| | Într-o altă formulare | In another form | 1 |

Table 8: Original (Romanian) and translated (English) linking phrases and transition words used to retrieve entailment sentence pairs, along with the number of retrieved sentences.

set is labeled manually, so overfitting to the linking phrases is not necessarily the best solution.

**Annotator selection.** We offered the possibility of annotating sentence pairs to students enrolled at the AI/NLP Master programs from the University of Bucharest, in exchange for bonus points awarded for their final grades. We specified the exact benefits awarded for annotating between 6K and 7K sentence pairs, so the students knew exactly what to expect. The students can have either computer science or linguistics background. We enrolled the first three Master students who volunteered for this optional assignment. All annotators are native Romanian speakers. Aside from the instructions presented below, all students attended a lecture in which the NLI and NLU tasks were extensively discussed.

**Instructions to annotators.** The instructions given to the annotators, translated from Romanian to English, are as follows:

*Natural language inference (NLI) is the task of recognizing the relationship in sentence pairs. In this labeling task, you will be presented with sentence pairs of the form (Sentence A, Sentence B),*

| Category | Romanian | English | Occurrences |
|---|---|---|---|
| Causal | Astfel | Therefore | 16245 |
| | Prin urmare | As a consequence | 5202 |
| | Ca urmare | As an outcome | 4433 |
| | În consecinţă | Consequently | 1010 |
| | Aşadar | Hence | 948 |
| | Drept urmare | As a result | 601 |
| | În acest fel | In this manner | 574 |
| | Ca rezultat | As a result | 528 |
| | Din această cauză | Because of this | 520 |
| | Astfel că | Thus | 230 |
| | În concluzie | In conclusion | 197 |
| | Rezultatul este | The result is | 105 |
| | În rezultat | In result | 36 |
| | Din această cauza | Because of this | 17 |
| | Concluzionând | Concluding | 14 |
| | Pentru a finaliza | To finalize | 7 |
| | Ca o consecinţă a acestui fapt | As a consequence of this | 4 |
| | Într-o concluzie | In a conclusion | 3 |
| | Ceea ce a dus la | This lead to | 2 |
| | Ducând la | Leading to | 2 |
| | Conducând la | Leading to | 1 |
| | Provocând astfel | Thus causing | 1 |
| | Se poate concluziona că | It can be concluded that | 1 |
| | Ţinând cont de acestea | Considering these | 1 |

Table 9: Original (Romanian) and translated (English) linking phrases and transition words used to retrieve reasoning sentence pairs, along with the number of retrieved sentences.

| Premise | Hypothesis |
|---|---|
| A man in a blue shirt is performing a skateboarding trick near stairs while two other men watch. | Nobody has a shirt. |
| Two young children, one wearing a red striped shirt, are looking in through the window while an adult in a pink shirt watches from behind. | Nobody has a shirt. |
| A boy in a yellow t-shirt and pink sweater talks on a cellphone while riding a horse through a crowd of people who are looking on. | Nobody has a shirt. |

Table 10: Examples of sentence pairs from SNLI that exhibit annotation artifacts (Gururangan et al., 2018).

*without any additional context. Your task is to determine the relationship between sentences A and B, choosing one of the following four options:*

- *Contrastive: Select this category if one sentence presents a viewpoint or fact that is different from the other. Sentence B does not have to directly oppose Sentence A. Hence, this category includes comparisons, criticisms, or pointing out a limitation or unique aspect in one sentence about the content of the other.*
- *Reasoning: Choose this option if Sentence A provides a basis or rationale that leads to or explains the information in Sentence B. Look for a logical sequence where the first sentence sets up a foundation that the second sentence builds upon or concludes from.*
- *Entailment: Use this category if one sentence expands on or specifies the information given in the other, essentially providing more details or a specific instance of the general idea of the premise.*
- *Neutral: Opt for this category if the two sentences are unrelated, with each standing independently without referring to, supporting, or elaborating on the other. A sentence pair that does not fit in the other categories should be labeled as neutral.*

The above instructions were followed by one pair of sentences from each category, to exemplify the four categories. The provided examples were manually labeled by the authors, to avoid any potential mistakes resulting from the automatic labeling process.

**Automatic vs. manual annotations.** In the main paper, we used Cohen's Kappa in order to determine the agreement between automatic and manual annotations. As an alternative way to estimate the alignment between automatic and manual annotations, we compute the micro and macro $F_1$ scores using automatic labels as predictions and manual labels as ground-truth. The resulting micro $F_1$ is $0.83$ and the macro $F_1$ is $0.69$. Notice that these

| Romanian | | English | | Class |
|---|---|---|---|---|
| **Premise** | **Hypothesis** | **Premise** | **Hypothesis** | |
| Sub umbra soacrei ei, Ulrica nu a fost niciodată fericită sau cel puţin viaţa ei de la curte nu a fost fericită. | S-a spus că viaţa ei privată cu regele şi copiii ei a fost una foarte fericită. | Under the shadow of her mother-in-law, Ulrica was never happy, or at least her life at court was not happy. | It was said that her private life with the king and her children was a very happy one. | Contrastive |
| În teoria jocurilor, un joc cu sumă zero sau nulă descrie o situaţie în care câştigul unui participant este perfect echilibrat pierderea unui alt participant. | Orice situaţie care funcţionează ca un joc cu sumă zero presupune că orice câştig al unui participant necesită o pierdere egală a altui participant. | In game theory, a zero sum or null sum game describes a situation in which the gain of a participant is balanced by the loss of another participant. | Any situation that works like a zero sum game assumes that any gain of a participant will require the equal loss of another participant. | Entailment |
| Cum pieţele bursiere au scăzut în septembrie 2008, fondul a putut să cumpere multe acţiuni la preţuri scăzute. | Pierderile suportate de turbulenţele de pe pieţe au fost recuperate până în noiembrie 2009. | As the stock market fell in September 2008, the fund was able to buy more stocks at lower prices | The losses incurred by the market turbulence were recovered by November 2009. | Reasoning |
| Conform Catalogue of Life specia “Drosophila gibberosa” nu are subspecii cunoscute. | Interesul său în domeniul meteorologiei l-a determinat să se ocupe de aviaţie. | According to the Catalog of Life, the species “Drosophila gibberosa” has no known subspecies. | His interest in meteorology led him to pursue aviation. | Neutral |

Table 11: Original and translated examples of sentence pairs extracted from the Romanian Wikipedia, which are automatically labeled via linking phrases and transition words. We show one example per class.

| Group | E2L | A | H2L |
|---|---|---|---|
| Fleiss Kappa | 0.75 | 0.72 | 0.69 |

Table 12: Inter-rater agreements for easy-to-learn (E2L), ambiguous (A) and hard-to-learn (H2L) samples from the validation set.

scores are much higher than the machine learning models reported in Table 5, confirming that the NLI task on our dataset is not yet saturated.

In Table 11, we show one randomly selected sentence pair per class. The examples are taken from the training set, being labeled automatically. We observe that the assigned labels are correct.

To determine if the H2L samples have a higher percentage of noisy labels, we applied data cartography to the validation set. After training Ro-BERT on the validation set to find the E2L, A, and H2L samples, we were able to compute the inter-rater agreement for each subgroup. We obtained the Kappa coefficients reported in Table 12. The reported inter-rater agreements confirm that the H2L group contains more examples with potentially wrong labels.

## B Hyperparameter Tuning

The hyperparameters of all models are determined via grid search. We used the following intervals for the various hyperparameters:

- Learning rates between $10^{-2}$ and $10^{-6}$.
- Mini-batch sizes between $8$ and $256$ (using only powers of two as options).
- Number of hidden units for the classification head in the set $\{128, 256, 512, 768\}$.
- Dropout rates between $0.1$ and $0.5$.
- Values for the regularization hyperparameter $C$ of the SVM and Softmax models between $0.1$ and $1000$.
- SVM kernel options between *linear* and *RBF*.
- Tolerance (for optimization) between $10^{-2}$ and $10^{-6}$.

For SVM, we obtain optimal validation results with $C = 0.5$, the maximum number of iterations set to $2500$, and the tolerance level equal to $10^{-5}$. For Softmax, we reach the best validation results with $l_2$ regularization with $C = 1$, and the tolerance level equal to $10^{-3}$.

The RoGPT2 model employs the AdamW optimizer (Loshchilov and Hutter, 2017) with a learning rate of $10^{-3}$, on mini-batches of 64 samples. The model is trained for at most 10 epochs, with early stopping. The fine-tuned multilingual BERT is optimized with AdamW for 10 epochs on mini-batches of 256 samples, using a learning rate of $10^{-3}$. As the multilingual BERT baseline, the models based on Ro-BERT employ the AdamW optimizer with a learning rate of $10^{-3}$, on mini-batches

| Romanian | | English | | Correct Class / Predicted Class |
|---|---|---|---|---|
| **Premise** | **Hypothesis** | **Premise** | **Hypothesis** | |
| Dacă nu autorităţile sunt cele care să decidă doctrina, şi dacă argumentele lui Martin Luther pentru preoţia tuturor credincioşilor sunt duse la extrem, caz în care Biserica ar fi condusă de cei aleşi, atunci a avea monarhul în fruntea Bisericii ar fi intolerabil. | Dacă monarhul e numit de Dumnezeu să fie în fruntea bisericii, atunci e intolerabil ca parohiile locale să îşi decidă singure doctrina. | If authorities are not the ones to decide the doctrine, and if the arguments of Martin Luther for the priesthood of all the believers are taken to the extreme, in which case the Church would be led by the chosen, then to have the monarch as the head of the Church would be intolerable. | If the monarch is named by God to be the head of the Church, then it is intolerable for local churches to decide their own doctrine. | Contrastive / Reasoning |
| Dar Biserica învaţă că realitatea şi eficacitatea Sfintelor Taine ale Bisericii realizate de preoţi nu depind de virtuţile lor personale, ci de prezenţa şi lucrarea lui Iisus Hristos, Care acţionează în Biserica Sa prin Sfântul Duh. | Preoţia este un dar al bisericii - trupul lui Hristos în istorie -, în slujba comunităţii ecleziale, iar nu un dar personal al celui care devine preot. | But the Church teaches that the efficacy of the Holy Teachings of the Church preached by priests do not depend on their personal virtues, but on the presence and work of Christ, Who acts in His Church through the Holy Spirit. | Priesthood is a gift of the Church, body of Christ in History - in the work of the ecclesiastical community, not a gift to the one who becomes a priest | Entailment / Reasoning |

Table 13: Original and translated examples of sentence pairs that are misclassified by our best performing model (Ro-BERT + Cart-Stra-CL++).

of 256 samples. The models are trained for at most 10 epochs, with early stopping.

All other hyperparameters are set to their default values. Please note that we release the code to reproduce all baselines, along with the RoNLI corpus[3].

## C Error Analysis and Task Complexity

To find interesting language-specific phenomena in Romanian NLI, we perform an error analysis of the best scoring model (Ro-BERT + Cart-Stra-CL++). We present two examples mislabeled by our best model in Table 13. The first example showcases a lack of understanding of Romanian sentence structure and formulation (sentences are mislabeled as reasoning instead of contrastive). The sentences exemplify complex Romanian phrasing structures with multiple clauses and conditional phrases. Such complexity can lead to classification errors, if the classifier is not apt at parsing and understanding the nuances in Romanian sentence construction. In the second example, both sentences are structurally complex, with multiple subordinate clauses. They share thematic elements related to Christianity, which may contribute to their classification error due to overlapping religious concepts. The error analysis reveals some interesting findings. For example, Romanian tends to have a complex sentence structure, which often confuses NLI models.

We note that our assertions are supported by the study of Dobrovie-Sorin (1994). We summarize the main aspects that influence the complexity of the Romanian language below:

- The clitic system of Romanian exhibits a level of complexity not found in other Romance languages, characterized by the presence of not only pronominal clitics, but also adverbial clitics and cliticized conjunctions. Significantly, Romanian includes auxiliary elements that serve as verbal clitics, distinguishing its syntactic structure further.

- The constituent structure of Romanian clauses, with particular emphasis on subjunctive clauses, displays marked distinctions from those observed in other Romance languages, as extensively discussed in Chapter 3 of (Dobrovie-Sorin, 1994). This divergence underscores the unique syntactic configurations inherent to Romanian.

- Romanian demonstrates a predilection for employing subjunctive constructions in contexts where other Romance languages would typically resort to infinitive forms. This syntactic preference highlights a notable deviation in structural utilization across Romance languages.

- The ambiguous nature of the particle ‘a’ holds substantial comparative significance. Such

[3] https://github.com/Eduard6421/RONLI

ambiguous particles are absent in other Romance languages and English, yet they are prevalent in Verb-Subject-Object languages, including Welsh. This phenomenon presents an intriguing area of study from a comparative linguistic perspective.