Final Report: Adding Multilinguality to LimitedInk

Jonathan Hus
jhus@gmu.edu

Wren McQueary
wmcquear@gmu.edu

1 Introduction

1.1 Task / Research Question Description

LimitedInk is a self-explaining model for generating rationales on language problems, specifically classification problems (Shen et al., 2022). It presents rationales as sets of words from the original prompt which best explain the output. LimitedInk's distinguishing feature is that it can produce rationales of a target length. Its creators chose to include this feature because they wanted to empirically interrogate the common attitude that shorter explanations are inherently more helpful for human understanding. Using LimitedInk, the authors found that brevity alone does not make an effective rationale. Rather, there is a balance to be struck between brevity and sufficiency.

1.2 Motivation and Limitations of Existing Work

LimitedInk is an explainable AI tool. The fields of explainable AI (Vilone and Longo, 2021) and interpretable AI (Rudin, 2019) represent efforts to redress and reduce cases where black-box models make decisions which adversely affect people's lives in dire ways (Obermeyer et al., 2019; Martin, 2020). Although interpretable models are more transparent and reliably understandable than explainable models like LimitedInk, explainable models are popular for their expressiveness and therefore unlikely to disappear anytime soon – especially because large neural networks are intrinsically uninterpretable (Rudin, 2019).

LimitedInk itself is a helpful tool for generating intuitive rationales, since a user can sweep through the rationale length space to find a rationale that works best (Shen et al., 2022). This is relevant to our group's interests in AI explainability and accountability. However, the tool isn't multilingual, since its datasets (known collectively as ERASER) consist only of English data (DeYoung et al., 2019), even though its pre-trained model, DistilBERT, is multilingual (Sanh et al., 2019). This limits the usefulness of the tool in explaining AI on non-English inputs, which is a waste of LimitedInk's potential.

1.3 Summary of Approach

We started this project by training English-only baseline models to verify the authors' reported results. Next, to broaden LimitedInk's usefulness, we aimed to train a multilingual version of the tool. Lacking a multilingual equivalent for any of the datasets used by the authors, we instead found a similar dataset, e-XNLI, which was suitable for training and had sufficient similarity to one of the authors' datasets, FEVER, to allow for rough comparison (Zaman and Belinkov, 2022). We used the same pre-trained model (DistilBERT) as the original authors and trained it on e-XNLI.

We also tested the robustness of the original English-only tool to data imperfections. Using CheckList (Ribeiro et al., 2020), we generated a battery of noisy movie reviews by creating and undoing contractions, creating typos, and swapping names of people with stock names from a list. We then compared a LimitedInk model's outputs on the noised inputs to their pre-noised counterparts. We also performed a sensitivity analysis to see how severely LimitedInk's performance might be changed by using different training seeds.

1.4 Summary of Results

Our baseline implementation mostly validated the authors' findings. Although each of our models scored worse than the authors' corresponding model, they did so by only a small margin (0-7%). We also found that our models achieved higher F1 scores versus human annotations when producing shorter rationales, which contradicts the authors'

findings, but again only by a small margin (4%).

Our multilingual model showed that LimitedInk can be modified to support multiple languages with minimal modification to the baseline architecture. Using our selected dataset, the model was able to achieve an F1 score of 0.62 on the English-only samples, compared to an F1 score of 0.60 on the full multilingual samples.

Our robustness study found that LimitedInk has high robustness to name changes, but suffers from occasional severe outliers when contraction changes or typos are introduced, even though the standard deviation of it output remains low. LimitedInk's sensitivity to training seeding is low; we saw at most a 2% difference in its F1 score across 5 seeds.

1.5 GitHub Repository

The repository for our project is located here: `https://github.com/WrenMcQueary/cs_678_final_project.git`

In addition to our own source files, the repository also contains files from the original LimitedInk repository (`https://github.com/huashen218/LimitedInk`) and e-XNLI repository (`https://github.com/KeremZaman/explaiNLI`) (Shen et al., 2022; Zaman and Belinkov, 2022).

2 Approach

Our approach consisted of three phases: baseline results replication 2.1, multilinguality 2.2, and robustness testing 2.3.

2.1 Approach: Baseline Results Replication

To replicate the authors' original results, we trained LimitedInk on all five datasets from the paper (Movies, BoolQ, Evidence Inference, FEVER, and MultiRC) using the same hyperparameters and rationale lengths from the original paper.

Our baseline replication consisted of two main replication tasks: (1) replicating the authors' most successful models, and (2) replicating the authors' findings that shorter rationales did not produce higher F1 scores.

To replicate the authors' most successful models, we reconfigured the authors' repository, particularly their config files, for use with SLURM, and then trained LimitedInk models using the same k (the total percentage of the input text to be highlighted as the rationale) and hyperparameters from the paper. The authors reported a separate best model for each of the five datasets (discussed in 3.1), each requiring a specific value of k, as shown in Table 1. We trained a separate model for each dataset, matching the k value used by the authors for that dataset, and compared the F1 values for our rationales to theirs.

To replicate the authors' findings that shorter rationales did not produce higher F1 scores versus human annotations, we used the Movies dataset because the authors primarily used this dataset when making this point (Shen et al., 2022). Using the same hyperparameters as the authors, we trained a model for each of the following k values: (0.1, 0.2, 0.3, 0.4, 0.5). We then observed the trends in precision, recall, and F1 with respect to the k value used.

2.2 Approach: Multilinguality

In the second phase, we tested the performance of LimitedInk when using multilingual data. To achieve this goal, two major modifications were required. First, we had to identify a multilingual dataset that was comparable to a dataset used in the original LimitedInk paper. LimitedInk is intended for classification tasks, so labeled training data were necessary, though the specific classification categories were not constrained. Additionally, the training data needed evidence annotations, identifying the portions of the training text that most influenced the classification. Obviously, the data had to be multilingual as well. There are not many available datasets that meet these requirements, but we selected e-XNLI, which is described below in Section 3.2.

The second modification to the LimitedInk baseline was substituting the model and associated tokenizer with a multilingual version. This trade space has more options than what is available for the dataset, with many multilingual models from which to choose. We opted for a multilingual DistilBERT model for reasons specified below. Mainly, the original LimitedInk model uses DistilBERT, so using a multilingual version provides a nice comparison with the original results.

2.3 Approach: Robustness

We investigated the robustness of an English-only LimitedInk model to noise in the validation data. We used CheckList (Ribeiro et al., 2020), a tool for robustness testing of language models. CheckList

has empirical evidence of being effective for creating tests and finding actionable bugs in models. It can test for a variety of capabilities, including vocabulary, parts of speech, taxonomy, fairness, robustness, and sequence of events. CheckList supports three main test types: minimum functionality tests (which are similar to unit tests), invariance tests (in which the model's output on two inputs are expected to be the same), and directional expectation tests (in which the model's output on one input is expected to change in a certain way compared to its output on a different input). To conduct our robustness study, we used CheckList's robustness functionality to compose a battery of invariance tests using pairs of the same input pre- and post-noising.

The LimitedInk model under test was a DistilBERT classification model for the Movies dataset using a rationale length ratio k of 0.5. The Movies dataset consists of movie reviews paired with sentiment labels (positive/negative), and is explained in greater depth in Section 3.3, along with the results of the robustness study. Given a movie review, the classification model outputs a value in the range $[0, 1]$, where a score of 0 indicates a negative sentiment with total confidence, and 1 indicates a positive sentiment with total confidence. Values below 0.5 indicate a negative review, and values above 0.5 indicate a positive review.

Our robustness study covered the following distortions of input data: perturbations in contractions (eg `don't` $\leftrightarrow$ `do not`), typos (eg `language` $\rightarrow$ `lanugage`), and name changes (eg `Ada Lovelace walked` $\rightarrow$ `Lynn Marshall walked`). We chose these perturbations because they are the most prevalent in real-world online texts, to best simulate the model's effectiveness in a practical scenario. For each type of perturbation, 250 perturbed inputs were generated. The model was then tested on both the original and perturbed input to measure how much the perturbation threw off its output score. A flowchart of our robustness study is shown in Figure 1.

We chose to measure disturbances in the model's output score, rather than class, because it was a binary classification model. The model's output scores still encode the estimated class, and also include the model's confidence, which is a much more detailed and useful piece of information. For example, using a score lets us distinguish between a case where an output changed from confident negative to confident positive, and a case where an output changed from unconfident negative to unconfident positive. The first case indicates less robustness, but both cases would appear identical if only the class label were used as a measurement.

Results Reproduction		
Dataset (k value)	**Their F1**	**Our F1**
Movies (50%)	0.90	0.90
BoolQ (30%)	0.56	0.49
Evidence Inference (30%)	0.50	0.45
FEVER (40%)	0.90	0.89
MultiRC (50%)	0.67	0.62

Table 1: Comparison of our models' performance to those presented in the paper.

To apply **contraction perturbations** to a review, CheckList attempted to find all current contractions, as well as all words that could be contracted together. CheckList then contracted or expanded all the detected words, inverting which were contracted versus uncontracted. To apply **typo perturbations** to a review, CheckList chose 10 pairs of neighboring characters to swap. To apply **name change perturbations** to a review, CheckList detected first and last names from two lists of known first and last names, and then swapped them for other names on the lists. When evaluating name change perturbations, only reviews where CheckList actually performed such a swap were used.

We also assessed the sensitivity of a training session to the random seed used. We once again used the Movies dataset, and a rationale of 50% with the same hyperparameters as the paper. We trained five models with unique seeds and compared their F1 scores on the validation dataset. The results are discussed in Section 3.3.3.

3 Experiments

3.1 Experiments: Baseline Results Replication

As our baseline, we used the LimitedInk repository to train LimitedInk models on the five datasets covered in the original paper (Shen et al., 2022). In order to verify the authors' results, we used the same hyperparameters and rationale lengths from the original paper. A comparison of the authors' models with our reproductions is shown in Ta-

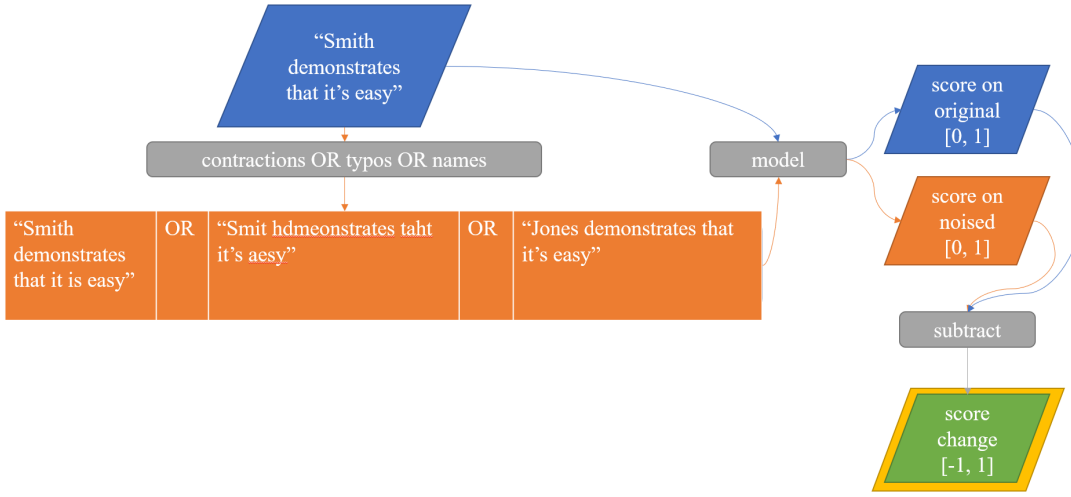

Figure 1: Flowchart of our robustness study. Original reviews are noised via changes in contractions, typos, or names. Then the difference between the model's score on the original review and the noised review is calculated. Smaller differences indicate greater robustness. For brevity, this figure uses a single sentence rather than a full-length movie review.

Movies Data Set Performance			
Method	**Precision**	**Recall**	**F1**
LimitedInk (50%)	0.91	0.90	0.90
Ours - 10%	0.89	0.89	0.89
Ours - 20%	0.89	0.88	0.88
Ours - 30%	0.86	0.86	0.86
Ours - 40%	0.87	0.87	0.87
Ours - 50%	0.87	0.86	0.86

Table 2: Model Performance Based on Rationale Length.

ble 1. Our results vary only marginally from the authors', although we note that all our reproduced results have a lower F1 score than those presented in the paper.

When evaluating their models, the authors used a number of rationale lengths: (10%, 20%, 30%, 40%, 50%) of the input text. We trained separate models with each of these lengths on the Movies data set, same hyperparameters as the authors (e.g., learning rate, number of epochs). The results, which are shown in Table 2, indicate that rationale length doesn't have a dramatic impact on the performance. In fact, our shorter rationales performed better than our longer ones. This conflicts with the LimitedInk paper, in which the authors stated that the 50% rationale length performed the best for the Movies dataset. However, our best and worst F1 scores differ by only 4%.

3.2 Experiments: Multilinguality

Our aim was to test how LimitedInk would perform in a multilingual environment. The original paper used five English-only datasets taken from the Evaluating Rationales And Simple English Reasoning (ERASER) benchmark (DeYoung et al., 2019). These datasets covered sentiment analysis (Movies) (Zaidan et al., 2007), question answering (BoolQ) (Clark et al., 2019), reading comprehension using multiple sentences (MultiRC) (Khashabi et al., 2018), claim or fact extraction and verification (FEVER) (Thorne et al., 2018), and evidence inference on medical data (Evidence Inference) (DeYoung et al., 2020). Other datasets from ERASER were not tested in the authors' paper. Notably, the e-SNLI dataset (Camburu et al., 2018) was not used. The e-SNLI data is a natural language inference dataset with natural language explanations. Each sample in the data contains a premise, a hypothesis, and a label indicating whether the hypothesis is entailed in, contradicted by, or neutral to the premise. In addition, each sample is human-annotated to indicate which parts of the premise supported the label. This additional aspect of the data (i.e., human-annotated evidences) is crucial for use in LimitedInk. Although the authors' chose not to use this data for their experiments, there is nothing that precludes its use.

For our multilinguality testing, a major challenge was to identify training data that was not only multilingual, but also had human-annotated

evidences. We surveyed available datasets for training data that met this criteria, but found the options lacking. The Cross-lingual Natural Language Inference (XNLI) corpus is a crowd-sourced collection of 5,000 test and 2,500 dev pairs for the MultiNLI corpus (Conneau et al., 2018). Sample data is shown in Figure 2. XNLI met the multilingual criterion, but lacked evidence annotations. However, e-XNLI has both the multilingual examples and the explanations. They are not human annotations, however, and instead obtain the non-English explanations "by extracting highlights for the English part of XNLI and projecting along word alignments to other languages."(Zaman and Belinkov, 2022) Regardless, e-XNLI has the attributes necessary for our testing and is the dataset used for our experiments.

e-XNLI consists of over 75K examples in 15 different languages: Arabic, Bulgarian, German, Greek, English, Spanish, French, Hindi, Russian, Swahili, Thai, Turkish, Urdu, Vietnamese, and Chinese. Similar to e-SNLI, the data consists of a premise and a hypothesis, as well as a label that indicates whether the premise entails, contradicts, or is neutral to the hypothesis. The samples also contain highlighted annotations signifying the words most relevant in determining the label. A sample is shown below.

Language: en
Label: contradiction
Premise: But anyway, the animals would get loose all the time, especially the goats.
Hypothesis: The goats were kept safe and secure.
Premise_Highlighted: But anyway , the *animals* would get *loose* all the time , especially the goats .
Hypothesis_Highlighted: The *goats* were *kept* *safe* and *secure* .

After formatting the data to be compatible with LimitedInk, we ran a number of experiments. First, we analyzed the performance of LimitedInk on just the English examples. This afforded us the opportunity to work through any issues specific to the new dataset and it provided a means of comparing accuracy between the published LimitedInk results and a novel dataset, which though similar had not been contemplated when the authors were coding their models. The results are shown in Table 3. Training and validation loss curves are shown in Figure 3. For comparison, similar charts for the Movies data (Figure 4), the FEVER data (Figure 5), and the Evidence Inference data (Figure 6) are provided, using the training logs from our reproduced results. As seen in the charts, while the Movies, FEVER, and Evidence Inference results have well-behaving loss curves indicating that the model is improving its performance with each epoch, the e-XNLI results have some variability with respect to the number of epochs, although in general it appears that the model is learning. One possible explanation for the validation loss behavior is the limited number of English examples that were provided during training. Since the e-XNLI data set has roughly 75K examples in 15 languages, that means that there are about 5K English samples. From that, 10% each are set aside for test and validation data. This quantity of data simply might not be large enough to obtain results similar to the other data sets. That being said, the results are promising.

e-XNLI English-Only Performance				
	Precision	**Recall**	**F1**	**Support**
contradiction	0.69	0.64	0.67	188
neutral	0.55	0.58	0.56	151
entailment	0.61	0.63	0.62	162
accuracy			0.62	501
weighted avg	0.62	0.62	0.62	501

Table 3: Model Performance on e-XNLI English-Only Data.

Once we were able to run LimitedInk on the English-only subset of e-XNLI and had obtained encouraging results, we then turned our attention to the full multilingual version of e-XNLI. One of our first decisions was selecting which multilingual model we wanted to use. Initially, we attempted to use multilingual BERT and XLM-RoBERTa, since they are both advertised as having superior performance compared to DistilBERT. However, when we tried them in our experiments, they did not perform well at all; in fact, after training they always degenerated into selecting only one of the three categories for every sample during inference. We were unable to identify the root cause, but our hypothesis is that this is due to how LimitedInk formats the training examples that are provided to the model. After much trial and error, we settled on using the multilingual version of DistilBERT. Our loss curves are shown in Figure 9.

Training a DistilBERT model yielded better re-

Language	Premise / Hypothesis	Genre	Label
English	You don't have to stay there. You can leave.	Face-To-Face	Entailment
French	La figure 4 montre la courbe d'offre des services de partage de travaux. Les services de partage de travaux ont une offre variable.	Government	Entailment
Spanish	Y se estremeció con el recuerdo. El pensamiento sobre el acontecimiento hizo su estremecimiento.	Fiction	Entailment
German	Während der Depression war es die ärmste Gegend, kurz vor dem Hungertod. Die Weltwirtschaftskrise dauerte mehr als zehn Jahre an.	Travel	Neutral
Swahili	Ni silaha ya plastiki ya moja kwa moja inayopiga risasi. Inadumu zaidi kuliko silaha ya chuma.	Telephone	Neutral
Russian	И мы занимаемся этим уже на протяжении 85 лет. Мы только начали этим заниматься.	Letters	Contradiction
Chinese	让我告诉你，美国人最终如何看待你作为独立顾问的表现。 美国人完全不知道您是独立律师。	Slate	Contradiction

Figure 2: XNLI Examples

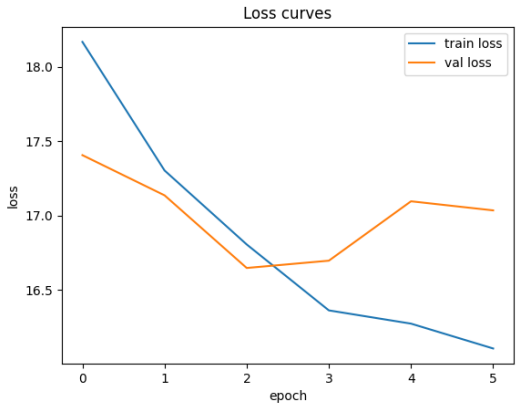

Figure 3: Training and Validation Loss on e-XNLI English-Only Data

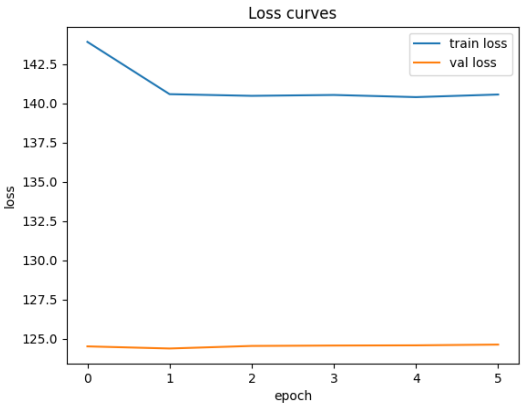

Figure 5: Training and Validation Loss on FEVER Data

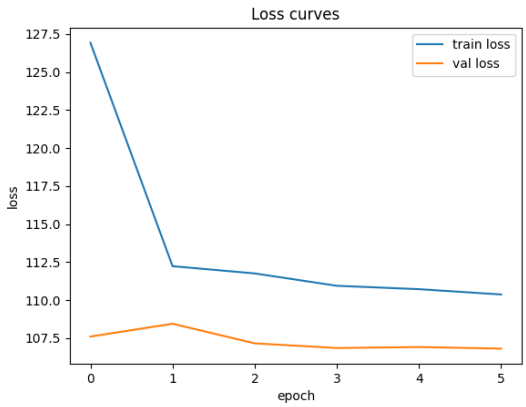

Figure 4: Training and Validation Loss on Movies Data

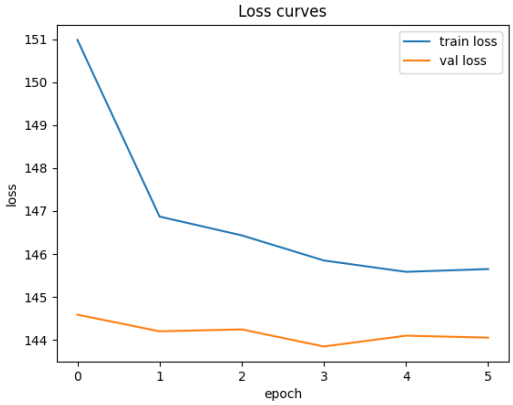

Figure 6: Training and Validation Loss on Evidence Inference Data

e-XNLI Performance				
	Precision	**Recall**	**F1**	**Support**
contradiction	0.65	0.59	0.62	2820
neutral	0.54	0.58	0.56	2265
entailment	0.60	0.62	0.61	2430
accuracy			0.60	7515
weighted avg	0.60	0.60	0.60	7515

Table 4: Model Performance on e-XNLI Data (all 15 languages).

sults, as shown in Table 4. The difference in the F1 scores between the e-XNLI English-only and the complete e-XNLI datasets is only 0.02, with weighted F1 scores of 0.62 and 0.60, respectively. Comparing these results to the baseline LimitedInk results is not straightforward, but the closest comparison is with the FEVER results. FEVER has two possible classifications (refutes and supports) while e-XNLI has three possible classifications (entails, contradicts, and neutral), so e-XNLI classification is an inherently harder problem. That said, the e-XNLI results are substantially worse, as shown in Table 5. As mentioned above, this could be partly due to the limited number of English examples available. Additionally, the quality of e-XNLI as a training set has not been as substantiated as the ERASER sets.

3.2.1 Evidence Highlights

The LimitedInk model outputs evidence annotations, indicating which tokens were most influential in determining whether the hypothesis logically followed from the premise. Figure 7 shows a case when the model accurately predicted the classification and a case when the model predicted the wrong classification for an English example. These results were derived from a training run on the English-only dataset where the evidence length hyperparameter was 0.4. As can be seen in the figures, the model selects reasonable though not perfect highlights. Figure 8 shows a similar set of examples on Greek samples, which were obtained from a training run on the full e-XNLI dataset with the evidence length hyperparameter set to 0.5. The results on the multilingual data were less impressive than those from the English data, which again could due to the quality of the e-XNLI data.

3.3 Experiments: Robustness

As discussed in Section 2.3, we used the Movies dataset for our robustness test. The Movies dataset

Dataset	**F1 Score**
FEVER	0.90
e-XNLI (English only)	0.62
e-XNLI (complete)	0.60

Table 5: e-XNLI Model Comparison Using Weighted F1 Scores. FEVER is chosen since the task and setup are most similar.

contains 2,000 movie reviews which are either positive or negative and have a typical length of 1-2 paragraphs. Section 2.3 also discusses the meanings of the model's output scores, whose displacements against noise we measured to assess the model's robustness.

An example entry from the Movies dataset is shown below. For brevity, only an excerpt from the review and evidences is shown. The query is the same for every sample in the dataset.

Review: ...but cinematographer Peter Deming (don't say a word) ably captures the dreariness...
Classification: positive
Query: What is the sentiment of this review?
Evidences: ...but cinematographer Peter Deming (don't say a word) *ably captures* the dreariness...

Figure 10 and Table 6 summarize the results of the results of the robustness study. We found the model to have high robustness to name changes, with a maximum score disturbance of 0.03949 and a standard deviation of 4.881E-03. However, for contractions and typos, the model was less robust. Its standard deviations remained quite low, at 0.09388 and 0.09658 respectively, but outliers were more numerous and severe. We qualitatively analyze some of these outlier cases in sections 3.3.1 and 3.3.2.

We believe the model had greater robustness to name changes because names carry less intrinsic sentiment, and because names can be cleanly swapped without altering the meaning of a sentence, except for extremely rare cases where a name that is already used in the passage might be inserted. Contractions and typos do not share these protections. For example, contracting `How easy it is!` to `How easy it's!` renders the sentence incoherent, and applying a typo to change `magic rap` to `magi crap` changes the sentiment from positive to negative.

3.3.1 Contraction Outliers

In one case, inverting a review's contractions increased the sentiment by 0.97, changing the es-

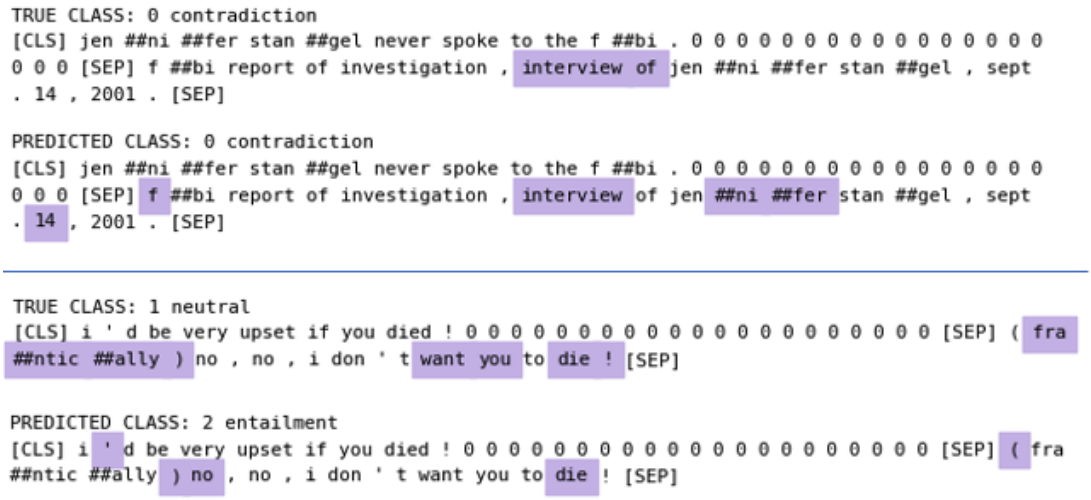

Figure 7: Evidence Highlights for English Examples. The figure shows a case when the correct classification was predicted and a case when an incorrect classification was predicted.

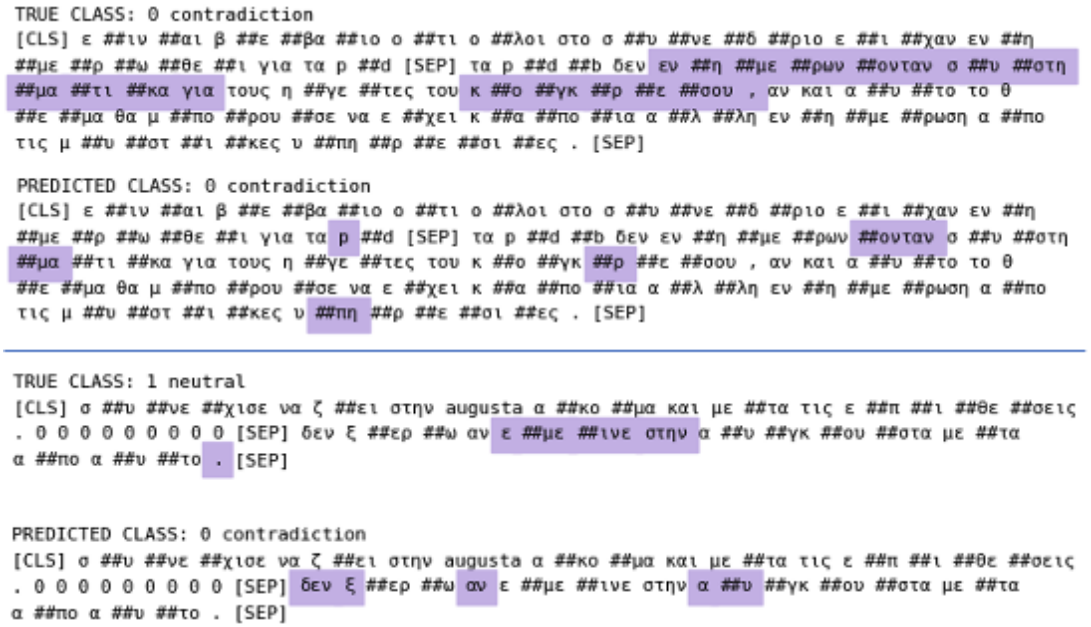

Figure 8: Evidence Highlights for Greek Examples. The figure shows a case when the correct classification was predicted and a case when an incorrect classification was predicted.

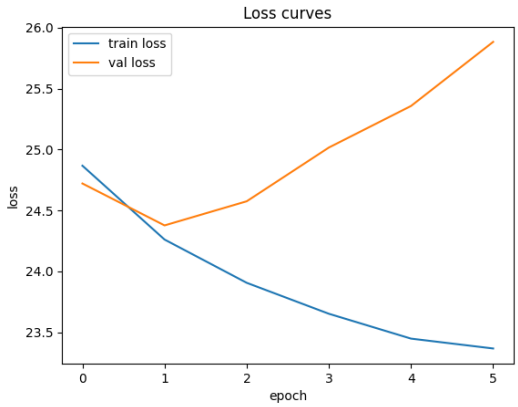

Figure 9: Training and Validation Loss on e-XNLI Data (all 15 languages)

timate from nearly 0 (the correct sentiment) to nearly 1. The perturbations which caused this change are shown in red in Figure 11. Two of the contraction changes (highlighted in yellow) distorted the meaning of text. It's possible that these play a significant role in the misclassification.

3.3.2 Typo Outliers

In another case, adding typos to a review changed the sentiment from nearly 1 (the correct sentiment) to nearly 0. The perturbations which caused this change are shown in Figure 12. Note that the review is relatively short, so the 10 typos are densely packed and do more damage to the overall meaning of the review. In particular, two typos (highlighted in yellow) make it difficult to understand the positive sentiment of a sentence that compliments the movie particularly lavishly. If this sentence had been understood, then the estimated sentiment would likely have been closer to the original.

3.3.3 Sensitivity to Random Seeding

As discussed in Section 2.3, we also ran a seed sensitivity analysis to investigate the effects of the training seed on overall model performance. We again used the Movies dataset. The F1 scores of each of these models are shown in Table 7. The F1 score varied across the seeds with a standard deviation of 0.006325. This standard deviation is quite small and suggests insensitivity to random seeding.

4 Related Work

Our work is immediately based upon (Shen et al., 2022), in which LimitedInk was created and evaluated. Whereas LimitedInk could originally only be used on English, our approach results in a multilingual model.

(Shen et al., 2022) questions an assumption that shorter rationales are intrinsically better. This assumption began with (Vafa et al., 2021), in which a new greedy algorithm for producing rationales was proposed. This algorithm is computationally efficient and had very high accuracy compared to the state of the art. It just so happened that the rationales produced by their model are particularly short, hence the assumption that shortness is causal to quality. Whereas their approach is greedy, LimitedInk (and thus our variant of it) is based on BERT.

An older paper, (Lei et al., 2016), proposes an approach for rationale highlighting in which two components, a generator and encoder, are trained in tandem. The generator learns a probability distribution over rationales, and this probability distribution is passed to the encoder to generate a prediction. At the time, this approach was more effective than attention-based approaches. This is interesting, since (Shen et al., 2022) is based on BERT, which is attention-based.

(Mosca et al., 2023) presents a tool for human-in-the-loop training finetuning of rationale-producing models. The tool is agnostic of architecture, and can be configured to work with a variety of HuggingFace models. This tool could allow for human finetuning of LimitedInk, as well as our multilingual variant of it.

5 Conclusions and Future Work

Our work consisted of three main parts: a reproducibility study of the authors' findings, an multilingual extension of the authors' tool, and a robustness study of the authors' model.

In our reproducibility study, we successfully replicated the accuracy of the authors' models. However, while attempting to replicate the authors' findings that shorter rationales *do not* lead to better F1 scores compared to human annotations, we instead found contradictory evidence. Our shortest rationales achieved the best F1 scores, although only marginally so (4% better than our worst rationales).

Our multilingual e-XNLI model was successful, achieving an F1 score of 0.60, compared to the English-only e-XNLI model's F1 score of 0.62. Direct comparison of our multilingual model to

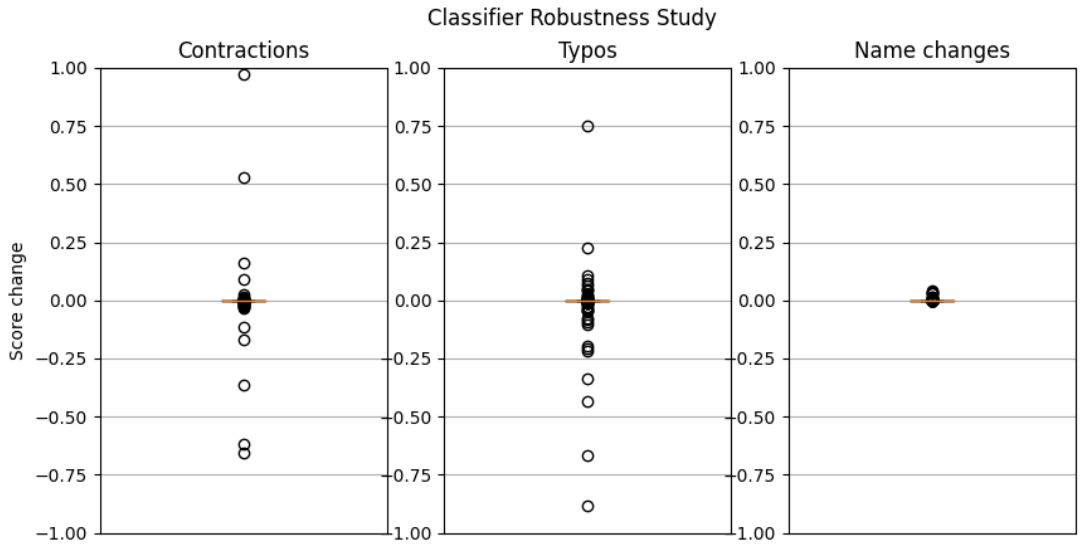

Figure 10: Box-and-whisker plot of changes in positivity score due to perturbations in contractions, typos, and name changes. Note that the boxes and whiskers are vanishingly small, but a small number of wide outliers exist.

Original	Altered
must realize her place and direction in life and understand that her life is not just a big party .\nthe film	must realize her place and direction in life and understand that her life isn\'t just a big party .\nthe film
million times .\nit is a classic cautionary tale .\nan echo .\na big , fat and expensive clich ? .\na shadow	million times .\nit\'s a classic cautionary tale .\nan echo .\na big , fat and expensive clich ? .\na shadow
failures and errors , that it is impossible to sum them all up in one review .\nfor some paranoid reason	failures and errors , that it\'s impossible to sum them all up in one review .\nfor some paranoid reason
place where happy alcoholics and cheerful drug addicts are not allowed to smoke , drink or watch tv after 11 .\nit is simply too light , simplified and unnecessary sweet to be taken seriously .\nbut the worst thing about it is that it actually thinks that it is saying something of significance .\nthat it actually tries to	place where happy alcoholics and cheerful drug addicts aren\'t allowed to smoke , drink or watch tv after 11 .\nit\'s simply too light , simplified and unnecessary sweet to be taken seriously .\nbut the worst thing about it\'s that it actually thinks that it\'s saying something of significance .\nthat it actually tries to
simple it is !\nin fact i would rather watch \' lost in space \' once again , than return to \' 28 days \' .\nas	simple it\'s !\nin fact i\'d rather watch \' lost in space \' once again , than return to \' 28 days \' .\nas
demonstrate that she is capable of more than \' speed \' .\nshe handles her part with a surprising	demonstrate that she\'s capable of more than \' speed \' .\nshe handles her part with a surprising

Figure 11: Contraction perturbations which caused the sentiment score for a movie review to increase by 0.97. Unchanged portions of the movie review are omitted. All changes are shown in red. Two of the changes (highlighted in yellow) distort the meaning of the text.

Original	Altered
"hilarious , ultra - low budget comedy from film school dropout kevin smith chronicles a day in the life of two convenience store slackers (brian o'halloran and jeff anderson) .\nthey spend most of their day ignoring customers while discussing everything from fellatio to self - fulfillment .\nthe premise is strictly sitcom and the photography is grainy as all get - out , but you could spend ten times the film 's budget (a reported $ 27 , 000) and still not get dialogue\nhalf as good as this .\noriginally rated nc-17 for language . not recommended for viewers with sensitive ears ."	"hilarious , ultra - low budgetc omedy from film school dropout kevin smith chrnoicles a day ni the life of two convenience sotre slackers (brian o'halloran and jeff anderson) .\nhtey spend most of their dya ignoring customers while discussing everythnig from fellatio to self - fulfillment .\nthe premis eis strictly sitcom and the photography is grainy as all get - out , but you coul dspend ten times the film 's budget (a reported $ 27 , 000) and still not get dialogue\nhal fas good as this .\noriginally rated nc-17 for language . not recommended for viewers with sensitive ears ."

Figure 12: Typo perturbations which caused the sentiment score for a movie review to increase by 0.97. Unchanged portions of the movie review are omitted. All changes are shown in red. Two of the changes (highlighted in yellow) obscure the sentiment of one of the most positive sentences in the review.

Classifier Robustness Study			
	Contractions	**Typos**	**Name changes**
Average score disturbance magnitude	1.590E-02	2.109E-02	1.012E-03
Average score disturbance	-1.256E-03	-8.524E-03	7.866E-04
Standard deviation	9.388E-02	9.658E-02	4.881E-03
Most negative score disturbance	-6.558E-01	-8.813E-01	-3.184E-03
Most positive score disturbance	9.710E-01	7.506E-01	3.949E-02

Table 6: Changes in positivity score due to perturbations in contractions, typos, and name changes.

Sensitivity Analysis	
Seed	**F1**
1234	0.86
1235	0.88
1236	0.87
1237	0.87
1238	0.87

Table 7: Comparison of F1 across different seeds for a LimitedInk identifier trained on the Movies dataset with $k = 0.5$ and the same hyperparameters as the paper.

authors' models is difficult, as the authors' most similar dataset to e-XNLI, FEVER, is easier to learn and has been more thoroughly verified to be high-quality. Attempting to compare our multilingual model to the authors' models regardless, we found our model performs substantially worse, as the FEVER model's F1 score is 0.90.

Our robustness study found the authors' model to be robust to noise in the input data. We measured the displacement of the model's score and confidence as a function of contraction changes, typo changes, and name changes. We found the standard deviation for all three types of changes to be low (on the order of 1E-01 for contractions and typos, and on the order of 1E-03 for names). However, we found that contraction changes and typos occasionally produce wide outlier changes in the model's label and confidence, and can be enough to sway a confident decision to an equally confident decision of the opposite label.

Future work could close the performance gap between multilingual LimitedInk models and the original LimitedInk models trained by the authors, and allow for easier comparison between the two. Resorting to e-XNLI as our dataset hampered our results compared to the authors', as e-XNLI has not been validated to the same level of rigor as FEVER, nor has it been human-annotated (Zaman and Belinkov, 2022; DeYoung et al., 2019). Additionally, e-XNLI's classification problem contains three classes instead of FEVER's two, which places e-XNLI models at a disadvantage. To enable easier comparison, a multilingual version of the FEVER dataset could be created in which annotations are done by hand.

References

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 9539–9549. Curran Associates, Inc.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *NAACL*.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. *arXiv preprint arXiv:1809.05053*.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. 2019. Eraser: A benchmark to evaluate rationalized nlp models. *arXiv preprint arXiv:1911.03429*.

Jay DeYoung, Eric Lehman, Benjamin Nye, Iain Marshall, and Byron C. Wallace. 2020. Evidence inference 2.0: More data, better models. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 123–132, Online. Association for Computational Linguistics.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface:a challenge set for reading comprehension over multiple sentences. In *NAACL*.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. *arXiv preprint arXiv:1606.04155*.

Cameron Martin. 2020. Facial recognition in law enforcement. *Seattle J. Soc. Just.*, 19:309.

Edoardo Mosca, Daryna Dementieva, Tohid Ebrahim Ajdari, Maximilian Kummeth, Kirill Gringauz, and Georg Groh. 2023. Ifan: An explainability-focused interaction framework for humans and nlp models. *arXiv preprint arXiv:2303.03124*.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of nlp models with checklist. In *Association for Computational Linguistics (ACL)*.

Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Hua Shen, Tongshuang Wu, Wenbo Guo, and Ting-Hao'Kenneth' Huang. 2022. Are shortest rationales the best explanations for human understanding? *arXiv preprint arXiv:2203.08788*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a Large-scale Dataset for Fact Extraction and VERification. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 809–819.

Keyon Vafa, Yuntian Deng, David M Blei, and Alexander M Rush. 2021. Rationales for sequential predictions. *arXiv preprint arXiv:2109.06387*.

Giulia Vilone and Luca Longo. 2021. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, 76:89–106.

Omar Zaidan, Jason Eisner, and Christine Piatko. 2007. Using "annotator rationales" to improve machine learning for text categorization. In *Proceedings of the conference of the North American chapter of the Association for Computational Linguistics (NAACL)*, pages 260–267.

Kerem Zaman and Yonatan Belinkov. 2022. A multilingual perspective towards the evaluation of attribution methods in natural language inference. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.