PromotionGo at LeWiDi-2025: Enhancing Multilingual Irony Detection with Data-Augmented Ensembles and L1 Loss

Ziyi Huang[1], **N. R. Abeynayake**[2], **Xia Cui**[2]

[1]Hubei University, Wuhan, China.
ziyihuang@hubu.edu.cn
[2]Manchester Metropolitan University, Manchester, UK.
{n.abeynayake, x.cui}@mmu.ac.uk

Abstract

This paper presents our system for the Learning with Disagreements (LeWiDi-2025) shared task (Leonardelli et al., 2025), which targets the challenges of interpretative variation in multilingual irony detection. We introduce a unified framework that models annotator disagreement through soft-label prediction, multilingual adaptation and robustness-oriented training. Our approach integrates tailored data augmentation strategies (i.e., lexical swaps, prompt-based reformulation and back-translation) with an ensemble learning scheme to enhance sensitivity to contextual and cultural nuances. To better align predictions with human-annotated probability distributions, we compare multiple loss functions, including cross-entropy, Kullback–Leibler divergence and L1 loss, the latter showing the strongest compatibility with the Average Manhattan Distance evaluation metric. Comprehensive ablation studies reveal that data augmentation and ensemble learning consistently improve performance across languages, with their combination delivering the largest gains. The results demonstrate the effectiveness of combining augmentation diversity, metric-compatible optimisation and ensemble aggregation for tackling interpretative variation in multilingual irony detection.

1 Introduction

Irony is a complex linguistic phenomenon in which the intended meaning of an utterance diverges from, or even contradicts, its literal expression. It often relies on contextual incongruity, implicit stance, or shared background knowledge, making it highly dependent on both linguistic and pragmatic cues. This complexity renders irony detection a particularly challenging task for computational systems, especially when extended to multilingual and multicultural contexts where the expression and interpretation of irony may vary substantially.

In such settings, human annotators frequently disagree on whether a given utterance is ironic. This disagreement stems not only from the inherent ambiguity of language but also from differences in cultural norms, humour styles, and pragmatic expectations. The *MultiPiCo* (Multilingual Perspectivist Irony Corpus, MP) (Casola et al., 2024), used in the LeWiDi-2025 shared task (Leonardelli et al., 2025), explicitly captures this variability by providing *soft labels* (i.e., empirical distributions over annotator judgments) rather than single hard labels. Modelling these distributions requires systems capable of representing annotation uncertainty and preserving distributional information in both training and inference phases.

We address this challenge by adopting a *perspectivist* framing of irony detection that emphasises multilingual generalisation and probabilistic supervision. Our system is built upon a multilingual transformer, such as `XLM-R` (Conneau et al., 2020), and incorporates several task-specific strategies: (1) a document representation pipeline that encodes post–reply pairs to preserve conversational context; (2) three targeted data augmentation methods to increase data diversity while maintaining semantic fidelity: *swap*, *prompt* and *translation*; (3) an ensemble training scheme to improve robustness and reduce variance; and (4) the use of an L1 loss function to directly optimise for the task's evaluation metric, *average Manhattan Distance*, which better reflects annotator agreement patterns than conventional cross-entropy loss.

Our contributions are as follows:

- A multilingual irony detection system that models soft labels using a transformer-based architecture aligned with human annotation distributions;
- Novel data augmentation techniques tailored to multilingual context-dependent irony detection;
- An ensemble-based training strategy that improves prediction stability under consistent modelling assumptions;

Proceedings of the The 4th Workshop on Perspectivist Approaches to NLP (NLPerspectivists 2025), pages 242–248
November 8, 2025 ©2025 Association for Computational Linguistics

- Empirical evidence supporting the use of Manhattan Distance as both an evaluation and optimisation target for soft-label learning.

Our system is designed to be robust, interpretable and adaptable across languages and cultural contexts, offering insights for future work on perspectivist approaches to subjectivity-driven language understanding tasks.

The source code for this paper is publicly available on GitHub[1].

2 Related Work

Irony and Sarcasm Detection. Detecting irony and sarcasm in text has been a long-standing challenge in computational linguistics due to its reliance on implicit meaning, context, and cultural cues. Early approaches relied on handcrafted features such as sentiment contrast or punctuation patterns (Davidov et al., 2010; Reyes et al., 2013). With the rise of deep learning, more robust methods using recurrent networks and attention mechanisms were introduced (Ghosh and Veale, 2016; Tay et al., 2018). Recent work has explored context-aware transformers, modelling not just the utterance but also conversational history or speaker intent (Bamman and Smith, 2021). While effective in monolingual settings, extending irony detection to multilingual and multicultural contexts remains an open problem, especially under limited annotated data.

Soft Labels and Annotator Disagreement. Standard supervised learning assumes a single ground truth label per instance, but tasks involving subjectivity, such as irony detection, frequently involve disagreement among annotators. This has motivated soft-label learning approaches that model the label distribution rather than a hard aggregated majority vote (Pavlick and Kwiatkowski, 2019). Soft supervision helps systems reflect uncertainty and align more closely with human perception. Galstyan and Cohen (2008) and Rizzi et al. (2024) provide comprehensive analyses of training objectives under soft labels, highlighting the inadequacy of cross-entropy loss and advocating for distance-based losses such as Manhattan Distance. These insights directly inform our use of L1 loss in both training and evaluation.

Multilingual Modelling and Data Augmentation. Multilingual pretraining has significantly advanced NLP systems' ability to generalise across languages. Models such as `mBERT` (multilingual BERT) and `XLM-R` (XLM-RoBERTa) have shown strong performance in zero-shot and few-shot cross-lingual transfer (Pires et al., 2019; Conneau et al., 2020). In tasks like irony detection, where training data may be imbalanced across languages, data augmentation becomes especially valuable. Prior work has applied back-translation (Sennrich et al., 2016), prompt-based reformulations (Bao et al., 2020) and contextual rewrites to enhance diversity. In line with these, we adopt a multilingual data augmentation framework that includes swapping discourse segments, prompt injection, and LLM-based translation to increase robustness across languages and cultural contexts.

3 System Overview

We consider the irony detection problem as a binary classification task. Given a set of dialogues D composed of post-reply pairs $(x^{(i)}_{\text{post}}, x^{(i)}_{\text{reply}})$, $\mathcal{D} = \{(x^{(i)}_{\text{post}}, x^{(i)}_{\text{reply}})\}_{i=1}^{N}$, N is the total number of instances in $\mathcal{D}$. Each instance is annotated with a probability distribution over labels $\mathbf{y}^{(i)} \in [0,1]^n$, where n is the number of annotators and $\sum_{j=1}^{n} y_j^{(i)} = 1$. The task is to train a model f_θ that maps each input pair to a predicted soft label distribution $\hat{\mathbf{y}}^{(i)} = f_\theta(x^{(i)})$, such that the average Manhattan Distance between predictions and target distributions is minimised.

3.1 Document Representation

To model the pragmatic and contextual signals that characterise irony, we use `xlm-roberta-base`[2], a multilingual transformer pretrained on 100+ languages. Each instance is represented as a concatenation of the post and its corresponding reply. Tokenisation is handled using the official Hugging Face tokeniser, preserving the consistency of subword units with the model's pretraining.

Let x = `[post] + [SEP] + [reply]` denote the tokenised input string. This is encoded into contextualised embeddings by the transformer encoder and passed through a linear projection followed by a softmax to produce the output distribution $\hat{\mathbf{y}}$.

3.2 Dataset Preprocessing

We use the official training and development splits provided by the shared task organisers. Each instance consists of a "*post*", a "*reply*", and a soft

[1] https://github.com/YhzyY/LeWiDi2025

[2] https://huggingface.co/FacebookAI/xlm-roberta-base

label distribution aggregated from multiple annotators. Instances are preprocessed and wrapped into a custom PyTorch dataset class, `MPDataset`, which encodes each post-reply pair jointly. This allows the model to account for discourse-level semantics essential for detecting irony.

3.3 Data Augmentation

To enhance robustness and reduce overfitting, we apply three task-specific data augmentation strategies, forming an augmented training set:

$$\mathcal{D}_{\text{train}} = \mathcal{D} \cup \mathcal{A}(\mathcal{D}) \tag{1}$$

where $\mathcal{A}$ denotes the augmentation pipeline. The following methods are used:

Swap. We reverse the order of the post and reply in each input sequence. This exposes the model to discourse variation and helps it focus on content and tone rather than fixed positional patterns.

Prompt-based Reformulation (Prompt). We prepend an instruction-style prompt to the input:

> "Given the following post: [post] and reply: [reply], determine whether irony can be detected."

This method conditions the model on the task and improves generalisation, particularly in multilingual settings where implicit task signals vary.

Translation. Using `gpt-3.5-turbo`[3], we translate the original input into the nine target languages (Arabic, Dutch, English, French, German, Hindi, Italian, Portuguese and Spanish). Each translated version is treated as an augmented instance, inheriting the original soft label. This expands the dataset 9-fold and promotes cross-lingual robustness. The prompt ensures consistency and tone preservation:

> "Translate its 'text' part into 9 languages: Arabic, Dutch, English, French, German, Hindi, Italian, Portuguese and Spanish. Pay attention: the translation should preserve the ironic tone in the original dialogues."

3.4 *K*-fold Ensemble Strategy

To further increase robustness, we use an ensemble training setup. The dataset is randomly shuffled and split into K equally sized subsets $\{\mathcal{D}_1, \ldots, \mathcal{D}_K\}$. For each subset, we train an independent model $f_\theta^{(k)}$. The final prediction for an instance is the unweighted average of all K outputs:

$$\hat{\mathbf{y}} = \frac{1}{K} \sum_{k=1}^{K} f_\theta^{(k)}(x) \tag{2}$$

This approach reduces variance and helps the system better handle ambiguity and soft supervision (Lakshminarayanan et al., 2017).

3.5 Training and Optimisation

We use Hugging Face's `Trainer` to train the model with a standard configuration (batch size, learning rate, epochs) tuned empirically. All models are fine-tuned on the task-specific data using the L1 loss.

Assume N denote the number of training instances and n for the number of classes, let $\hat{\mathbf{y}}^{(i)} = f_\theta(x^{(i)})$ be the predicted distribution and $\mathbf{y}^{(i)}$ the target distribution. Following Rizzi et al. (2024), we use an evaluation metric between these two distributions, *Average Manhattan Distance* (AvgMD), defined as:

$$\text{AvgMD} = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{n} \left| \hat{y}_j^{(i)} - y_j^{(i)} \right| \tag{3}$$

To align the optimisation with this metric, we train the model using the *L1 Loss*:

$$\mathcal{L}(\hat{\mathbf{y}}, \mathbf{y}) = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{n} \left| \hat{y}_j^{(i)} - y_j^{(i)} \right| \tag{4}$$

This loss provides a more faithful learning signal than cross-entropy in soft-label scenarios, especially for modelling disagreement among annotators (Rizzi et al., 2024).

4 Dataset and Experimental Setup

We use the *MultiPiCo* (MP) corpus (Casola et al., 2024), a multilingual dataset comprising short post–reply exchanges collected from Twitter and Reddit. Each reply is annotated in the context of the preceding post by approximately five crowd workers. Annotators label whether the reply is ironic, resulting in a binary classification task. Instead of collapsing annotations into hard labels, the corpus provides soft labels (i.e., distributions over the two classes) to preserve inter-annotator disagreement and enable models to learn from nuanced and perspectivist supervision.

[3] https://platform.openai.com/docs/models/gpt-3.5-turbo

The MP corpus spans nine languages: Arabic, Dutch, English, French, German, Hindi, Italian, Portuguese and Spanish. This makes it particularly well-suited for evaluating systems in multilingual and cross-cultural settings. The distribution of training, development, and test instances for each language is shown in Table 1.

Table 1: Number of training, development, and test instances per language in the MP corpus.

Language	#Train	#Dev	#Test
Arabic	1,399	363	419
Dutch	637	147	216
English	1,920	489	590
French	1,137	276	347
German	1,513	358	504
Hindi	505	132	149
Italian	646	159	195
Portuguese	1,286	325	383
Spanish	2,974	756	953

We strictly adhered to the official data splits provided by the shared task organisers for training, development and evaluation. No external resources or additional data were used at any stage. This ensures that our system is evaluated under the same constraints as other submissions, and that performance comparisons remain valid.

We use the hyperparameters provided in the `transformers` library for the `xlm-roberta-base` model. Training is performed using the Adam optimiser with a linear learning rate schedule and early stopping based on validation loss. All preprocessing, tokenisation and batching are handled using the Hugging Face framework.

5 Results

Our system achieved competitive performance in the LeWiDi-2025 shared task, highlighting the effectiveness of its multilingual architecture and soft-label modelling approach. In the following subsections, we present detailed evaluations, including augmentation and loss function ablations, ensemble comparison, as well as cross-lingual analyses. All the following experiments are using the same training arguments, the only differences between them are the data augmentation methods, loss function, and ensemble method. Due to the absence of publicly released gold labels for the test set, most validation results were obtained via the Codabench evaluation platform. Language-wise analyses (Section 5.4) could not be conducted on the hidden test set; for these, we report results on the development set instead.

5.1 Effect of Data Augmentation

We evaluated the impact of each proposed augmentation strategy (*swap*, *prompt* and *translation*), as well as their combinations. Table 2 reports the average Manhattan Distance (AvgMD; lower is better) for systems trained under different augmentation settings. All experiments in this section use the L1 loss function without ensemble methods, ensuring that the effects of data augmentation are measured in isolation. Results show that combining augmentation methods consistently outperforms individual ones, with the best performance obtained by using all three techniques or the *swap+translation* pairing. Among single strategies, *translation* yields the largest gain over the baseline, while *swap* and *prompt* produces only marginal improvements. This suggests that semantic-preserving transformations (e.g., translation) contribute more than structural manipulations when modelling irony across languages.

Table 2: AvgMD for different data augmentation configurations.

Augmentation	AvgMD
All Combined	**0.407**
Swap + Translation	**0.407**
Prompt + Translation	0.410
Translation	0.411
Swap + Prompt	0.428
No Augmentation	0.451
Prompt	0.464
Swap	0.473

5.2 Effect of Loss Function

We next compared three loss functions, cross-entropy (CE), L1 and KL divergence, on the baseline system without data augmentation and ensemble. (Figure 1). L1 loss achieved the lowest AvgMD, outperforming CE by 0.034 absolute points, confirming its suitability for aligning predictions with human-annotated distributions in the presence of label uncertainty. KL divergence performed substantially worse, likely due to over-sensitivity to distribution mismatches in low-resource or highly ambiguous cases. These results motivate our final system design, which integrates L1 loss with combined data augmentation

to maximise robustness and cross-lingual generalisation.

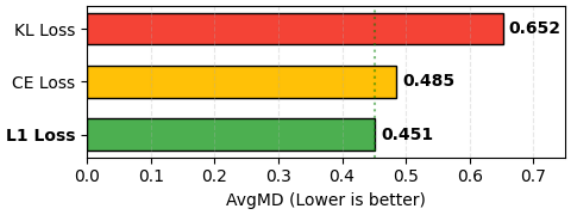

Figure 1: Performance using different loss functions (no augmentation).

5.3 Effect of Ensemble Training

We assess the impact of the ensemble approach using the L1 loss function in combination with different data augmentation settings. The ensemble is constructed by randomly shuffling the training set and partitioning it into five equally sized subsets ($K = 5$), each used to train an independent model. During inference, all models produce soft-label predictions on the test set, which are then averaged to form the final output. This averaging mitigates variance, reduces prediction noise, and improves robustness, particularly in the presence of noisy or ambiguous labels such as those found in irony detection.

Table 3 compares the ensemble results with their single-model counterparts under identical loss and augmentation settings. In all configurations, the ensemble achieves a lower AvgMD. The largest gain occurs with *swap* augmentation, which achieves the highest AvgMD. When no ensemble was applied, its AvgMD drops by 0.049 (from 0.473 to 0.424). Even the smallest gain, observed with *swap+prompt* augmentation, still yields a reduction of 0.008. These consistent improvements highlight the ensemble's ability to capture complementary decision patterns from models trained on different data partitions, leading to more stable and accurate soft-label estimations.

Table 3: AvgMD improvements from applying the ensemble method under different data augmentation settings.

Augmentation	w/o Ensemble	w/ Ensemble	ΔAvgMD
All Combined	**0.407**	0.396	-0.011
Swap + Translation	**0.407**	0.396	-0.011
Prompt + Translation	0.410	0.394	-0.016
Translation	0.411	**0.392**	-0.019
No Augmentation	0.451	0.417	-0.034
Swap + Prompt	0.428	0.420	-0.008
Prompt	0.464	0.428	-0.036
Swap	0.473	0.424	**-0.049**

5.4 Cross-lingual Performance

We assess the system's ability to generalise across the nine target languages using the best-performing configuration for single model(L1 loss with all combined data augmentations). Figure 2 reports AvgMD per language. Performance is better for English and Dutch (AvgMD < 0.4), which have relatively larger training sets and higher lexical similarity to other European languages in the corpus. Spanish and Arabic also perform well despite linguistic differences, suggesting the model effectively leverages cross-lingual transfer. In contrast, Portuguese, French, German and Italian exhibit higher AvgMD values, indicating reduced agreement with human annotations. These discrepancies may stem from smaller data sizes, domain-specific lexical variation, or cultural differences in the expression of irony. Overall, results highlight both the promise and the unevenness of cross-lingual generalisation in perspectivist irony detection.

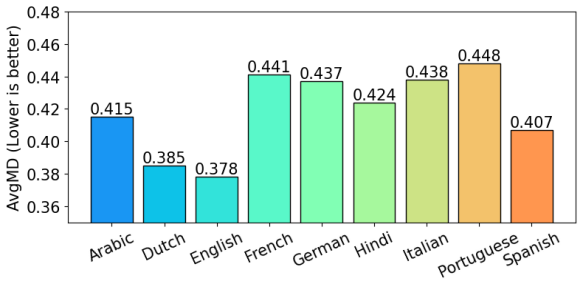

Figure 2: AvgMD per language (lower is better) using L1 loss and all augmentations.

5.5 Overall Result

Under the L1 loss setting without any data augmentation or ensemble, the baseline system achieves an AvgMD of 0.451. As shown in the augmentation ablation study, incorporating data augmentation yields consistent performance improvements, with the best-performing augmentation strategy being the one combined with swap, prompt and translation, which attains an AvgMD of 0.407. This confirms that certain augmentation combinations, particularly those leveraging complementary linguistic variations, can effectively reduce the divergence from human soft labels.

When further integrating the ensemble method, the results demonstrate an even more pronounced improvement. In all cases, the ensemble consistently reduces AvgMD compared to their non-ensemble counterparts, as discussed in the ensemble ablation analysis. The optimal configuration is

achieved using the translation augmentation with ensemble, which delivers the lowest AvgMD of 0.392 across all experiments. This outcome aligns with our earlier explanation that ensembles, by aggregating predictions from models trained on diverse data subsets, capture richer and more complementary decision patterns, thereby achieving superior alignment with annotator distributions.

6 Conclusions

We presented a unified system for the LeWiDi-2025 shared task that addresses the challenges of annotator disagreement in multilingual irony detection. Our approach integrates complementary data augmentation strategies, loss functions tailored to the evaluation metric, and an ensemble framework to improve alignment with human-annotated soft labels. Experiments demonstrate that the combination of augmentation and ensemble learning yields substantial reductions in Average Manhattan Distance over strong baselines, with L1 loss proving particularly effective for soft-label prediction under this metric. These findings underscore the value of jointly leveraging data diversity, metric-compatible optimisation, and model aggregation to better capture interpretative variation in multilingual and culturally nuanced NLP tasks. Future work will explore more context-aware augmentation methods and adaptive ensemble schemes to further enhance cross-lingual robustness.

Limitations

While our system demonstrates strong performance in reducing divergence from annotator distributions, several limitations remain. First, our data augmentation strategies, though effective, are primarily heuristic and may not fully capture the full range of linguistic or cultural variability present in real-world irony. Second, the ensemble approach, while improving performance, increases computational cost during both training and inference, which may limit scalability in resource-constrained settings. Third, our experiments focus on multilingual but not truly cross-lingual transfer scenarios; future work should investigate whether the proposed framework generalizes effectively to unseen languages or domains. Finally, although L1 loss proved advantageous for the given metric, its effectiveness for other evaluation criteria remains to be systematically assessed.

Ethical Statements

This study builds upon publicly released datasets provided by the competition organizers, which include multilingual social media content originally collected from platforms such as X and Reddit. All data used are anonymised and intended for research purposes only. We do not introduce any additional user-generated content or external datasets beyond the official competition resources.

Given the subjective nature of irony and the perspectivist framing of this task, we acknowledge the potential for cultural and linguistic biases to influence both human annotations and model predictions. Our system is trained to align with aggregated soft labels that reflect annotator disagreement, however, it may still reflect dominant cultural interpretations embedded in the training data. Additionally, the use of multilingual large language models such as xlm-roberta-base and gpt-3.5-turbo may introduce biases inherited from their pretraining corpora. We encourage careful downstream use of such models and stress the importance of transparency, cultural sensitivity, and critical evaluation when deploying irony detection systems in real-world applications.

References

David Bamman and Noah Smith. 2021. Contextualized sarcasm detection on twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, 9(1):574–577.

Yujia Bao, Menghua Wu, Shiyu Chang, and Regina Barzilay. 2020. Few-shot text classification with distributional signatures. In *International Conference on Learning Representations*.

Silvia Casola, Simona Frenda, Soda Marem Lo, Erhan Sezerer, Antonio Uva, Valerio Basile, Cristina Bosco, Alessandro Pedrani, Chiara Rubagotti, Viviana Patti, and Davide Bernardi. 2024. MultiPICo: Multilingual perspectivist irony corpus. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16008–16021, Bangkok, Thailand. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Semi-supervised recognition of sarcasm in Twitter and Amazon. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pages 107–116, Uppsala, Sweden. Association for Computational Linguistics.

Aram Galstyan and Paul R. Cohen. 2008. Empirical comparison of "hard" and "soft" label propagation for relational classification. In *Inductive Logic Programming*, pages 98–111, Berlin, Heidelberg. Springer Berlin Heidelberg.

Aniruddha Ghosh and Tony Veale. 2016. Fracking sarcasm using neural network. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 161–169, San Diego, California. Association for Computational Linguistics.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6405–6416, Red Hook, NY, USA. Curran Associates Inc.

Elisa Leonardelli, Silvia Casola, Siyao Peng, Giulia Rizzi, Valerio Basile, Elisabetta Fersini, Diego Frassinelli, Hyewon Jang, Maja Pavlovic, Barbara Plank, and Massimo Poesio. 2025. Lewidi-2025 at nlperspectives: third edition of the learning with disagreements shared task. In *Proceedings of the 4th Workshop on Perspectivist Approaches to NLP (NLPerspectives)*. Association for Computational Linguistics.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Antonio Reyes, Paolo Rosso, and Tony Veale. 2013. A multidimensional approach for detecting irony in twitter. *Lang. Resour. Eval.*, 47(1):239–268.

Giulia Rizzi, Elisa Leonardelli, Massimo Poesio, Alexandra Uma, Maja Pavlovic, Silviu Paun, Paolo Rosso, and Elisabetta Fersini. 2024. Soft metrics for evaluation with disagreements: an assessment. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 84–94, Torino, Italia. ELRA and ICCL.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Yi Tay, Anh Tuan Luu, Siu Cheung Hui, and Jian Su. 2018. Reasoning with sarcasm by reading in-between. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1010–1020, Melbourne, Australia. Association for Computational Linguistics.