How Much Noise Can BERT Handle? Insights from Multilingual Sentence Difficulty Detection

Nouran Khallaf, Serge Sharoff
Centre for Translation, Localisation and Interpreting Studies
School of Languages, Cultures and Societies
University of Leeds, UK
{n.khallaf,s.sharoff}@leeds.ac.uk

Abstract

Noisy training data can significantly degrade the performance of language-model-based classifiers, particularly in non-topical classification tasks. In this study we designed a methodological framework to assess the impact of denoising. More specifically, we explored a range of denoising strategies for sentence-level difficulty detection, using training data derived from document-level difficulty annotations obtained through noisy crowdsourcing. Beyond monolingual settings, we also address cross-lingual transfer, where a multilingual language model is trained in one language and tested in another. We evaluate several noise reduction techniques, including Gaussian Mixture Models (GMM), Co-Teaching, Noise Transition Matrices, and Label Smoothing. Our results indicate that while BERT-based models exhibit inherent robustness to noise, incorporating explicit noise detection can further enhance performance. For our smaller dataset, GMM-based noise filtering proves particularly effective in improving prediction quality by raising the Area-Under-the-Curve score from 0.52 to 0.92, or to 0.93 when de-noising methods are combined. However, for our larger dataset, the intrinsic regularisation of pre-trained language models provides a strong baseline, with denoising methods yielding only marginal gains (from 0.92 to 0.94, while a combination of two denoising methods made no contribution). Nonetheless, removing noisy sentences (about 20% of the dataset) helps in producing a cleaner corpus with fewer infelicities. As a result we have released the largest multilingual corpus for sentence difficulty prediction: see https://github.com/Nouran-Khallaf/denoising-difficulty.

Keywords: Noise reduction; Crowdsourcing; Multilinguality; Readability; Non-topical classification

1. Introduction

Modern Natural Language Processing (NLP) methods, driven by Pre-Trained Language Models (PLMs) and Large Language Models (LLMs), have achieved impressive results across a wide range of tasks. Nevertheless, their performance remains uneven in *non-topical classification* tasks, such as predicting text genre (Rönnqvist et al., 2022; Kuzman et al., 2023), demographic properties (Kang et al., 2019), or text difficulty (North et al., 2022). Such tasks require sensitivity to subtle stylistic, structural, and syntactic cues rather than reliance on topical keywords, while LLMs struggle to capture these linguistic distinctions (Kuzman et al., 2023).

In contrast to topical classification, which benefits from the presence of explicit domain-related terms, non-topical classification demands that models identify latent stylistic features and register-level indicators (Dewdney et al., 2001). Both PLMs and LLMs can be misled by irrelevant topical content in these settings (Roussinov and Sharoff, 2023). Moreover, recent studies demonstrate that larger generative LLMs such as GPT or LLaMA do not consistently outperform smaller PLMs such as BERT on text classification tasks (Edwards and Camacho-Collados, 2024). Accordingly, our experiments focus on fine-tuning BERT-like PLMs; we do not include LLM-based classifiers due to their high computational costs and marginal benefits for this task (see Section 10 for discussion of computational and ethical considerations).

A major challenge in non-topical classification tasks lies in the scarcity of high-quality, reliable training data, especially in multilingual contexts. Existing datasets frequently suffer from annotation noise and label inconsistencies. Some of this noise stems from the variability of crowd-sourced annotations and from the use of document-level difficulty labels for sentence-level predictions. For example, a sentence taken from a Wikipedia article (the default source of more complex texts in our experiments) may in fact be linguistically simple, while some sentences from an ostensibly "simple" crowdsourced corpus may be structurally or semantically challenging because of annotation inconsistencies. This can seriously degrade classifier performance and obscure genuine cross-lingual trends.

In this study, we systematically analyse the effects of annotation noise on sentence-level difficulty classification by exploring a range of denoising techniques aimed at improving robustness under realistic data conditions. With this paper we:

1. identify effective noise reduction methods that enhance the stability and accuracy of non-topical classification;

2. evaluate intersection of data points identified as noise by several noise reduction methods;

Language	#Docs	Wikipedia			Vikidia		
		#Words	#Sentences	IQR	#Words	#Sentences	IQR
Catalan	179	396,277	16,813	(15, 29)	19,394	1,000	(13, 23)
English	2585	8,281,625	340,924	(16, 29)	424,306	22,462	(13, 22)
Spanish	3875	7,946,169	301,241	(16, 33)	607,990	27,825	(14, 26)
French	33438	46,618,143	1,945,046	(15, 29)	6,643,567	320,372	(14, 25)
Italian	3902	6,790,163	263,271	(16, 32)	537,723	25,202	(14, 26)
Russian	136	334,416	17,618	(13, 22)	10,454	604	(12, 20)

Table 1: Statistics for paired Wikipedia–Vikidia corpora. #Docs is the document count (paired Wikipedia–Vikidia articles by topic). #Words and #Sentences are totals across all documents in each subset (restricted to paired topics). IQR gives the inter-quartile range of sentence lengths (in words) in each subset.

3. evaluate the interaction between noise and cross-lingual transfer, thus providing evidence for how multilingual PLMs handle noise across languages; and
4. release a publicly available de-noised multilingual corpus for sentence-level difficulty classification, accompanied by all scripts and models for reproducible experimentation.

2. Dataset

The datasets used in our experiments were sourced from Vikidia and Wikipedia,[1] covering multiple languages, see Table 1. The simple versions have been crawled from Vikidia,[2] a website that maintains Wikipedia-style content aimed at "children and anyone seeking easy-to-read content". The range of languages in our experiments reflects the availability of languages in Vikidia and the availability of annotators for quality control. We have removed the entries marked as stubs (with little content at the moment) and collected the corresponding main Wikipedia entries for the respective languages. Therefore, the documents in our dataset address exactly the same topics. This removes topic biases, which often impact non-topical classification tasks due to unreasonable performance through just learning the keywords (Roussinov and Sharoff, 2023). In the end, we have obtained a document-level resource for text-difficulty detection. However, our aim is to develop more granular classifiers on the sentence level. This formulation introduces noise, since many sentences may be simple despite their provenance from Wikipedia.

The Inter-Quartile Range (IQR) values in Table 1 indicate that the length of most sentences in either source falls between roughly 13 and 33 words. Across all languages, Vikidia consistently features shorter sentences, typically *2–7 words shorter*. While the documents are paired between the Wikipedia and Vikidia in each language, the number of sentences for the classification task is severely skewed towards more complex examples, as Wikipedia documents are considerably longer, leading to having 15 times more complex-labelled sentences for English and 6 times more for French.

3. Methodology

All of our experiments aim at a binary classification task: predicting whether a sentence is complex, i.e., in our definition of complexity, the sentence is not suitable for inclusion in Vikidia. According to Table 1 this is the majority class, so a classifier may naturally produce a high proportion of false positives (simple sentences predicted as complex) simply because most sentences are complex and the prior pushes predictions toward that class: a degenerate classifier which only predicts the majority class will achieve about 0.93 F1-Score. A false negative means a complex sentence is predicted as simple, which is a critical error in our scenario. However, allowing false positives means the classifier unnecessarily flags simple sentences. To mitigate this, we use the Area Under the ROC Curve (AUC) as the primary evaluation metric, as it is less sensitive to class imbalance. For example, the degenerate classifier which only predicts the majority class will achieve the AUC score of 0.5, reflecting the absence of meaningful discrimination (Li, 2024).

3.1. Models and Training Setup

Our experiments use the multilingual BERT-base model (Devlin et al., 2019), as well as multilingual SBERT for sentence-level embedding representations (Reimers and Gurevych, 2019). Preliminary experiments with mBERT-large and XLM-RoBERTa (Conneau et al., 2020) revealed similar performance trends; therefore, we focus on the base architectures for clarity and consistency.

Baseline mBERT models were fine-tuned independently on the English and French datasets using

[1]The datasets from Wikipedia and Vikidia are available under the (CC BY-SA) license. Our use aligns with their intended purpose, with our modifications limited to preprocessing and data selection.

[2]https://www.vikidia.org/

the Hugging Face Transformers library, trained for up to ten epochs with early stopping. The English corpus serves as a standard benchmark, while the larger French dataset (nearly two million sentences) enables analysis of how corpus size influences model robustness to noise. For evaluating cross-lingual transfer we applied the English and French models to Catalan, Spanish, Italian, and Russian: Catalan is a low-resource language for PLM pre-training, so we expect weaker results, while Russian is distant to English and French, thus these languages provide interesting cases for assessing the multilingual transfer gap (Hu et al., 2020). Since the PLMs were pre-trained primarily on English, stronger English performance is expected. However, the French Vikidia dataset is much larger, so it allows us to test whether data scale can compensate for pre-training bias.

3.2. Noise Reduction Pipeline

We define noise as either incorrect labels or low-quality text. The incorrect labels are Vikidia sentences that are actually complex or Wikipedia sentences that are simple, as distinct from correctly labeled sentences (either simple or complex), which we seek to keep. Low-quality text concerns data-extraction artifacts, such as list-like or malformed segments, encoding issues or markup.

Therefore, we applied several noise reduction methods to identify unreliable samples in the training data with the aim of removing them to avoid annotation inconsistencies and to improve prediction accuract. After detecting noisy instances, only samples identified as clean were retained for fine-tuning, using the same hyperparameters as the baselines to ensure comparability. For evaluation, we manually cleaned the test sets (removing/repairing clear mislabels or broken segments) to ensure reliable measurement; results on the original noisy test are lower, as expected.

In addition to assessing each denoising method individually, we conducted an *intersection analysis* to identify sentences consistently flagged as noisy across multiple approaches, revealing the most recurrent noisy examples and their effect on downstream performance.

We compare five denoising methods:

Gaussian Mixture Models (GMMs). GMMs cluster high-dimensional sentence representations into two distributions corresponding to clean and noisy samples (Bishop, 2006). We experimented with representations from BERT CLS embeddings (GMM-B) and from SBERT (GMM-SB). Initially, hyperparameters—such as the number of components, covariance type, and threshold—were tuned using Optuna over 100 trials (Akiba et al., 2019), targeting maximal separation between clean and noisy clusters. Based on these trials, we selected two fixed, cross-lingual configurations to standardise experiments and reduce computational costs, so that we set the number of GMM components to 9 for both languages and models to keep the clustering consistent and stable. For SBERT, we used *full* covariance matrices because its embeddings generates more semantically meaningful embeddings using a siamese architecture and pooling strategies (Reimers and Gurevych, 2019). In contrast, BERT [CLS] embeddings are unevenly distributed in space and can lead to unstable clusters if full covariance is used. Therefore, the best model uses *tied* covariance for BERT, which shares the same shape across all clusters and improves stability (Ethayarajh, 2019). We set the noise threshold by applying kernel density estimation to the GMM noise scores and selecting the dip between peaks (Silverman, 1981). This dynamic thresholding strategy captures data-specific separation between noisy and clean points and adapts to the inherent variability of each corpus. We apply this unified GMM configuration across English and French to ensure a consistent modeling approach, enhancing reproducibility and comparability across multilingual settings.

Small-Loss Trick (ST). ST assumes that examples with high training losses are more likely to contain noise (Arpit et al., 2017; Han et al., 2018). During training, the model retains only a subset of samples with lower losses to detect data points likely to be mislabeled (Yu et al., 2019). A key hyperparameter in this method is the *loss threshold percentile*, which determines the fraction of lowest-loss examples retained at each selection step: at each epoch, losses are computed for all current examples, for example, a 75% threshold thus means that at each selection point we retain the bottom 75% of examples ranked by loss and exclude 25% with the highest loss. For the next epoch another 75% of the lowest losses from the full dataset are selected for training. After testing the range from 10% to 90% we have selected the 75% threshold for the best balance, effectively excluding uncertain data points while retaining a sufficient number of reliable instances for model training. Finally, we define *noisy* sentences as those repeatedly assigned to the excluded (high-loss) set across the five epochs (i.e., the intersection of high-loss selections over epochs).

Co-Teaching (CT). CT extends the Small-loss Trick method by training two prediction models in parallel, each selecting the lowest-loss samples from the other's batch, under the assumption that these are more likely to be correctly labeled (Han et al., 2018). This cross-filtering mechanism en-

Language	Baseline	GMM-B	GMM-SB	ST	CT	NTM	LS	Intersection CT/LS	LS/NTM	CT/NTM	CT/G-S	LS/G-S	CT/NTM/G-S
en	0.5209	0.9206	0.9211	0.8296	0.7790	0.8343	0.8116	0.9235	0.9259	0.9114	0.9096	0.9196	**0.9261**
ca	0.5099	0.7693	0.7504	0.7239	0.7494	0.7277	0.8003	0.7764	0.7558	0.7692	0.7560	0.7605	0.7525
es	0.5180	0.7753	0.7760	0.7509	0.7724	0.7407	0.7955	0.7770	0.7750	0.7806	0.7528	0.7697	0.7735
fr	0.5207	0.7736	0.7733	0.7255	0.7358	0.7262	0.6612	0.7859	0.7735	0.7760	0.7637	0.7678	0.7702
it	0.5104	0.7718	0.7699	0.7502	0.7259	0.7212	0.7974	0.7714	0.7719	0.7738	0.7628	0.7699	0.7678
ru	0.5447	0.7968	0.8015	0.7301	0.7899	0.7976	0.8475	0.7797	0.7982	0.7917	0.7756	0.8146	0.8067
Average	0.5207	0.7774	0.7742	0.7361	0.7547	0.7427	**0.7804**	0.7781	0.7749	0.7783	0.7622	0.7765	0.7741
Train (s)	19,977	25,536	13,658	22,706	36,723	24,239	30,348	74,291	62,627	75,457	58,443	52,471	89,115
# Noisy	–	3037	37070	24,525	157,979	72,128	72,678	476	14,612	31,328	16,200	7,429	3,215
# Percent	–	0.84%	10.20%	6.75%	43.47%	19.85%	20.00%	0.13%	4.02%	8.62%	4.46%	2.04%	0.88%
%noisy Wiki	–	98.58%	93.88%	100%	88.99%	94.26%	94.33%	99.58%	94.55%	89.34%	88.57%	94.45%	88.52%
%noisy Viki	–	1.42%	6.12%	0%	11.01%	5.74%	5.67%	0.42%	5.45%	10.66%	11.43%	5.55%	11.48%

(a) Trained on English

Language	Baseline	GMM-B	GMM-SB	ST	CT	NTM	LS	Intersection CT/LS	LS/NTM	CT/NTM	CT/G-S	LS/G-S	CT/NTM/G-S
fr	0.9198	0.9213	0.9193	0.9182	0.5432	0.7694	**0.9402**	0.9240	0.9265	0.9216	0.9221	0.9235	0.9238
en	0.8523	0.8585	0.8623	0.8480	0.5045	0.7530	0.8516	0.8296	0.8430	0.8344	0.8332	0.8360	0.8411
ca	0.7677	0.7660	0.7628	0.7801	0.5089	0.6836	0.7697	0.7550	0.7704	0.7722	0.7607	0.7585	0.7700
es	0.8033	0.8090	0.8044	0.8047	0.5033	0.7167	0.7991	0.7909	0.7996	0.7899	0.7898	0.7940	0.7913
it	0.8076	0.8137	0.8123	0.8127	0.5086	0.6982	0.8080	0.7940	0.8072	0.7928	0.7931	0.7972	0.7969
ru	0.7783	0.7807	0.7670	0.7587	0.4946	0.7443	0.7478	0.7456	0.7675	0.7618	0.7439	0.7533	0.7632
Average	0.8018	**0.8056**	0.8018	0.8008	0.5040	0.7192	0.7952	0.7830	0.7975	0.7902	0.7841	0.7878	0.7925
Train (s)	58,386	69,490	110,424	93,637	80,342	174,600	69,899	243,540	320,370	330,813	270,111	259,931	441,237
# Noisy	–	1,108,670	349,441	434,377	747,823	279,472	566,356	183,487	56,468	81.349	40,732	25,144	30,482
# Percent	–	48.96%	15.43%	19.18%	33.01%	12.34%	25.02%	8.10%	2.49%	3.59%	2.09%	1.11%	1.35%
%noisy Wiki	–	88.87%	86.19%	89.14%	69.08%	85.78%	86.04%	65.55%	83.83%	66.17%	70.41%	85.59%	69.56%
%noisy Viki	–	11.13%	13.81%	10.86%	30.92%	14.22%	13.96%	34.45%	16.17%	33.83%	29.59%	14.41%	30.44%

(b) Trained on French

Table 2: Comparative performance of noise reduction methods (measured via ROC-AUC) with models trained on English (top) and French (bottom). "Average" is the macro-average ROC-AUC over transfer languages (excluding the training language). "Baseline" refers to standard fine-tuning with no noise filtering. # Noisy is the count of sentences detected as noisy by the respective method. G-B stands for GMM-Bert, G-S stands for GMM-SBert. The underlined values are not significantly different from the best ones.

sures that each model learns from cleaner examples, reducing the influence of noisy data. The key hyper-parameter in CT is the *forget rate*, which determines the fraction of high-loss samples to remove from training at each training step. Dynamic Loss Thresholding (DLT) enables the model to adapt gradually, preventing premature discarding of difficult samples and reducing excessive data loss in early training. This outperforms static thresholding by enhancing robustness against noisy labels (Yang et al., 2024). In our implementation, we used a dynamic forget rate schedule, increasing linearly from 0.0 to 0.3 across epochs:

$$r_t = r_{\min} + (r_{\max} - r_{\min})\frac{t}{T}, \tag{1}$$

where r_t is the forget rate at epoch t, $r_{min} = 0.0$ is the initial forget rate, $r_{max} = 0.3$ is the maximum forget rate, and T is the total number of epochs. At each training step, the forget rate determines the fraction of high-loss samples to discard within each mini-batch. In the early epochs, all samples are used ($r_t = 0$), while in later epochs, up to 30% of the higher-loss samples are discarded ($r_t \to 0.3$).

Noise Transition Matrix (NTM). NTMs explicitly model the probability of label corruption under class-dependent noise by estimating each matrix element $T_{ij} = P(\tilde{y} = j \mid y = i)$, where y and $\tilde{y}$ are the true and observed labels, respectively (Patrini et al., 2017). Unlike CT and ST, which discard noisy samples based on loss, noise transition-based methods retain all data and adjust model predictions to compensate for systematic label noise. Using data identified as noisy by GMM-SB, we estimated the matrix T_{ij} from the empirical confusion matrix and computed its inverse T_{inv}. During training, predicted probabilities $\hat{P}$ are adjusted as:

$$\hat{P}_{\text{adjusted}} = \hat{P} \cdot T_{\text{inv}} \tag{2}$$

followed by cross-entropy loss computation. This procedure retains all samples while correcting for systematic label noise.

Label Smoothing (LS). Label smoothing regularises the model by softening categorical targets. Instead of assigning a probability of 1 to the correct class and 0 to all others, LS redistributes a

small fraction of this probability across all classes, helping the model handle mislabeled or ambiguous data (Szegedy et al., 2016; Müller et al., 2019). This reduces overconfidence and has shown to improve generalisation on noisy or imbalanced datasets (Lukasik et al., 2020; Ren et al., 2024). The smoothed label is computed as:

$$y_{\text{smooth}} = (1 - \epsilon) \times y + \frac{\epsilon}{k}, \tag{3}$$

where ϵ is the smoothing factor and k is the number of classes. The noise-rejection threshold τ was tested in the range $0.50 \leq \tau \leq 0.70$, incremented by 0.05, while the smoothing factor was varied between $0.0 \leq \epsilon \leq 0.2$, also incremented by 0.05.

The results indicate that higher smoothing factors ($\epsilon \geq 0.15$) excessively redistributed probabilities, leading to degraded predictions due to increased uncertainty in class assignments. Conversely, the absence of smoothing ($\epsilon = 0.0$) resulted in overconfident predictions, increasing the risk of misclassification. A moderate smoothing factor of $\epsilon = 0.1$ provided the optimal balance, enhancing generalisation while maintaining well-calibrated predictions. Similarly, higher noise-rejection thresholds ($\tau = 0.70$) yielded superior performance by effectively filtering uncertain predictions.

4. Noise Detection and Reduction

4.1. Impact de-noising on performance

Our evaluation focuses on the impact of denoising techniques on sentence-level classification robustness, particularly in cross-lingual settings. Our goal is to assess how effectively different noise-reduction strategies enhance model stability and transfer performance across languages, using English and French as the primary training datasets. The experiments use the sentence-level datasets described in Table 1.

When trained on English (Table 2a), the baseline model achieved a very poor AUC value, which has been substantially improved by all de-noising methods. Both Gaussian Mixture Models (on the BERT embeddings and on the Sentence-BERT embeddings) achieved higher and more stable discrimination. As expected, cross-lingual transfer decreases the performance. However, de-noising can keep it to acceptable values with the best method of GMM-Bert and GMM-SBert and Label Smoothing.

The number of sentences flagged as noisy varies widely across methods: from 3,037 for GMM-Bert (less than 1% of the original dataset) to over 157,000 for CT (43%), which leads to a substantially less accurate classifier trained on CT-filtered data. This indicates that the de-noising methods have different sensitivities.

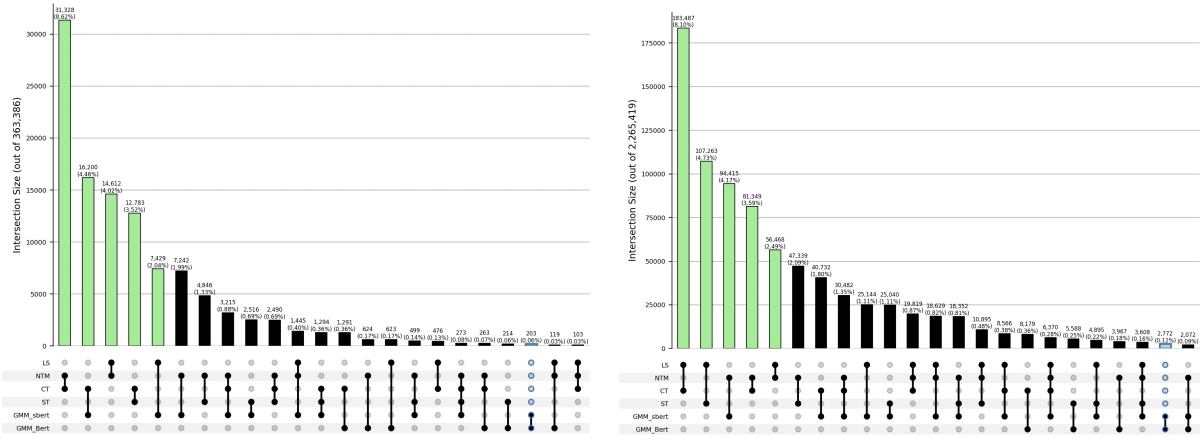
Beyond the individual de-noising methods, we experimented with their intersection, i.e. by removing the data detected as noisy by two classifiers. The resulting intersection identifies smaller but more reliable subsets of noise. The best-performing intersections—CT/LS and LS/NTM (0.93 for English, 0.78 for the cross-lingual transfer)–demonstrate that integrating even relatively weak denoising methods can yield fairly balanced outcomes. However, the computational costs are more significant for computing the intersection; for example, computing both CT and GMM-SBert on the English dataset using a single L40S GPU adds 43,395 seconds to the time of fine-tuning the baseline model (19,977 sec).

In contrast to English, models trained on the much larger French dataset (Table 2 b) exhibited a different trend. The baseline already achieved the level of performance comparable with the best English model after de-noising. De-noising for French improves over the baseline model only marginally, namely from 0.9189 to 0.9402 for the single method (LS) or to 0.9265 for the intersection (LS/NTM), the same de-noising techniques which worked well for English. This indicates diminishing returns from denoising when training data is abundant. The close linguistic proximity between French and the Romance test languages (Catalan, Spanish, Italian) helps in achieving better cross-lingual transfer from French in comparison to Russian, which has seen no benefits from the bigger French dataset.

From a computational efficiency perspective, both training and de-noising become more expensive for French. While the classification models provide slightly better performance on French and in cross-lingual transfer, the training time increased, both for the baseline model (58,386 sec) and for the best de-noising (69,899 sec). At the same time, the de-noising step helps with removing less relevant examples, so a combination of de-noising with training is less that the sum of de-noising and baseline training.

4.2. Intersection analysis

We also investigated the agreement across the de-noising techniques through their intersection (Figure 1), which shows the size of intersections, i.e., sentences jointly flagged by the methods indicated by the connected dots. For the English-trained model 1a, noise detection is dominated by combinations involving CT, particularly with NTM (8.62%), GMM-SBERT (4.46%), and ST (3.52%), indicating that CT consistently overlaps with other methods in identifying noisy samples. The LS–NTM intersection (4.02%) also contributes substantially, suggesting that loss-based and transition-based modelling approaches converge with probabilistic clustering methods on similar subsets of noisy data.



(a) Noisy data intersections for English.

(b) Noisy data intersections for French.

Figure 1: Comparison of noisy data intersections across denoising techniques for English and French. Each vertical bar represents the proportion of sentences identified as noisy by the combination of methods in the connected dots below. The small bar chart at the bottom left indicates the contribution of each method to the total pool of sentences identified as noisy.

These high-overlap regions represent the most reliable noise detections, as multiple independent techniques agree on the same examples.

The analysis shows a clear concentration of cross-method agreement in noise around CT- and NTM-based methods, with GMM-SBert appearing consistently across several high-agreement intersections and contributing to multiple overlapping regions, alongside smaller yet meaningful overlaps involving LS and ST.

For the French-trained model 1b, noise detection is likewise dominated by combinations involving CT, particularly with LS (8.10%), the ST (4.73%), and NTM (3.59%), indicating that CT consistently overlaps with other methods in identifying difficult or mislabeled samples. The NTM–GMM-SB intersection (4.17%) also contributes substantially, suggesting that probabilistic clustering and loss-based modelling converge on similar subsets of noisy data. These high-overlap regions represent the most reliable noise detections, as multiple independent techniques agree on the same examples. The analysis shows a clear concentration of noise consensus around LS and CT, with additional meaningful overlaps involving NTM and GMM-SB. This pattern suggests that CT and LS form a robust backbone for noise identification in the large-scale French dataset, while NTM and GMM-based methods provide complementary filtering signals. Overall, these intersections confirm that multi-method agreement serves as a strong indicator of genuine annotation noise, improving both reliability and interpretability of the denoising process.

Category	English	French
Total sentences	230	237
Not noisy	92	78
Noisy	138	159

Table 3: Accuracy analysis at the intersection

5. Manual Error Analysis

To validate our denoising pipeline, we analyse errors in two stages: (i) selecting a subset of sentences flagged jointly by CT, NTM, and GMM-SBert to check whether there are reasons to treat them as noisy, and (ii) reviewing the noisy sentences to develop a noise taxonomy. Guided by Table 2, we select the three-way intersection (CT∩NTM∩GMM-SB) as a high-confidence subset of noisy predictions. This intersection ranks among the strongest-performing intersections in both English- and French-trained settings. This subset captures agreement between three complementary signals: loss-based filtering (*CT*), explicit modelling of label transitions (*NTM*), and embedding-based clustering (*GMM-SB*). As a result, it is less likely to include sentences that are simply difficult but clean, because an instance must be supported by all three mechanisms to be retained. Moreover, this intersection represents the strongest multi-method agreement observed consistently across both English- and French-trained settings. For the sentences in which there was a reason to treat them as noise, we have also performed analysis of the possible categories. This has led to detecting three main categories: *Structural* noise (artifacts which lack clausal structure), *Content* noise (distributionally atypical for the classifier), and *Label* noise. The subcategories of *Structural* noise have

been defined as:

SF: Structural fragments (truncation / broken segmentation), e.g., *"The dwarf planet Ceres is by far the largest asteroid, with a diameter of ."*; *", 85 HTCs have been observed, compared with 664 identified JFCs."*

MS: Markup or symbolic artifacts (wiki/template spillover, formulas/symbols), e.g., *"value:rgb (1,0.7,0.7) Legend:Xbox_360_..."*; *"+ 4 CO → 3 Fe + 4 . . . C + HO → CO + H."*

EN: Enumerations and list spillovers (category/tag lists), e.g., *"Apple Inc.|1976 establishments in California|American brands|. . . |Steve Jobs|Technology companies . . . "*; *"Game Boy Color games|Mario platform games|. . . |Virtual Console games for Wii U|. . . "*

LA: Language switch or encoding (non-English scripts or Unicode characters embedded inline), e.g., *"(Hong Kong) Co., Ltd. (*松下信*(*香港*)*有限公司*) and Panasonic SH. . . "*; *"The word physics comes from the Greek word* φύσις *('nature')."*

CI: Citation or reference fragments (dangling citations, references, or file/link spillover), e.g., *"↑ 'Sky & Telescope: March 2008', Southern Hemisphere Highlights . . . Allen, R. H. (1899) . . . "*; *"Internet map 1024.jpg | Partial map of the Internet . . . Structure of the Universe.jpg | Galactic. . . "*

The subcategories of *Content* noise are:

NE: Density of named entities, e.g., *"Cover athlete: Kaká (World), Wayne Rooney (United Kingdom), Mesut Özil & René Adler (Germany), Tim Cahill (Australia) . . . "*

NU: Numerical or measurement density, e.g., *"In currency, there are pennies ($0.01), nickels ($0.05), dimes ($0.1), quarters ($0.25), half dollars ($0.5), and dollar coins ($1)."*;

TE: Technical terminology density, e.g., *". . . Theoretical and Experimental Nuclear Physics, Nonlinear Optics, Thin film Magnetism, Neutron Scattering, Neutron Activation Analysis . . . "*

AC: Acronym or abbreviation density, e.g., *"Consumers are given the option to have any URL ending in .weebly.com, .com, .net, .org, .co, .info, or .us."*

The *Label* noise category covers two cases: sentences from Wikipedia are simple to be included in Vikidia (coded as MC) and sentences from Vikidia should have been considered complex (coded as MS).

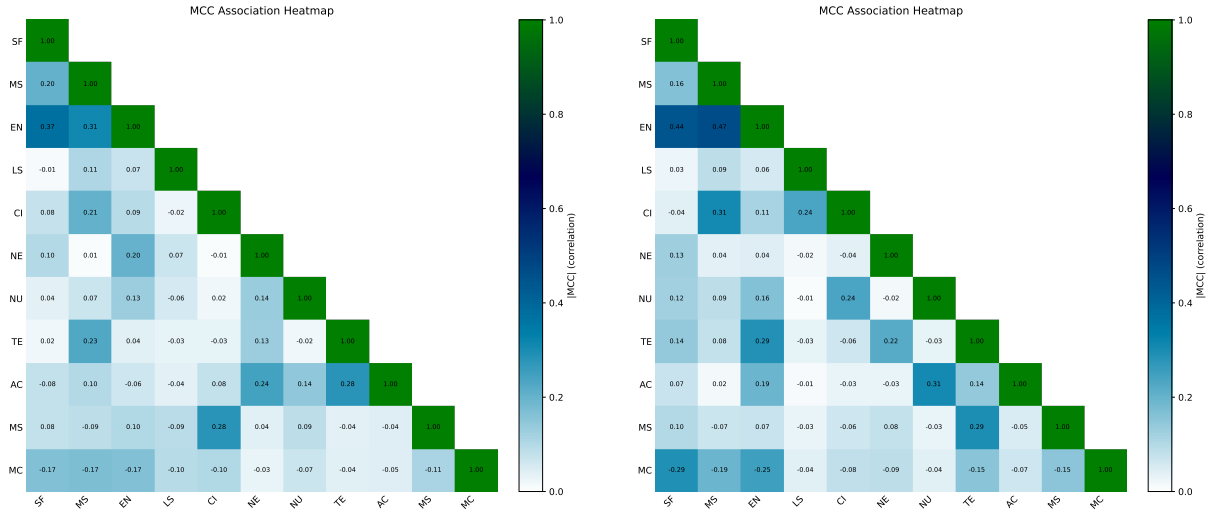
Table 3 shows the contribution of errors to the sentences detected as noisy. Even though, the sample was selected as agreement between noise detection methods, there was no reason to suspect noise in 38.7% of them for English and 33.1% for French. However, Table 4 shows that the rate of mislabelling (0.200 for English, 0.262 for French) in the case of what has been detected as noise was far higher than the error rate of the respective classifiers (0.043 for English and 0.083 for French). In practical terms, this means that a substantial portion of the flagged sentences in both languages are incorrect labels, which is the reason for improvement in performance once they are removed.

The most common category in both languages is the presence of structural artifacts from data processing, such as fragments, markup residue, list spillovers, or truncated segments. A second frequent source of noise is label noise, where sentences appear well-formed but assigned with the wrong complexity label (46 in English; 62 in French). These cases likely arise from annotation inconsistencies when document-level judgements are projected onto individual sentences. Content noise is the least common category, where sentences are dominated by standalone lists of names, numbers, or domain terminology (30 in English; 41 in French). These cases are often well-formed, but they are out-of-distribution for sentence-level difficulty prediction: when a line is mostly a list of entities or technical terms, it lacks normal sentence structure and its dense token pattern can be mistaken for high complexity.

While structural artifacts constitute the largest share of flagged noise, the presence of wrongly labeled sentences is particularly critical. These cases reflect annotation inconsistencies caused by projecting document-level complexity labels onto individual sentences. These findings provide strong motivation for applying denoising methods, as filtering or down-weighting such instances reduces the risk of learning incorrect patterns and improves the reliability of sentence-level difficulty classification.

To further investigate how the annotated noise categories interact, Figure 2 presents the pairwise Matthews Correlation Coefficient (MCC) (Matthews, 1975) between subcategories for English and French. While Table 4 reports category frequencies, the MCC heatmaps reveal patterns of co-occurrence, indicating which noise phenomena tend to appear together within the same sentence.

Across both languages, the strongest positive associations are concentrated among the structural noise. In particular, *SF*, *MS*, and *EN* tend to co-occur, with the highest correlations appearing in French (peaking at MCC $= 0.47$ for *MS–EN*, and 0.44 for *SF–EN*). This suggests that "noise" is often not a single isolated issue: markup spill-over



(a) English (b) French

Figure 2: Pairwise associations between noise categories measured using the Matthews Correlation Coefficient (MCC) for English (left) and French (right).

Category	**English**	**French**
Structural: Total	*185*	*186*
SF:Structural fragment	59	83
MS:Markup or symbolic artifact	46	36
EN:Enumeration or list artifact	44	57
LA:Language switch or encoding	17	2
CI:Citation or reference fragment	19	8
Content: Total	*30*	*41*
NE:Named entity density	14	9
NU:Numerical or measurement density	9	2
TE:Technical terminology density	3	25
AC:Acronym or abbreviation density	4	5
Label: Total	*46*	*62*
MS: Mislabeled as Simple	22	24
MC: Mislabeled as Complex	24	38
Total sentences	230	237

Table 4: Manual analysis of noise for English and French. Counts reflect annotation assignments (multiple labels per sentence allowed). Totals may exceed the number of noisy segments.

and formatting artifacts frequently trigger segmentation failures and fragment-like outputs, creating a compact structural core of errors. In contrast, content-density noise appears more independently. These categories (e.g., *NE*, *AC*, *TE*) only co-occur in a few specific pairs—such as *TE–AC* and *NE–AC* in English—rather than showing strong links across the whole taxonomy. Similarly, *LA* is near zero with most other tags.

A key observation is that *MC* (label noise) has mostly weak, and often negative, correlations with the structural cluster (e.g., as low as -0.29 with **SF** in French). This suggests that many mislabelled instances are actually well-formed sentences: they are problematic mainly because their assigned complexity label is incorrect, not because the text is corrupted in the usual sense. Overall, these results show that supervision errors in both languages come from different sources: structural issues, density-related effects, and annotation inconsistencies. This supports the need for denoising methods that can handle different types of errors, rather than assuming that all noise follows the same pattern.

The qualitative patterns help explain the complementary behaviour of our methods:

1. GMM-based filters mostly detect *distributional outliers* (domain NE density, symbolic tokens), yielding high-precision removal of atypical segments.

2. CT/ST preferentially downweight *structural fragments and corrupted strings* that produce unstable or persistently high losses across epochs.

3. LS does not *detect* noise per se, but improves calibration under mixed uncertainty (e.g., residual encoding noise or context loss), reducing overconfidence without aggressive data deletion.

These observations align with our quantitative results: intersection between GMMs and Co-Teaching captures the most consistent artifacts,

while Label Smoothing achieves competitive reliability at low computational cost.

Content noise categories are often linked to the lack of discourse context. This supports our earlier finding that longer segments improve robustness by reducing under-contextualised cases.

6. Related Studies

Noise in training data poses a significant challenge in NLP, especially in non-topical classification tasks such as genre prediction (Rönnqvist et al., 2022; Roussinov and Sharoff, 2023), demographic property detection (Kang et al., 2019), and text difficulty classification (North et al., 2022). These tasks rely on language style rather than explicit topical keywords, making them sensitive to noise and annotation errors.

Noise reduction techniques like majority agreement between the classifiers have been effective. Studies by Di Bari et al. (2014) and Khallaf and Sharoff (2021) show that leveraging consensus between the predictions of different models can significantly reduce noise, resulting in more reliable classifiers. Additionally, Zhu et al. (2022) provide a baseline by evaluating BERT models' robustness to label noise, without a clear outcome on which denoising methods are more useful. Our study goes further, by selecting a non-topical classification task and real-life settings by shifting prediction from document- to sentence-level predictions.

Bayesian learning has also been applied to handle noise, as discussed by Papamarkou et al. (2024) and Miok et al. (2020), focusing on managing uncertainty and noise in large-scale AI tasks. This approach is particularly relevant for semi-supervised text annotation, where it enhances noise reduction efficacy. Given the amount of unlabeled data in our domain, we will apply the experiments to the Bayesian framework.

Calibration of model predictions is crucial for handling noise, particularly using softmax outputs. Proper calibration ensures lower probabilities correspond to a higher likelihood of errors, aiding in producing well-calibrated classifiers. Methods for uncertainty estimates in BERT-like models can improve robustness to noise at the inference stage (Vazhentsev et al., 2023; Khallaf and Sharoff, 2026), which we need to investigate further.

Cross-lingual transfer learning, where models like BERT are trained on one language and applied to another, is particularly challenging in noisy environments due to linguistic differences and resource variability (Conneau et al., 2020; Zhao et al., 2021).

7. Conclusions

This paper examined the impact of various noise reduction techniques on cross-lingual sentence difficulty classification, providing insights into their effectiveness across languages and datasets. Our findings show that noise reduction can improve model performance, although its effectiveness depends on dataset characteristics and cross-lingual transfer conditions.

In our smaller dataset, Gaussian Mixture Models (GMMs) proved effective in mitigating noise. By contrast, in our larger dataset, the inherent regularisation properties of pretrained language models provide a strong baseline, with more computationally intensive denoising methods yielding only marginal additional gains. However, having a cleaner dataset reduces the training efforts and overall helps with future experiments.

These findings have practical implications for improving the robustness of cross-lingual applications in domains such as language education, text simplification, and language learning tools. By tailoring noise reduction strategies to dataset size and structure, developers can enhance the reliability and interpretability of sentence complexity models across languages.

The released dataset is the largest multilingual dataset for language difficulty prediction on the sentence level.

8. Acknowledgments

This document is part of a project that has received funding from the European Union's Horizon Europe research and innovation program under Grant Agreement No. 101132431 (iDEM Project). The University of Leeds was funded by UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee (Grant Agreement No. 10103529). The views and opinions expressed in this document are solely those of the author(s) and do not necessarily reflect the views of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

9. Limitations

This study aims at a specific non-topical classification task, for which there have been no prior experiments on denoising. The specific setup of moving from document- to sentence-level annotation relies on available Vikidia-Wikipedia pairs and on the availability of annotators, which limited the number of languages for testing. Our definition of difficulty relies on the Vikidia/Wikipedia proxy

and, despite test-set cleaning, may encode corpus-specific biases beyond difficulty. Future work will explore the applicability of these denoising techniques to other multilingual datasets and classification tasks, particularly in low-resource settings and for multiclass classification scenarios, as well as for other non-topical classification tasks. Additionally, investigating alternative sources of sentence-level annotations or adapting methods for diverse text genres (e.g., social media, news, or educational content) could further assess the robustness of our approach.

10. Ethical Impact

The potential societal benefits of our findings are substantial, particularly in improving the quality of communication by detecting complex sentences across languages. This study will also contribute to production of cleaner de-noised datasets.

In conducting the study we have been careful with the environmental impact of NLP research. Large Language Models are more computationally expensive, while they have been shown to be not better than BERT-like PLMs in text classification tasks. For each of the methods we estimated the computational costs of running the models (on NVIDIA L40S GPUs), with a total training time of ≈ 19 hours for English and ≈ 201 hours for French. We are not aware of potential risks in deploying the study discussed in the paper.

11. Bibliographical References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 2623–2631, New York, NY, USA. Association for Computing Machinery.

Devansh Arpit, Stanislaw Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. 2017. A closer look at memorization in deep networks. In *International Conference on Machine Learning*, pages 233–242. PMLR.

Christopher M. Bishop. 2006. *Pattern Recognition and Machine Learning*. Springer, New York.

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.

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*, pages 4171–4186, Minneapolis, Minnesota.

Nigel Dewdney, Carol VanEss-Dykema, and Richard MacMillan. 2001. The form is the substance: classification of genres in text. In *Proc. Human Language Technology and Knowledge Management*, pages 1–8.

Marilena Di Bari, Serge Sharoff, and Martin Thomas. 2014. Multiple views as aid to linguistic annotation error analysis. In *Proceedings of LAW VIII - The 8th Linguistic Annotation Workshop*, pages 82–86, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Aleksandra Edwards and Jose Camacho-Collados. 2024. Language models for text classification: Is in-context learning enough? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10058–10072, Torino, Italia. ELRA and ICCL.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. 2018. Co-teaching: Robust training of deep neural networks with extremely noisy labels.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: a massively multilingual multi-task benchmark for evaluating cross-lingual generalization. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Dongyeop Kang, Varun Gangal, and Eduard Hovy. 2019. (male, bachelor) and (female, Ph.D) have different connotations: Parallelly annotated stylistic language dataset with multiple personas. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1696–1706, Hong Kong, China. Association for Computational Linguistics.

Nouran Khallaf and Serge Sharoff. 2021. Automatic difficulty classification of Arabic sentences. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 105–114, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Nouran Khallaf and Serge Sharoff. 2026. To predict or not to predict? Towards reliable uncertainty estimation in the presence of noise. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Palma, Mallorca, Spain. European Language Resources Association (ELRA).

Taja Kuzman, Nikola Ljubešić, and Igor Mozetič. 2023. ChatGPT: Beginning of an end of manual annotation? use case of automatic genre identification. *arXiv preprint arXiv:2303.03953*.

Jing Li. 2024. Area under the ROC curve has the most consistent evaluation for binary classification. *PloS one*, 19(12):e0316019.

Michal Lukasik, Srinadh Bhojanapalli, Aditya Krishna Menon, and Sanjiv Kumar. 2020. Does label smoothing mitigate label noise? In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6448–6458. PMLR.

Brian W. Matthews. 1975. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, 405(2):442–451.

Kristian Miok, Gregor Pirs, and Marko Robnik-Sikonja. 2020. Bayesian methods for semi-supervised text annotation. In *Proceedings of the 14th Linguistic Annotation Workshop*, pages 1–12, Barcelona, Spain. Association for Computational Linguistics.

Rafael Müller, Simon Kornblith, and Geoffrey Hinton. 2019. When does label smoothing help? *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 4694 – 4703.

Kai North, Marcos Zampieri, and Matthew Shardlow. 2022. An evaluation of binary comparative lexical complexity models. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 197–203, Seattle, Washington. Association for Computational Linguistics.

Theodore Papamarkou, Maria Skoularidou, Konstantina Palla, Laurence Aitchison, Julyan Arbel, David Dunson, Maurizio Filippone, Vincent Fortuin, Philipp Hennig, Aliaksandr Hubin, et al. 2024. Position: Bayesian deep learning in the age of large-scale AI. *arXiv preprint arXiv:2402.00809*.

Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. 2017. Making deep neural networks robust to label noise: A loss correction approach. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2233–2241.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Han Ren, Yajie Zhao, Yong Zhang, and Wei Sun. 2024. Learning label smoothing for text classification. *PeerJ Computer Science*, 10:e2005.

Samuel Rönnqvist, Aki-Juhani Kyröläinen, Amanda Myntti, Filip Ginter, and Veronika Laippala. 2022. Explaining classes through stable word attributions. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1063–1074, Dublin, Ireland. Association for Computational Linguistics.

Dmitri Roussinov and Serge Sharoff. 2023. BERT goes off-topic: Investigating the domain transfer challenge using genre classification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Singapore.

B. W. Silverman. 1981. Using kernel density estimates to investigate multimodality. *Journal of the Royal Statistical Society: Series B*, 43(1):97–99.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826.

Artem Vazhentsev, Gleb Kuzmin, Akim Tsvigun, Alexander Panchenko, Maxim Panov, Mikhail Burtsev, and Artem Shelmanov. 2023. Hybrid uncertainty quantification for selective text classification in ambiguous tasks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11659–11681, Toronto, Canada. Association for Computational Linguistics.

Hao Yang, You-Zhi Jin, Zi-Yin Li, Deng-Bao Wang, Xin Geng, and Min-Ling Zhang. 2024. Learning from noisy labels via dynamic loss thresholding. *IEEE Transactions on Knowledge and Data Engineering*, 36(11):6503–6516.

Xiyu Yu, Tongliang Wu, Chaochao Wei, Bo Liu, Wei Liu, and Dacheng Tao. 2019. How does disagreement help generalization against label corruption? In *International Conference on Machine Learning*, pages 7164–7173. PMLR.

Mengjie Zhao, Yi Zhu, Ehsan Shareghi, Ivan Vulić, Roi Reichart, Anna Korhonen, and Hinrich Schütze. 2021. A closer look at few-shot crosslingual transfer: The choice of shots matters. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5751–5767, Online.

Dawei Zhu, Michael A. Hedderich, Fangzhou Zhai, David Ifeoluwa Adelani, and Dietrich Klakow. 2022. Is BERT robust to label noise? a study on learning with noisy labels in text classification. In *Proceedings of the Third Workshop on Insights from Negative Results in NLP*, pages 62–67, Dublin, Ireland. Association for Computational Linguistics.