From Babel to Brilliance: De-Noising Techniques for Cross-Lingual Sentence-Difficulty Classifiers

Anonymous ACL submission

Abstract

Noisy training data can significantly degrade the performance of classifiers using language models, particularly in applications such as readability assessment, content moderation, and language learning tools. This study investigate the use of several de-noising techniques in sentence-level difficulty detection, using a training set derived from document-level difficulty annotation. In addition to monolingual de-noising, we address the cross-lingual transfer gap when a multilingual language model is trained on one language and tested on another. We have examined the influence of segment lengths and have studied a wide range of noise reduction techniques, such as Gaussian Mixture Models, Co-Teaching, Noise Transition Matrices, and Label Smoothing. Results reveal that, while BERT-like models are robust to noise, incorporating noise detection can further enhance performance. For a smaller dataset, Gaussian Mixture Models can be especially helpful to reduce noise and improve prediction quality, especially in the cross-lingual transfer. However, for a larger dataset the inherent regularisation of the PLMs provides a good baseline, which (fairly expensive) de-noising methods cannot improve further.

1 Introduction

Modern NLP methods such as Pre-Trained Language Models (PLMs) and Large Generative Language Models (LLMs) have markedly advanced the performance across a wide range of tasks. Nevertheless, significant challenges persist in non-topical classification tasks, such as predictions of genre (Rönnqvist et al., 2022; Kuzman et al., 2023), demographic properties (Kang et al., 2019), or the difficulty of a text (North et al., 2022). Zhang et al. (2023) further highlight the limitations of large language models in handling subtle stylistic features in such tasks. In comparison to more standard topical classification tasks, which rely on explicit keywords, non-topical classification requires models to discern stylistic features rather than overt domain-specific content (Dewdney et al., 2001). Both PLMs and LLMs can be distracted by topical keywords in such tasks (Roussinov and Sharoff, 2023). Also larger LLMs, such as GPT, do not outperform smaller PLMs, such as BERT, on text classification tasks (Edwards and Camacho-Collados, 2024), while the LLMs might be better at sentence-level simplification tasks (Kew et al., 2023).

The scarcity of high-quality data across languages, coupled with prevalent annotation noise, presents a significant challenge in sentence difficulty classification. This study addresses noise originating from the variability of crowd-sourced annotations and the transition from document-level to sentence-level predictions. For instance, a sentence extracted from Wikipedia–which we categorise as a source of more complex texts–can often be simple, whereas individual sentences from crowd-sourced simple language resources may occasionally exhibit complexity or become difficult to interpret without broader context. This inconsistency further challenges the effectiveness of non-topical classification models.

Accurately predicting the complexity of a single short sentence is challenging because individual sentences often lack the contextual richness required for proper assessment. In order to solve this problem, it is essential to investigate the impact of aggregating multiple short sentences when determining the complexity of a text segment.

To address these challenges, this study makes the following contributions:

1. Investigate the impact of segment length on handling noise in sentence-level difficulty prediction.

2. Identify effective methods for filtering noisy datasets to improve classification robustness.

3. Evaluate the effect of noise in both monolingual and cross-lingual classification settings.

4. Release a de-noised multilingual dataset, along with scripts and models, in the camera-ready version.

2 Dataset

The datasets used in this experiment 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". 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 negatively impact non-topical classification tasks. 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. The Inter-Quartile Range (IQR) column in Table 1 shows that the majority of sentences in either collection are of about 8 to 21 words, with longer sentences being about 4-6 words shorter in Vikidia.

3 Methodology

All experiments listed below aim at a binary task of predicting whether a sentence (or a short segment) is complex, i.e., it is NOT suitable to be included in Vikidia. However, this makes it a noisy task, as some Wikipedia segments can be legitimately simple. This also makes the Area Under Curve (AUC) score as the main evaluation measure, as we are specifically interested in reducing the impact of False Negatives (complex segments which have not been identified).

Our experiments have been conducted using multilingual BERT-base (Devlin et al., 2019) with some experiments using multilingal SBERT (Reimers and Gurevych, 2019). We have also conducted preliminary experiments with mBERT-large and XLM-Roberta (Conneau et al., 2020), which demonstrated comparable patterns, so the focus of this paper is on the core models for clarity and consistency.

Our baseline models were fine-tuned separately on the English and French datasets using the Huggingface Transformers library with the default hyper-parameters over 10 epochs with early stopping. The English dataset serves as a widely used benchmark, while the French dataset, with over 2 million sentences, provides an opportunity to investigate how larger training corpora influence model robustness. The performance of the English and French models was subsequently evaluated in other languages to assess cross-lingual transferability. The PLMs used (mBERT and XLM-Roberta) have English as the biggest corpus for pre-training, thus leading to better weight estimation for English. However, for our specific downstream task the French dataset is much bigger, so we investigated how larger training corpora influence noise detection. Additionally, we test across a range of languages from Vikidia, including Catalan and Russian. Catalan has the smallest amount of pre-training corpora, so we can expect lower quality weights, while Russian is more remote from languages prevalent in Vikidia, so we expect greater cross-lingual transfer gap (Hu et al., 2020).

Traditional approaches often rely on sentence-based segmentation, where each sentence is treated as an independent input. However, this can introduce fragmentation and fail to capture contextual dependencies. First, we compared two strategies to investigate the effects of text segmentation:

1. Splitting texts into single sentences using NLTK and SpaCy, as these libraries use different algorithms for sentence boundary detection, potentially influencing performance. Sentences from Wikipedia also present in Vikidia were removed from the datasets.

2. Combining adjacent sentences to create chunks of varying lengths (approximately 50, 70, 100, 150, 200 and 250 tokens), as the original sentences are likely to be too short for reliable predictions.

The objective was to assess how text segmentation affects the results of a text classification task with the aim of determining the optimal segment length performance in the original language and in cross-lingual applications.

After determining the optimal segment length, we experimented with noise reduction techniques

[1] The datasets from *Wikipedia* and *Vikidia* are available under the **(CC BY-SA)** license. Our use aligns with their intended purpose of open knowledge sharing, with modifications limited to filtering and preprocessing for research.

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

Language	#Texts	Wikipedia			Vikidia		
		#Words	#Sentences	IQR	#Words	#Sentences	IQR
Catalan	125	434344	807	(7, 22)	77223	803	(7, 16)
English	1726	8843600	16431	(11, 25)	560226	17712	(9, 19)
Spanish	2738	8504129	21609	(8, 25)	685636	19034	(9, 22)
French	21515	49689858	1998934	(14, 31)	7293514	374537	(11, 24)
Italian	2456	7479678	23432	(8, 25)	604046	19325	(9, 21)
Russian	104	396786	692	(5, 15)	13877	551	(5, 14)

Table 1: Statistics of the dataset with information on the number of documents, words and sentences for each dataset, as well as the Inter-Quartile Range of the sentence lengths as from Spacy's tokenization and segmentation.

to remove potentially noisy segments from the training data, i.e., the segments with incorrect or not useful gold-standard annotations because of the process of dataset construction. After detecting the noisy data points, we retained only the datapoints identified as clean, and fine-tuned with the same hyperparameters as the baseline to maintain consistency. Both the validation and test datasets have been manually cleaned, ensuring reliable evaluation.

To address label noise, we compare five established denoising methods:

Gaussian Mixture Models (GMMs): GMMs can be used to cluster high-dimensional sentence representations into two Gaussian distributions, corresponding to clean and noisy data, respectively (Bishop, 2006). We evaluated sentence representations derived from SBERT (GMM-SB) and CLS embeddings from BERT (GMM-B).

To optimise the GMM parameters—such as the number of components, covariance type, and convergence tolerance—we employed Optuna (Akiba et al., 2019) for hyperparameter tuning over 100 trials. The optimization objective was to maximize the score, ensuring more effective differentiation between noisy and non-noisy data (see Table 2).

Table 2: Optimal GMM Configurations

Parameter	SBERT EN	SBERT FR	BERT EN	BERT FR
Components	9	8	10	3
Covariance	Spherical	Full	Tied	Full
Tolerance	5.69×10^{-5}	1.04×10^{-3}	9.89×10^{-5}	5.00×10^{-4}
Max Iterations	234	278	172	50
Threshold	0.456	0.759	0.503	0.479

Small-loss Trick (ST): ST identifies noisy data points by leveraging the assumption that difficult-to-learn examples (i.e., those with high training losses) are more likely to contain noise (Arpit et al., 2017; Han et al., 2018). During each training iteration, the model computes the loss for each sentence and selects a subset with the smallest losses, assuming these correspond to correctly labeled instances. The model is then updated using only this subset, thereby reducing the impact of mislabeled or noisy data on training (Malach and Shalev-Shwartz, 2017; Yu et al., 2019).

A key hyperparameter in this method is the *loss threshold percentile*, which determines how many low-loss samples are retained for training. We experimented with four different threshold values—10th, 25th, 50th, and 75th percentiles—to evaluate their impact on filtering noisy data. Our results indicate that using a 75% threshold provided the best balance, effectively excluding uncertain data points while retaining a sufficient number of reliable instances for model training.

Co-Teaching (CT): Similar to ST, this method relies on small losses, but it involves simultaneous training of two models, each initialised independently (Han et al., 2018). During training, each model selects the lowest-loss samples from batches of the other model, under the assumption that these are more likely to be correctly labeled. This cross-filtering mechanism ensures that each model learns from cleaner examples, reducing the influence of noisy data. By defining, the key parameter in CT, the *forget rate*, which determines the fraction of high-loss samples to remove 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. Progressive filtering outperforms static thresholding by maintaining a more diverse and informative dataset, enhancing robustness against noisy labels (Yang et al., 2023).

In our implementation, we use a linearly scheduled forget rate ranging from 0.0 to 0.3 over training epochs, following:

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

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 highest-loss samples per batch are progressively discarded ($r_t \to 0.3$)

Additionally, to further refine noise removal, we implemented a post-training filtering mechanism based on maximum predicted probability. This mechanism evaluates each sample's highest softmax probability p assigned to a predicted class. Samples with $p^* < 0.6$ are flagged as potentially noisy and removed. Our analysis showed that at 0.50, no samples were filtered, whereas at 0.69, 0.90% of samples were flagged, providing a balanced trade-off between noise removal and data retention.

Noise Transition Matrix (NTM): NTMs model the probability of labeling errors, enhancing feedback for misclassifications during training (Patrini et al., 2017). Each element T_{ij} in this matrix indicates the likelihood that a true label i is mistakenly assigned the label j. We can estimate this matrix using prior knowledge about the dataset or by analysing predictions from a clean subset. Unlike CT and ST, which discard noisy samples based on loss, NTM does not remove data but instead adjusts model predictions to compensate for label noise. In our experiment, we used the noisy data identified by GMM-B to derive the Noise Transition Matrix. Specifically, we derive T_{ij} from the confusion matrix created from the noisy dataset, and then we normalise it to show the probabilities of mislabelling. To effectively utilise the NTM, we compute its inverse, known as the Inverse Noise Transition Matrix (T_{inv}).

During the training process, instead of filtering out noisy samples as in CT or ST, we multiply the model's raw predicted probabilities, denoted as $\hat{P}$, by T_{inv}. This adjustment corrects the predictions to account for label noise, resulting in a more accurate estimation of the true labels.

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

Next, we compute the loss function for the adjusted predictions $\hat{P}_{\text{adjusted}}$ using the cross-entropy loss. By integrating T_{inv} into the training loop, the model can dynamically correct noisy predictions during training.

Label Smoothing (LS): LS is a regularization technique that reduces model overconfidence. 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). LS has been shown to improve model performance by helping the model generalise and reducing overfitting, especially when dealing with noisy or unbalanced datasets (Khan and B., 2023; Adikari and Draper, 2023). The smoothed label y_{smooth} 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 probability 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 probability thresholds ($\tau = 0.70$) yielded superior performance by effectively filtering uncertain predictions, leading to improved model reliability.

4 Segmentation Strategy Evaluation

The choice of segmentation strategy plays a critical role in classification performance, probability calibration, and computational efficiency. We evaluate longer segments against single sentences. As shown in Figure 1, our findings confirm that fixed-length token segmentation consistently outperforms sentence-based segmentation by preserving input consistency, reducing fragmentation, and ensuring more stable representations across languages. Sentence-based or short-length segmen-

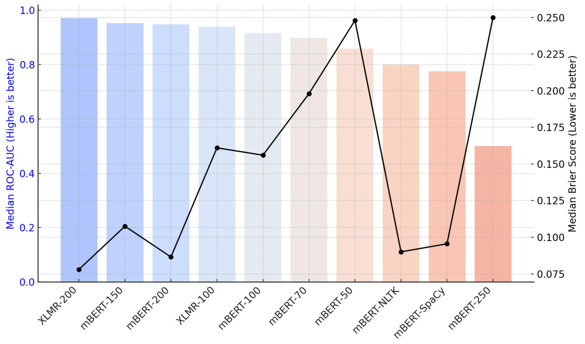


Figure 1: Comparison of ROC-AUC and Brier Score across segmentation strategies. Higher ROC-AUC and lower Brier Score indicate better classification performance and better probability calibration.

tation yields lower ROC-AUC in comparison to longer segments. Additionally, models trained on shorter segments exhibit poor probability calibration, as indicated by higher Brier Scores and greater uncertainty, measured by Prediction Entropy. This suggests that shorter segments lead to overconfident yet incorrect probability estimates, reducing classification reliability. Interestingly, the 250-token segmentation does not yield improvements over 200-token; instead, its ROC-AUC drops, suggesting that longer segments may introduce unnecessary noise or redundant context that does not contribute to classification performance.

Between mBERT-200 and XLMR-200, key trade-offs emerge. XLMR-200 achieves a higher ROC-AUC (+0.014) and Precision (+0.026), indicating stronger ranking ability and fewer false positives. However, mBERT-200 compensates with a higher Recall (+0.033). mBERT-200 also offers superior efficiency, with faster training (498s vs 1327s) and a slighlty faster inference (12.4 ms vs. 13.8 ms per segment).

5 Noise Detection and Reduction

Building upon our previous experiment, where we determined that a 200-token segmentation provided the most effective results, we expanded our study to evaluate the impact of the noise reduction techniques on classification performance. Our objective was to assess how these techniques influence model robustness in the cross-lingual setting, using English and French as the primary training datasets. The English dataset consists of 4,644 training segments and 1,162 validation segments, while the French dataset comprises 65,803 training segments and 16,451 validation segments. To evaluate cross-lingual generalisation, we tested models trained on English and French against each other, in addition to multilingual evaluation across Catalan (346 segments), Spanish (8,076 segments), Italian (7,348 segments), and Russian (266 segments).

In addition to investigation of the de-noising techniques, we performed Intersection Analysis by examining noise labels flagged by multiple methods to determine the most consistently noisy data points.

When trained on English (Table 3), the baseline model achieved the AUC score of 0.991. Both Gaussian Mixture Models (GMM-B and GMM-SB) improved this performance to 0.996 and 0.997 respectively. Notably, their intersection performed even better, as it achieved the highest AUC score of 0.998. This suggests that combining noise reduction methods can yield superior classification outcomes. Performance has also improved in the cross-lingual setting with again GMM-B and GMM-SB outperforming the baseline model and improving further at their intersection. Intersecting GMM-B with Co-Teaching led to the best cross-lingual performance averaging over the Vikidia languages (0.971 AUC) with the transfer gap of merely 0.02.

However, when trained on a much bigger French dataset (Table 4), the baseline performance remained the best both for French itself (0.998) as well as in the cross-lingual setting (0.979), outperforming the best English models. In itself it demonstrates that English-centric training procedure does not always help for multilingual models if better data is available for another language. Catalan, Italian and Spanish are typologically related to French, so we can expect better performance there. As for the noise detection, our baseline training on a significantly larger French dataset improved generalisation across languages even without the need for de-noising. Here, GMM-B/SB emerged as the best standalone methods (0.997) with their intersection getting the same performance as the baseline but additional improvements.

Balancing effectiveness with computational efficiency is crucial. While French-trained models generalised better, they also introduced higher computational costs. Training the baseline model was the most efficient (498 seconds for English, 2466 seconds for French), while denoising with GMM-based methods, despite their strong performance, were computationally demanding (3356 seconds for English, 47,736 seconds for French), highlighting scalability challenges.

								Intersection			
Language	Baseline	GMM-B	GMM-SB	ST	CT	NTM	LS	CT/ST	ST/GMM-SB	GMM-B/SB	CT/GMM-B
en	0.991	0.996	**0.997**	0.876	0.988	0.988	0.995	0.989	0.998	**0.999**	0.998
ca	0.927	0.930	0.945	0.990	0.896	0.902	0.926	0.908	0.942	0.942	0.955
es	0.947	0.962	0.966	0.881	0.950	0.949	0.963	0.924	0.956	0.972	0.974
fr	0.935	0.953	0.959	0.877	0.916	0.925	0.947	0.894	0.938	0.965	0.966
it	0.948	0.962	0.963	0.849	0.929	0.931	0.954	0.901	0.956	0.970	0.969
ru	1.000	0.999	0.998	0.847	0.984	0.990	0.991	0.966	0.998	0.996	0.993
Average	0.951	0.961	**0.966**	0.888	0.935	0.939	0.956	0.918	0.958	0.969	**0.971**
Train (s)	498	3356	3331	706	762	3588	**360**	2085	4012	5744	4097
# Noisy		432	1234	1430	1704	100	2060	433	265	228	368

Table 3: Comparative performance of noise reduction methods using ROC-AUC when trained on English.

								Intersection			
Language	Baseline	GMM-B	GMM-SB	ST	CT	NTM	LS	CT/ST	ST/GMM-SB	GMM-B/SB	CT/GMM-B
fr	**0.998**	0.997	0.997	0.918	0.902	0.753	0.993	0.996	0.997	0.997	**0.998**
en	0.991	0.991	0.992	0.918	0.970	0.753	0.991	0.988	0.990	0.992	0.991
ca	0.952	0.954	0.950	0.918	0.902	0.645	0.937	0.932	0.943	0.945	0.949
es	0.979	0.976	0.978	0.921	0.888	0.855	0.982	0.974	0.972	0.973	0.978
it	0.978	0.975	0.975	0.907	0.959	0.781	0.978	0.971	0.970	0.973	0.978
ru	0.995	0.983	0.993	0.732	0.962	0.907	0.989	0.978	0.988	0.994	0.991
Average	**0.979**	0.975	0.977	0.879	0.936	0.788	0.975	0.968	0.972	0.975	0.977
Train (s)	2466	47736	48578	**719**	8512	46115	1131	9831	48446	93732	57123
# Noisy		23647	11030	19129	29615	1193	2060	22133	16717	4509	9212

Table 4: Comparative performance of noise reduction methods using ROC-AUC when trained on French.

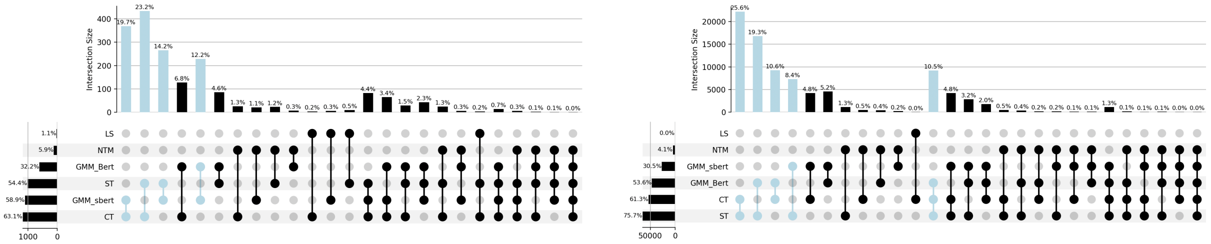
Conversely, the Small-Loss Trick remained efficient (706 and 719 seconds). Co-Teaching, though effective in English (762 seconds), became more costly in French (8512 seconds), indicating that larger datasets require more iterations for achieving agreement between the two models. NTM was the most expensive method, while it did not lead to improvements over GMMs. Label Smoothing, the least costly method (360 and 1131 seconds), performed adequately improving over the baseline almost on par with the much more expensive GMM-B/SB methods.

To better understand the overlap between noise detection techniques, the intersection analysis of noisy segments across denoising techniques (Figures 2a and 2b) reveals distinct trends between English- and French-trained models. These plots illustrate how different denoising methods agree or diverge in identifying noisy segments.

In English, noise detection is dominated by ST (84.2%) and CT (55.5%), suggesting a more aggressive filtering approach, while higher-order intersections remain sparse, indicating that methods largely detect distinct subsets. In contrast, French-trained models exhibit a more balanced distribution, with ST (80.3%), CT (57.1%), and GMM-BERT (52.0%) showing greater agreement. The increased frequency of multi-method intersections suggests that larger datasets allow for more consistent noise identification, making hybrid approaches more effective. Notably, Label Smoothing plays a minimal role in both cases, reinforcing its function as a regularisation tool rather than a primary noise filter.

Figure 3 provides a comparative visualization of model strengths and weaknesses across Accuracy, Precision, Recall, F1-score, ROC AUC, and Brier Score. A clear Precision-Recall trade-off emerges: *GMM-based models* favor high precision at the expense of recall, while LS and ST maintain a more balanced profile. Additionally, ROC AUC and Brier Score discrepancies indicate differences in probability calibration, with some models ranking instances well (measured with ROC AUC) but producing less reliable probability estimates (measured with the Brier Score).

While LS and CT offer stable performance, models like CT, ST and GMM-B improve precision at the expense of recall, making them more suitable for high-confidence classification tasks. The radar chart further illustrates these trade-offs, with circular patterns indicating balanced performance and irregular shapes reflecting metric-specific biases.



(a) Noisy data intersections for English.

(b) Noisy data intersections for French.

Figure 2: Comparison of noisy data intersections across denoising techniques for English and French. Bars indicate detected noisy segments, and connections show agreement among techniques. Greater overlap suggests higher agreement across techniques.

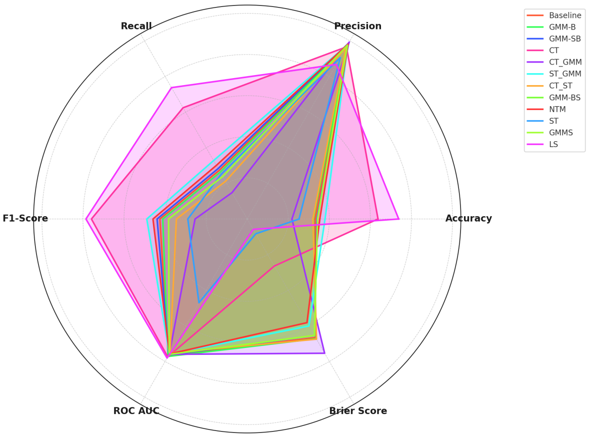


Figure 3: Radar Chart: Model Strengths and Weaknesses across Key Metrics

6 Error Analysis

Considering the intersection of CT and GMM-B as the best-performing de-noising model, we conducted a manual error analysis to further assess its effectiveness. A total of 368 noisy segments were identified at the intersection, as summarized in Table 5, with 151 coming from simpler Vikidia and 217 from more complex Wikipedia.

To evaluate the noise classification performance, we randomly selected a sample of 50 instances, consisting of 20 sentences from Vikidia and 30 from Wikipedia. Within this subset, 26 sentences were classified as noisy, while 24 were considered not noisy, as detailed in Table 5.

To better understand the nature of noise in the dataset, we categorised the noisy segments into different issue types (see Table 6). It is important to note that the same sentence may fall under more than one category, as multiple types of noise can coexist within a single sentence.

Our manual error analysis highlights seven key noise factors that affect text processing in both Simple Vikidia and Complex Wikipedia. The identified noise types and their distribution are detailed in Table 6. The most frequent issue was misclassification as simple, followed by irregular encoding problems, which likely affected sentence structure and classification accuracy.

Category	Count
Total Noisy Sentences	368
Simple Vikidia (Noisy)	151
Complex Wikipedia (Noisy)	217
Sample Size	50
Simple Vikidia (Sample)	20
Complex Wikipedia (Sample)	30
Noisy in Sample	26
Not Noisy in Sample	24

Table 5: Summary of Noisy Segment Analysis

Issue Category	Count
Misclassified as Simple	21
Misclassified as Complex	1
Irregular Encoding	18
Incomplete	1
Mostly a list of Named Entities	7
Mostly Referencing	6
Listing Numbers (mostly years)	7

Table 6: Categorization of Noisy Segments Based on Identified Issues

The heatmap in Figure 4 illustrates the correlation between these noise categories and classification errors. We observe the following key trends:

- Certain noise categories exhibit strong positive correlations with misclassification. For example, encoding issues and excessive named entities appear to contribute more frequently to sentences being misclassified as complex.

- Sentences containing excessive referencing and multiple years mentioned show a notice-

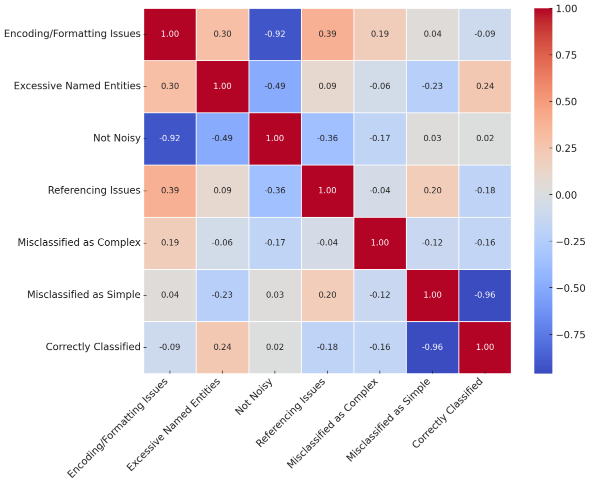


Figure 4: Heatmap showing the correlation between different noise categories and classification outcomes. Positive correlations (red) indicate a stronger association, while negative correlations (blue) suggest an inverse relationship.

able association with misclassification as simple.

- The strongest correlation with correctly classified sentences appears in cases with minimal formatting issues or named entities, indicating that clean, structured text is less prone to misclassification.

7 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 de-noising 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 (Kuleshov et al., 2018; Vazhentsev et al., 2023), which we need to investigate further, while this study only assesses the degree of calibration via the Brier score.

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).

8 Conclusions

This paper investigates the impact of various noise reduction techniques on cross-lingual sentence difficulty classification, providing insights into their effectiveness across different languages and datasets. Our findings demonstrate that these methods can enhance model performance, with their effectiveness varying based on specific data characteristics and cross-lingual transfer settings. Classification of complex language for longer segments is consistently better than for single sentences. For a smaller dataset, Gaussian Mixture Models can be helpful to reduce noise, while for a bigger dataset the inherent regularisation of the PLMs provides a good baseline, which more expensive de-noising methods cannot improve further.

These findings have practical implications for enhancing the reliability of cross-lingual applications in areas such as language education, language simplification, and language learning tools.

9 Limitations

This study aims at a specific non-topical classification task, for which there have been no prior experiments on de-noising. However, the specific setup of moving from document- to sentence-level annotation relies on freely available Vikidia-Wikipedia pairs, which also offer a limited number of languages for testing. Future work will explore the applicability of these de-noising techniques to other multilingual datasets and classification tasks, particularly in low-resource settings. 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 PLMs in several text classification tasks. For each of the methods we provided the computational costs of running the models (on NVIDIA L40S GPUs), with a total training time of ≈ 9 hours for English and ≈ 101 hours for French. We are not aware of potential risks in deploying the study discussed in the paper.

References

D. Adikari and S. Draper. 2023. Label smoothing for enhanced text sentiment classification. *arXiv preprint arXiv:2312.06522*.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *The 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2623–2631.

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.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034, 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 *Proc LAW VIII - The 8th Linguistic Annotation Workshop*, Dublin, Ireland.

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.

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. In *Advances in neural information processing systems*, pages 8527–8537.

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. *arXiv preprint arXiv:2003.11080*.

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.

Tannon Kew, Alison Chi, Laura Vásquez-Rodríguez, Sweta Agrawal, Dennis Aumiller, Fernando Alva-Manchego, and Matthew Shardlow. 2023. BLESS: Benchmarking large language models on sentence simplification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13291–13309, Singapore. Association for Computational Linguistics.

Nouran Khallaf and Serge Sharoff. 2021. Automatic difficulty classification of Arabic sentences. In *Proceedings of the Sixth Arabic Natural Language Pro-*

cessing Workshop, pages 105–114, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Z. Khan, A. Wang and Liu B. 2023. Learning label smoothing for text classification. *PeerJ Computer Science*, 9:e1102.

Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. 2018. Accurate uncertainties for deep learning using calibrated regression. In *International Conference on Machine Learning*, pages 2796–2804.

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*.

Eran Malach and Shai Shalev-Shwartz. 2017. Decoupling" when to update" from" how to update". In *Advances in neural information processing systems*, pages 960–970.

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? In *Advances in 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1944–1952.

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. Association for Computational Linguistics.

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.

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.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 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, Tongliang Liu, Bo Han, Gang Niu, and Masashi Sugiyama. 2023. Learning with noisy labels via dynamic loss thresholding. *IEEE Transactions on Knowledge and Data Engineering*, 35(1):348–361.

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.

Wei Zhang, Ming Liu, Hao Wang, and Yinan Li. 2023. Fine-grained evaluation of large language models for non-topical classification tasks. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1234–1245.

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 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.