Parallel Sentence Filtering for Low-Resource Language Pairs: A Case Study for Upper Sorbian, German, and Czech

Ruiyang Jiang[1], Shu Okabe[1,2], Alexander Fraser[1,2,3]
[1]Technische Universität München (TUM), [2]Munich Center for Machine Learning,
[3]Munich Data Science Institute
[1,2]Munich, Germany, [3]Garching, Germany
{ruiyang.jiang, shu.okabe}@tum.de

Abstract

As parallel corpora for low-resource languages are scarce, and automatic approaches to mine sentence pairs can lead to noisy datasets, parallel sentence filtering aims to detect only actual translations. We study here two language pairs: Upper Sorbian–German and Czech–German to represent both high and low availability of data resources. To evaluate filtering performance, we generate synthetic datasets by combining existing parallel corpora with synthetic non-parallel pairs, notably with five types of local semantic changes on the German side, such as negation or modality transformations. We represent sentences using three multilingual language models, XLM-R, Glot500m, and LaBSE, and train classifiers for the task. All three model representations led to worse filtering quality when pairs were altered more subtly, such as an antonym replacement. We still observed that a language model pre-trained on the considered language achieves more robust classification performance when sentence pairs are more ambiguous. We also evaluated a cross-lingual approach where the classifier is trained on the Czech–German pair and then applied to the Upper Sorbian–German pair. Such a language transfer paves the way for filtering other low-resource language pairs in the future.

Keywords: parallel sentence filtering, language representation, low-resource languages

1. Introduction

Parallel corpora represent a crucial resource, as they notably help improve the multilingual capabilities of language models or the training of a machine translation system. Such datasets are widely available for high-resource language pairs (e.g., English–German), but become scarcer when they involve low-resource languages. To address this lack of large corpora, parallel sentence mining (or bitext mining) aims to discover translation pairs from two monolingual texts, ideally from a similar domain, such as news or Wikipedia articles. While this can lead to massive multilingual corpora, such as WikiMatrix (Schwenk et al., 2021), the automatic extraction may also contain noisy pairs. This can be critical, especially for low-resource languages, as they negatively affect downstream performance (Lin et al., 2025).

To tackle this issue, several WMT shared tasks focused on the task of parallel corpus filtering, especially in low-resource conditions (Koehn et al., 2020; *inter alia*). The objective is then to predict whether a given sentence pair from a noisy corpus is a translation or not. One of the notable approaches trains classification models to distinguish parallel from non-parallel sentences, improving the quality of the filtering and, hence, of the final bilingual corpus.

This article thus uses a classification approach to study the task of parallel sentence filtering. We focus on two language pairs: Upper Sorbian–German (HSB–DE) to represent a low-resource data condition and Czech–German (CS–DE) for a higher-resourced case. Recent shared tasks on Machine Translation (Weller-Di Marco and Fraser, 2022; Okabe et al., 2025a) concentrate on the former language pair, underlining the current utility of clean parallel corpora.

As no existing datasets fit our task (with sentence pairs classified based on translation quality), we created synthetic corpora: sentence pairs labelled as 'parallel' come from existing parallel corpora, and 'non-parallel' pairs from randomly paired monolingual sentences. As such random pairs appeared too simple to detect, we also generated more challenging non-parallel sentence pairs. We follow the methodology of (Chen et al., 2023) and newly extend it to our two non-English-centric language pairs. This also gives us full control over the type and proportion of noise in the pairs. In practice, we altered the sentences on the German side with a set of five word-level replacements: antonym, negation, modal verb transformations, as well as entity and number replacements. We strove to give a more balanced representation of all transformation types, contrary to (Chen et al., 2023).

We rely on a simple classification pipeline: first, it encodes sentence pairs thanks to multilingual language models (XLM-R (Conneau et al., 2020), Glot500m (Imani et al., 2023), and LaBSE (Feng et al., 2022)). Then, based on their representation and the cosine similarity, we train standard classifiers to predict a binary label. On clearly different sentence pairs, high classification accuracy can

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 8883–8896
11–16 May 2026.

be achieved with a reliable language model or a strong classifier. As expected, our word-level transformations made the dataset more complex for all models.

Moreover, as Upper Sorbian is a low-resource language but related to Czech, we devise and evaluate a cross-lingual strategy where we train on the higher-resourced CS–DE pair and apply the classifier to the unseen HSB–DE. We investigate here whether language transfer can help filtering when parallel data is too scarce to train a classifier.

We aim to answer the following research questions, mainly for the low-resource Upper Sorbian–German language pair:

- **RQ1**: How well are the language pairs represented in the multilingual space by the language models for parallel sentence filtering?
- **RQ2**: To what extent does the word-level perturbation introduced in the sentence impact the classification quality?
- **RQ3**: Can a classifier trained on a better-resourced language pair be used for a low-resourced pair?

Our contributions are as follows: (i) we create a synthetic classification dataset for parallel sentence filtering with five sentence transformations, (ii) we train and analyse standard classifiers to distinguish parallel from non-parallel sentences based on the sentence embeddings, and (iii) we evaluate the robustness of the classifiers trained on the same or related language pair. We publicly release the created datasets alongside the code material.[1]

2. Related work

Two tasks aim to create parallel corpora automatically, as they are less widely available compared to their monolingual counterparts: parallel sentence mining and filtering. They can be regarded as successive steps: mining gathers sentence pairs, which can then be filtered to obtain a higher-quality corpus.

Parallel sentence mining The BUCC shared tasks (Zweigenbaum et al., 2017, 2018) notably studied the parallel sentence mining task. These only considered high-resource pairs, such as English–German, which prompted other works on low-resource language pairs. Besides, submitted or concurrent systems usually relied on bilingual and static embeddings (Hangya and Fraser, 2019), while later approaches use contextual representations as in (Artetxe and Schwenk, 2019a). Further works improved the sentence-level representation of low-resource languages for the task using distillation and parallel sentences (Heffernan et al., 2022; Tan et al., 2023) or post-processing methods (Okabe et al., 2025b). All mining approaches rely on a filtering process, which is usually based on a similarity score to keep or discard sentence pairs.

Parallel corpus filtering The task of parallel corpus filtering aims to improve the last step of the mining pipeline or, more generally, the quality of a bilingual corpus. The goal is to remove noisy sentence pairs from an allegedly parallel corpus, for instance, scraped automatically through mining. Several editions of WMT shared tasks focused on improving the quality of such corpora (Koehn et al., 2018), specifically for low-resource languages and pairs (Koehn et al., 2019, 2020; Sloto et al., 2023). Methods based on multilingual pre-trained models, such as LASER (Artetxe and Schwenk, 2019b), are competitive approaches for the task (Chaudhary et al., 2019). Parallel corpus filtering is frequently treated as a classification task, with models trained on clean parallel sentences and synthetic noisy pairs (Xu and Koehn, 2017; Zaragoza-Bernabeu et al., 2022). A sentence classification approach can be used within a larger filtering pipeline, in combination with other heuristics, such as language identification or length ratio comparison. Finally, our evaluation method also differs from the shared tasks. They evaluate the quality of a filtered corpus on the downstream machine translation performance, while we only compute accuracy scores.

3. Language pairs and datasets

3.1. Two language pairs

The main language we study is Upper Sorbian (ISO: `hsb`; Glottocode: `uppe1395`), a Slavic language spoken in eastern Germany. It is classified as endangered by Ethnologue (Eberhard et al., 2025) and is considered low-resource in NLP according to (Joshi et al., 2020), as it lies in the 'scraping-bys' (1) cluster.

The closest higher-resource language to Upper Sorbian is Czech (`cs`; `czec1258`), also a Slavic language. It is the official national language of the Czech Republic and thus thrives in both data and NLP tools in comparison.

As Upper Sorbian is a minority language in Germany, we pair it with German (`de`; `stan1295`; a Germanic language). This language pair was notably considered by successive editions of Machine Translation Shared Tasks at WMT (Fraser, 2020; Libovický and Fraser, 2021; Weller-Di Marco and Fraser, 2022; Okabe et al., 2025a). Hence, we focus on the following two language pairs in our

[1] https://github.com/TUM-NLP/lr-parallel-sentence-classification/

case study: Upper Sorbian–German (HSB–DE) for the low-resource scenario and Czech–German (CS–DE) for the high-resource setting.

3.2. Original corpus creation

The central goal is to assess how well standard classification models can distinguish parallel from non-parallel sentence pairs based on sentence representation. As such labelled datasets do not exist for our two pairs, we construct a balanced dataset with both categories. First, we combined parallel sentences drawn from existing corpora and generated non-parallel sentence pairs in equal proportion. To create this second category, we randomly match sentences from an unrelated monolingual German corpus. This widely used perturbation simulates a misalignment during pairing.

Dataset source For the Upper Sorbian–German pair, we used the bilingual training corpus[2] from the WMT 2022 Shared Task on Unsupervised MT and Very Low Resource Supervised MT (Weller-Di Marco and Fraser, 2022). The parallel corpus for Czech–German comes from the EUBookshop corpus[3] combined with the WikiMatrix corpus[4] in OPUS (Tiedemann, 2012).

To create the non-parallel sentence pairs, we randomly selected sentences from the German news 2024 monolingual dataset[5] in the Leipzig Corpora Collection (Goldhahn et al., 2012).

Pre-processing We assume both initial parallel corpora to be of high quality. Still, we pre-processed them to create our synthetic datasets. We filtered the sentences that were shorter than 10 words and that contained URLs or a large number of non-linguistic symbols (e.g., equations). We also removed duplicated pairs and pairs that have vastly differing lengths. Finally, we discarded non-standard characters and formatting artefacts to normalise the sentences.

3.3. Sentence transformation

Although the constructed datasets already contain both positive and negative sentence pairs, randomly paired sentences are often too simple for the classification task. Other methods to create noisy sentence pairs notably include shuffling words within a sentence (Xu and Koehn, 2017; Zhang et al., 2020), removing words or replacing them with words of similar frequency (Zaragoza-Bernabeu et al., 2022). We opt to create more challenging and realistic conditions by applying a set of five sentence transformation techniques inspired by (Chen et al., 2023).

We categorised the transformations into two main strategies: (i) modifying causal and logical relations within sentences, and (ii) replacing surface-level tokens. Each transformation follows a two-step procedure: first, candidate sentences are identified using rule-based linguistic filters, and second, a word-level change is applied to alter the sentence meaning to varying degrees. Only sentences that were actually modified are kept and labelled as non-parallel examples.

The transformations are applied to the German side of the actual parallel sentence pairs, while keeping the corresponding sentence in the other language unchanged, thereby constructing controlled non-parallel examples. German is chosen for transformation as it is a high-resource language and thus features reliable NLP tools.

All German sentences were processed using spaCy's German pipeline model (Honnibal et al., 2020) to perform tokenisation, Part-of-Speech tagging, and Named Entity Recognition (NER). These three annotations were essential to identify the words to add, remove, or replace in our five transformations. We notably needed to recognise adjectives, verbs, auxiliaries, modal verbs, entities, and numbers. Table 1 illustrates each transformation type with example sentences.

Antonyms transformation We manually compiled a list of 145 frequently used German antonym pairs based on standard German dictionaries and replaced selected words with their opposites. For example, '*gut*' (good) was replaced by '*schlecht*' (bad). The full list is provided in our GitHub repository.

Negation transformation We apply negation to the original German sentence by inserting the particle '*nicht*' (not) for positive sentences in appropriate positions. For instance, '*Das Wetter ist schön*' (The weather is nice) was transformed into '*Das Wetter ist **nicht** schön*' (The weather is not nice). Conversely, we also removed negation markers such as *kein* and its variants to revert sentences from negative to positive.

Modality transformation We created mapping pairs to change the modal verbs: for instance, the verb '*kann*' (can) becomes '*muss*' (must). As verbs are also inflected in German, we list all the possibilities in the conjugation for the modal verbs (e.g,

[2] `HSB-DE_train.tsv.gz`; primarily news domain https://github.com/mariondimarco/WMT22_UnsupVeryLowResMT_Data/tree/main.

[3] https://opus.nlpl.eu/datasets/EUbookshop?pair=cs&de; v2.

[4] https://opus.nlpl.eu/datasets/WikiMatrix?pair=cs&de; v1.

[5] https://corpora.uni-leipzig.de?corpusId=deu_news_2024.

transformation category	original sentence	transformed sentence
antonym transformation	Denke daran, was im Augenblick **wichtig** ist! *Remember what is **important** right now!*	Denke daran, was im Augenblick **unwichtig** ist! *Remember what is **unimportant** right now!*
negation transformation	Das ist ein gutes Vorzeichen für unsere künftigen Beziehungen *This is a good sign for our future relations.*	Das ist **kein** gutes Vorzeichen für unsere künftigen Beziehungen *This is **not** a good sign for our future relations.*
modality transformation	Ich **kann** genauso gut Ski fahren wie mein Bruder. *I **can** ski just as well as my brother.*	Ich **muss** genauso gut Ski fahren wie mein Bruder. *I **must** ski just as well as my brother.*
entity replacement	Unsere besten Freundinnen stammen aus **Neudorf**. *Our best friends come from **Neudorf***	Unsere besten Freundinnen stammen aus **Köln**. *Our best friends come from **Cologne***
number replacement	Ich bin nicht zu jung, ich bin **127** Jahre, **2** Monate und **22** Tage! *I'm not too young, I'm **127** years, **2** months and **22** days old!*	Ich bin nicht zu jung, ich bin **15** Jahre, **26** Monate und **63** Tage! *I'm not too young, I'm **15** years, **26** months and **63** days old!*

Table 1: Examples of transformation categories with original and transformed German sentences.

'*könnten*' with '*müssten*', another tense from the same verb pair). This ensures that the transformed sentences remain grammatically plausible.

Entity replacement We replaced three categories of named entities: people, locations, and organisations. We use the NER tool from spaCy for German. For each sentence, we convert the first occurrence of an entity with another name (or a combination) from a pre-defined set. For example, '*Berlin*' can be replaced with '*Prag*' (Prague).

Number replacement Similarly, we replace numerical values with other numbers from a predefined set. This also holds for spelt numbers or numerals to have a similar yet different quantity on the German side. For example, a sentence containing the year *2020* could be changed to *2038*.

3.4. Final datasets

After applying all sentence transformation and preprocessing steps, we obtain four datasets for our experiments, which differ in the nature of the non-parallel pairs. We call the first type of dataset with randomly matched sentences, original (HSB–DE and CS–DE, respectively), while the transformed variant will be denoted with a trailing _TR (HSB–DE_TR and CS–DE_TR). The complexity induced by linguistic rules for word-level transformation constitutes the main restriction on the dataset size.

Table 2 displays the total number of sentences. We split each dataset into training (80%), validation (10%), and test (10%) sets. We ensure that all sets contain an equal proportion of both labels. Each transformed dataset consists of 15,000 sentence pairs for the entity and number replacement transformations, and 10,000 sentence pairs each for the antonym, modality, and negation transformations. The difference in number stems from the fewer available candidate sentences for these three categories.

dataset	parallel	non-parallel	total
HSB–DE	60,000	60,000	120,000
HSB–DE_TR	60,000	60,000	120,000
CS–DE	60,000	60,000	120,000
CS–DE_TR	60,000	60,000	120,000

Table 2: Number of sentences in the final datasets.

4. Methodology

We present the classification pipeline in Figure 1. First, we represent sentence pairs in the same multilingual space using pre-trained language models. We then create a feature vector for a sentence pair using dimensionality reduction and computing the cosine similarity. Finally, we train classification models to predict the binary label.

4.1. Language models

We evaluate three multilingual language models. XLM-RoBERTa base or XLM-R (Conneau et al., 2020) is a multilingual pre-trained language model covering more than 100 languages. Glot500m (Imani et al., 2023) extends XLM-R with a training over 500 languages, specifically covering low-resource ones. We also consider a state-of-the-art sentence encoder, LaBSE (Feng et al., 2022),

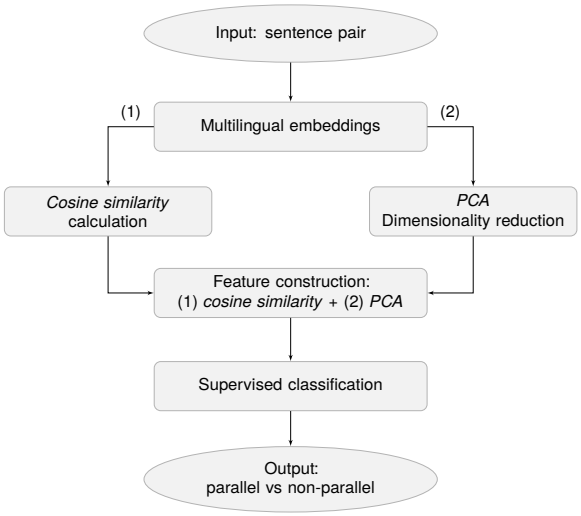


Figure 1: Base pipeline for parallel sentence filtering as a classification task.

which is optimised for the sentence similarity and translation retrieval tasks.

For XLM-R and Glot500m, sentence-level representations are obtained by mean pooling over the final-layer token embeddings. For LaBSE, we directly use the provided sentence-level embeddings. We note here that, while German and Czech were seen by all three models during their pre-training, only Glot500m was trained on Upper Sorbian.

4.2. Dimensionality reduction

As the sentence embeddings are large (768 for all three models), we normalised and reduced them using Principal Component Analysis (PCA). We perform dimensionality reduction here to decrease the computational complexity for downstream supervised classifiers, while focusing on the most significant components in the multilingual space.

We choose to retain 95% of the original variance and present the final number of dimensions in Table 3. We note that applying PCA leads to vectors of around 200 dimensions in most cases.

4.3. Classification setting

Sentence pair representation For each sentence pair, we construct one feature vector F by concatenating the embeddings in the two languages after applying PCA and their cosine similarity, as in Equation (1) for HSB–DE:

$$F = [\mathrm{PCA}(\vec{v}_{\mathrm{HSB}}) \,\|\, \mathrm{PCA}(\vec{v}_{\mathrm{DE}}) \,\|\, \mathrm{sim}(\vec{v}_{\mathrm{HSB}}, \vec{v}_{\mathrm{DE}})] \quad (1)$$

For example, if we consider the PCA component numbers for XLM-R reported in Table 3, the embeddings are reduced to 234 dimensions for Upper Sorbian and 262 for German, yielding a 234+262+1 = 497-dimensional feature vector including the cosine similarity score.

model	HSB–DE		CS–DE	
	HSB	DE	CS	DE
original datasets				
XLM-R	234	262	267	272
Glot500m	212	233	275	266
LaBSE	171	196	181	195
transformed datasets				
XLM-R	235	256	260	262
Glot500m	177	200	205	260
LaBSE	169	185	186	200

Table 3: The number of PCA components retained (95% variance) for each language model and corpus.

Classifiers We evaluate six standard classifiers implemented in the scikit-learn library (Pedregosa et al., 2011): Logistic Regression, Random Forest, LightGBM, Support Vector Machine (SVM), XGBoost, and Multi-Layer Perceptron (MLP). We fine-tune the hyperparameters on the HSB–DE dataset and keep them fixed across all subsequent datasets and experimental settings. Table 4 displays the final values of hyperparameters for the classifiers.

classifier	parameters
Logistic Regression	max_iter = 1000
Random Forest	n_estimators = 500 n_jobs = −1
LightGBM	n_estimators = 500 learning_rate = 0.05
Linear SVM	max_iter = 1000 dual = True
XGBoost	n_estimators = 500 learning_rate = 0.05
MLP	hidden_layers = (128, 64) max_iter = 400

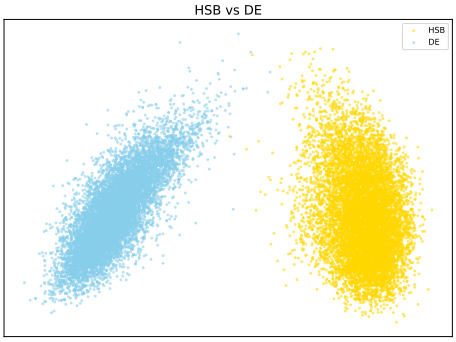
Table 4: Classifier configurations tuned on the validation split of the HSB–DE dataset.

4.4. Evaluation method

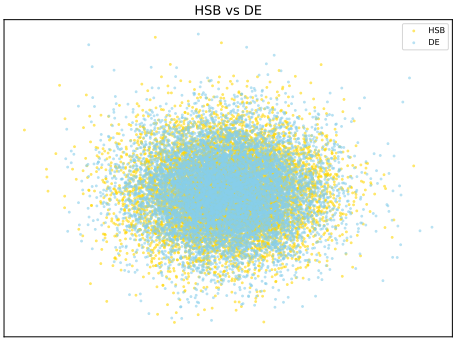
We evaluated all the experiments with the usual Accuracy, Precision, Recall, F1 score, and the Area Under the ROC Curve (ROC-AUC) since we have a standard classification task with balanced datasets. However, we only report accuracy, which we considered to be the primary evaluation metric in the

main part of the article. Experimental results are displayed in Appendix A with all five metrics.

4.5. Cluster-based isotropy enhancement



(a) Glot500m



(b) Glot500m with CBIE

Figure 2: Parallel sentences in HSB and DE with Glot500m before and after using CBIE.

We display 10,000 parallel sentence pairs (20,000 sentences in total) for the HSB–DE language pair using the PCA representation obtained with Glot500m in Figure 2. The two languages are represented separately with the model; with LaBSE (in Appendix B) we only see a small overlap between the two clusters. Both seem to struggle to match the two languages to a varying degree and hence present anisotropy.

Cluster-Based Isotropy Enhancement (CBIE) is known for tackling this issue for a set of multilingual sentence representations (Rajaee and Pilehvar, 2021). First, it clusters the dataset, then applies PCA to each, and removes the top principal components. Hämmerl et al. (2023) have shown that CBIE substantially improves performance on cross-lingual similarity and retrieval tasks. Okabe et al. (2025b) also found that this transformation noticeably helps increase parallel sentence mining quality for low-resource pairs, including Upper Sorbian–German. As shown in Figure 2, CBIE significantly improves the alignment between HSB–DE embeddings with overlapping distributions for Glot500m.

Hence, we apply the CBIE post-processing to the sentence embeddings of the *transformed* dataset (_TR) before dimensionality reduction. Subsequently, we perform 95% PCA on the CBIE-adjusted representations when constructing the feature vectors for the sentence pairs. As the process depends on the dataset, we apply it separately per language and split.

4.6. Cross-lingual setting

As Upper Sorbian and Czech are related, we also evaluate how effective a classifier trained on a better-resourced and related language pair can be for a low-resource pair. We examine cross-lingual transfer between the Upper Sorbian–German and Czech–German datasets in both directions. In the high-to-low setting, we train the classification model on the CS–DE_TR dataset and evaluate directly on the HSB–DE_TR test set (CS → HSB). The low-to-high (HSB → CS) does the opposite. The former enables us to answer our third research question, while the latter, although not realistically needed, aims to assess the extent to which the representation of the low-resource language impacts the classification quality.

5. Results

5.1. Preliminary experiments

We first present our experimental choices for the trained machine learning models and the classification approach.

5.1.1. Classifier comparison

classifiers	XLM-R	Glot500m	LaBSE
Logistic Reg.	78.90	74.72	99.66
Rand. Forest	86.14	89.11	98.37
LightGBM	92.14	95.34	99.50
Linear SVM	89.66	96.19	99.67
XGBoost	92.41	95.25	99.52
MLP	**97.43**	**98.20**	**99.68**

Table 5: Comparison of the six classifiers across the three language models on the HSB–DE dataset.

Table 5 compares the results of the six classifiers on the original HSB–DE dataset. We note that the sentence representations from LaBSE are the most robust overall, while the mean-pooled representations lag further behind. On average, Glot500m leads to better accuracy values than XLM-R, thanks to its pre-training on Upper Sorbian. Besides, all three language models achieve their highest accuracy with the MLP classifier. Hence, we only report

the results of MLP in the remaining sections; full results are presented in Appendix A.

5.1.2. Filtering pipeline choice

model	cos. similarity	PCA	PCA+cos+MLP
XLM-R	54.73	62.72	**97.43**
Glot500	61.86	73.60	**98.20**
LaBSE	94.35	97.44	**99.68**

Table 6: Comparison of the three embedding models across three different experimental settings on the HSB–DE dataset: cosine similarity baseline, PCA without classifier, and PCA with MLP (our base pipeline).

We compare three filtering settings on the HSB–DE dataset to assess the effects of two components in our pipeline: the PCA and the supervised classifier. The first setting uses the raw sentence embeddings of the models and directly computes the cosine similarity for the final decision, without employing any classifier. The second approach carries out PCA on the embeddings and then performs classification using cosine similarity and a threshold. These two methods, hence, represent an unsupervised approach, with no trained classifier. The threshold is selected here to maximise filtering quality on the validation split. The last setting corresponds to our base pipeline (cf. Figure 1), which combines PCA representations with cosine similarity that are fed into a supervised classifier for binary classification.

Table 6 shows that the supervised approach (third column) performs the best across classification settings, as expected. Moreover, LaBSE gives the most robust representation, reaching accuracies above 90% consistently for all three pipeline settings. We note that although using PCA reduced dimensionality, it improved the overall performance on our dataset. Using the classifier further enhanced the results, even compensating for the weaker sentence representation of XLM-R and Glot500. These results confirm the utility of the components in our base pipeline (third setting).

5.2. Monolingual setting

5.2.1. Original vs transformed datasets

Table 7 presents the performance of the three language models with our base pipeline across the four datasets: the original and transformed versions of HSB–DE and CS–DE. The results show that all models achieve high accuracy scores on the original datasets (above 98%), while performance drops noticeably on the transformed sentence pairs. In the latter case, Glot500m outperforms the other two models in both language pairs. It sees the smallest decrease in accuracy (around 9–15%) compared to LaBSE (more than 20%), which was the strongest option when the sentences were randomly matched.

model	HSB–DE		CS–DE	
	origin.	transf.	origin.	transf.
XLM-R	97.43	87.60	98.98	80.10
Glot500m	98.20	**89.98**	98.80	**83.71**
LaBSE	**99.68**	78.37	**99.32**	74.93

Table 7: Classification accuracy with and without sentence transformation on the HSB–DE and CS–DE datasets.

We note that the representations for both language pairs seem sufficient to carry out our classification task, with high accuracies in the simple and challenging scenarios. In the latter case, pretraining on Upper Sorbian seems to help achieve a more robust representation. This answers RQ1.

5.2.2. Details per transformation

transformation	XLM-R	Glot500m	LaBSE
HSB–DE_TR			
antonym	71.52	**75.94**	68.25
negation	96.57	**96.72**	80.40
modality	**95.07**	94.70	71.17
entity	75.46	83.76	**91.10**
number	95.03	**95.96**	83.36
CS–DE_TR			
antonym	80.74	**84.97**	75.90
negation	81.05	**85.14**	76.01
modality	80.99	**85.08**	75.92
entity	80.06	**84.29**	74.94
number	80.10	**84.36**	75.05

Table 8: Accuracy (%) of language models across the five sentence transformations on **HSB–DE** and **CS–DE**. Best results per transformation are in **bold**.

We further investigate how difficult each transformation type is for the three language models in Table 8. We observe that for the high-resource language pair, all transformations seem to affect the classification performance similarly, while the accuracy values fluctuate more with the HSB–DE pair. Some language models are more robust to certain perturbations: LaBSE gives the most reliable sentence representation on the *entity* replacement category, while Glot500m outperforms it for the other modifications. For instance, on the *modality* transformation, it remains at 95%, which is around 24%

above LaBSE's score. We suppose that the named entities are more challenging, as they may be represented similarly by word encoders that were not explicitly trained to discriminate them. Therefore, classification performance is impacted by our word replacements, and low-resource languages are more prone to variability, for which pre-training (cf. Glot500m) can again help. This answers RQ2.

5.2.3. CBIE results (monolingual)

model	HSB–DE_TR	CS–DE_TR
Glot500m	69.74 (↓20.24)	63.32 (↓20.39)
LaBSE	58.18 (↓20.19)	58.34 (↓16.89)

Table 9: Accuracy with CBIE transformation on the transformed HSB–DE and CS–DE datasets.

We assess the impact of anisotropy of the multilingual representation on the classification quality by applying CBIE to sentence representations. Since XLM-R and Glot500m gave similar trends, we only present the results of Glot500m alongside LaBSE. Table 9 shows that although we know CBIE generally improves the language alignment, it does not help the parallel sentence classification quality. All accuracy scores drop significantly by around 20% on the transformed datasets. We hypothesise that reduced anisotropy makes the task harder, because the classifier actually relied on imperfect language alignment.

5.3. Cross-lingual setting

5.3.1. Cross-lingual transfer

model	CS → HSB	HSB → CS
XLM-R	80.64 (↓6.96)	74.27 (↓5.83)
Glot500m	**84.11** (↓5.87)	**78.16** (↓5.55)
LaBSE	69.73 (↓8.64)	68.62 (↓6.31)

Table 10: Cross-lingual performance of the classifiers on the *transformed* dataset without re-training in both directions.

In the cross-lingual setting, we train the classification model on a language pair (e.g., CS–DE) and evaluate it on the other language pair (e.g., HSB–DE) without re-training. Table 10 compares the classification performance on the transformed dataset in both directions of language transfer. Glot500m is the most robust model in both directions of this setting, as it maintains its edge against the other two models. We observe that all accuracy scores are lower than in the (simpler) monolingual setting, but the representation obtained with Glot500m is the least affected. This seems to be thanks to its pre-training on Upper Sorbian, which helps with cross-lingual transfer in the classification task. Although the HSB → CS direction is purely experimental, the decrease in accuracy is of comparable level in all settings. This seems to suggest that poorer language representation quality did not greatly affect classification performance, probably because of the quality of the Czech representation. To answer RQ3, a reliable approach seems to be to use a language model pre-trained on the low-resource language (as in Glot500m) when training a classifier only on a high-resource and related language pair.

5.3.2. Details per transformation

transformation	XLM-R	Glot500m	LaBSE
high to low: CS → HSB			
antonym	58.42	54.09	**62.50**
negation	92.83	**93.65**	76.93
modality	**93.23**	90.95	61.63
entity	66.60	66.13	**79.83**
number	**90.47**	89.93	67.87
low to high: HSB → CS			
antonym	76.16	**78.98**	69.30
negation	76.59	**78.99**	69.13
modality	76.36	**78.85**	69.10
entity	76.68	**78.81**	67.87
number	76.69	**78.91**	69.35

Table 11: Accuracy (%) of language models across various sentence transformations on HSB → CS and CS → HSB. Best results per transformation are in **bold**.

Table 11 presents the classification accuracy per transformation. As in the monolingual setting, the variance is high when the (test) language is low-resource, compared to a high-resource pair. We notably see that Glot500m is the best approach on the HSB → CS direction for all five transformation types. Besides, accuracy remains within the same range for a given model, suggesting that the word-level perturbations do not impact the classification performance more in the cross-lingual setting.

In the (main) opposite direction, Glot500m loses its edge against the other two models for certain transformations, but remains the overall best approach. It is indeed the only language model pre-trained on all three languages involved. We notably see a larger drop for the antonym transformation comparatively. Besides, LaBSE remains the best-performing model for the entity transformation, as in the monolingual setting.

5.3.3. CBIE results (cross-lingual)

model	CS → HSB	HSB → CS
Glot500m	58.62 (↓25.49)	64.20 (↓13.96)
LaBSE	57.87 (↓11.86)	55.74 (↓12.88)

Table 12: Cross-lingual classification performance with CBIE transformation applied on the transformed HSB–DE_TR and CS–DE_TR datasets.

As in the previous section, we also evaluate the influence of CBIE in the cross-lingual setting in Table 12, when applied to both the training and test sets. We observe the same trend as in the monolingual experiment: CBIE transformation does not help improve the classification performance and leads to a decrease in accuracy for all settings. The drop, however, is smaller in the cross-lingual setting, especially for HSB → CS. Only Glot500m on the CS → HSB direction suffers more than in the monolingual setting; the pre-training alone does not seem to suffice here.

6. Conclusion

We considered parallel sentence filtering as a classification task, where we train a model to distinguish parallel from non-parallel sentence pairs. We created synthetic balanced corpora from existing parallel sentences in two non-English-centric language pairs: Upper Sorbian–German and Czech–German. Beyond randomly matching sentences, we also applied five types of word-level transformations to sentences to obtain non-parallel pairs.

We find that the multilingual models we choose can correctly distinguish parallel from clearly non-parallel sentence pairs (original dataset), but struggle with the dataset we synthetically corrupt (transformed dataset). We note that the drop in performance can be slightly mitigated by a better language representation, i.e., with dedicated pre-training, as it was the case with Glot500m on HSB–DE_TR. Yet, without a stronger encoding capability (as in LaBSE), the classification task is more prone to variability depending on the injected ambiguity. Besides, CBIE, which is known for improving multilingual alignment, did not help the classifier and actually made the task harder.

Furthermore, we explored a cross-lingual transfer scenario for the classifier, notably training it on the high-resource pair and applying it to the low-resource pair. Despite the drop in accuracy, we see that classification performance remains relatively high, even on the more complex transformed dataset. Here again, pre-training on the low-resource language proved helpful.

Future work naturally includes an extension to more languages to assess the stability of our results. The cross-lingual setting also enables us to carry out parallel sentence filtering on pairs where classifiers cannot be directly trained due to a lack of data. Related languages from the same language family (as it was the case for Upper Sorbian and Czech) are the straightforward option in such cases; we could also rely on locally dominant and better-resourced languages.

7. Limitations

Our main limitation comes from the small language coverage, as we focused on only one low-resource (HSB–DE) and one better-resourced (CS–DE) language pairs. Although this restricts the scope of the findings, we aimed to evaluate the language representation along with the classification capacity in a controlled environment. We hope that this work will foster further studies on a broader range of languages to assess the robustness of our results. A related point is that the classification accuracy is bound by the datasets we chose: due to data scarcity on the low-resource side, we could not test over varying domains. Nevertheless, the cross-lingual setting features a domain mismatch in terms of both language and content.

Moreover, our experiments only rely on synthetic noise, primarily targeting semantic and logical consistency. This setting may not fully reflect the diversity of noise in real-world mined corpora. Yet, as the difference lies in one word within the full sentence, it represents a more subtle and, hence, challenging setup. Additionally, while we demonstrate high classification accuracy on our synthetic dataset, the direct impact of this filtering on downstream Machine Translation (MT) performance remains to be quantified.

Furthermore, this study focused on standard classifiers, while more advanced models exist. Our goal was to observe both the quality of the representation of a language pair and the classification performance itself. Our base pipeline results from a trade-off from that perspective. Besides, for well-resourced pairs, such as Czech–German, more data could have been considered for training to evaluate their robustness across domains and languages.

8. Acknowledgements

We thank the anonymous reviewers for their insightful comments. This work was funded by the European Research Council (ERC) under grant agreements No. 101113091 - Data4ML (a Proof of Concept Grant) and No. 101141712 - EPICAL. Views and opinions expressed are, however, those

of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

9. Bibliographical References

Mikel Artetxe and Holger Schwenk. 2019a. Margin-based parallel corpus mining with multilingual sentence embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3197–3203, Florence, Italy. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2019b. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Vishrav Chaudhary, Yuqing Tang, Francisco Guzmán, Holger Schwenk, and Philipp Koehn. 2019. Low-resource corpus filtering using multilingual sentence embeddings. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 261–266, Florence, Italy. Association for Computational Linguistics.

Mingda Chen, Kevin Heffernan, Onur Çelebi, Alexandre Mourachko, and Holger Schwenk. 2023. xSIM++: An improved proxy to bitext mining performance for low-resource languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 101–109, Toronto, Canada. Association for Computational Linguistics.

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

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2025. *Ethnologue: Languages of the World*, twenty-eighth edition. SIL International, Dallas, Texas. Online version.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Alexander Fraser. 2020. Findings of the WMT 2020 shared tasks in unsupervised MT and very low resource supervised MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 765–771, Online. Association for Computational Linguistics.

Katharina Hämmerl, Alina Fastowski, Jindřich Libovický, and Alexander Fraser. 2023. Exploring anisotropy and outliers in multilingual language models for cross-lingual semantic sentence similarity. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7023–7037, Toronto, Canada. Association for Computational Linguistics.

Viktor Hangya and Alexander Fraser. 2019. Unsupervised parallel sentence extraction with parallel segment detection helps machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1224–1234, Florence, Italy. Association for Computational Linguistics.

Kevin Heffernan, Onur Çelebi, and Holger Schwenk. 2022. Bitext mining using distilled sentence representations for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2101–2112, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Ayyoob Imani, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André Martins, François Yvon, and Hinrich Schütze. 2023. Glot500: Scaling multilingual corpora and language models to 500 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1082–1117, Toronto, Canada. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Philipp Koehn, Vishrav Chaudhary, Ahmed El-Kishky, Naman Goyal, Peng-Jen Chen, and Francisco Guzmán. 2020. Findings of the WMT 2020 shared task on parallel corpus filtering and alignment. In *Proceedings of the Fifth Conference on Machine Translation*, pages 726–742, Online. Association for Computational Linguistics.

Philipp Koehn, Francisco Guzmán, Vishrav Chaudhary, and Juan Pino. 2019. Findings of the WMT 2019 shared task on parallel corpus filtering for low-resource conditions. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 54–72, Florence, Italy. Association for Computational Linguistics.

Philipp Koehn, Huda Khayrallah, Kenneth Heafield, and Mikel L. Forcada. 2018. Findings of the WMT 2018 shared task on parallel corpus filtering. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 726–739, Belgium, Brussels. Association for Computational Linguistics.

Jindřich Libovický and Alexander Fraser. 2021. Findings of the WMT 2021 shared tasks in unsupervised MT and very low resource supervised MT. In *Proceedings of the Sixth Conference on Machine Translation*, pages 726–732, Online. Association for Computational Linguistics.

Peiqin Lin, Andre Martins, and Hinrich Schuetze. 2025. A recipe of parallel corpora exploitation for multilingual large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4038–4050, Albuquerque, New Mexico. Association for Computational Linguistics.

Shu Okabe, Daryna Dementieva, Marion Di Marco, Lukas Edman, Katharina Haemmerl, Marko Měškank, Anita Hendrichowa, and Alexander Fraser. 2025a. Findings of the WMT 2025 shared task LLMs with limited resources for Slavic languages: MT and QA. In *Proceedings of the Tenth Conference on Machine Translation*, pages 503–519, Suzhou, China. Association for Computational Linguistics.

Shu Okabe, Katharina Hämmerl, and Alexander Fraser. 2025b. Improving parallel sentence mining for low-resource and endangered languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 196–205, Vienna, Austria. Association for Computational Linguistics.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85):2825–2830.

Sara Rajaee and Mohammad Taher Pilehvar. 2021. A cluster-based approach for improving isotropy in contextual embedding space. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 575–584, Online. Association for Computational Linguistics.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online. Association for Computational Linguistics.

Steve Sloto, Brian Thompson, Huda Khayrallah, Tobias Domhan, Thamme Gowda, and Philipp Koehn. 2023. Findings of the WMT 2023 shared task on parallel data curation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 95–102, Singapore. Association for Computational Linguistics.

Weiting Tan, Kevin Heffernan, Holger Schwenk, and Philipp Koehn. 2023. Multilingual representation distillation with contrastive learning. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1477–1490, Dubrovnik, Croatia. Association for Computational Linguistics.

Marion Weller-Di Marco and Alexander Fraser. 2022. Findings of the WMT 2022 shared tasks in unsupervised MT and very low resource supervised MT. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 801–805, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Hainan Xu and Philipp Koehn. 2017. Zipporah: a fast and scalable data cleaning system for noisy web-crawled parallel corpora. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2945–2950, Copenhagen, Denmark. Association for Computational Linguistics.

Jaume Zaragoza-Bernabeu, Gema Ramírez-Sánchez, Marta Bañón, and Sergio Ortiz Rojas. 2022. Bicleaner AI: Bicleaner goes neural.

In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 824–831, Marseille, France. European Language Resources Association.

Boliang Zhang, Ajay Nagesh, and Kevin Knight. 2020. Parallel corpus filtering via pre-trained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8545–8554, Online. Association for Computational Linguistics.

Pierre Zweigenbaum, Serge Sharoff, and Reinhard Rapp. 2017. Overview of the second BUCC shared task: Spotting parallel sentences in comparable corpora. In *Proceedings of the 10th Workshop on Building and Using Comparable Corpora*, pages 60–67, Vancouver, Canada. Association for Computational Linguistics.

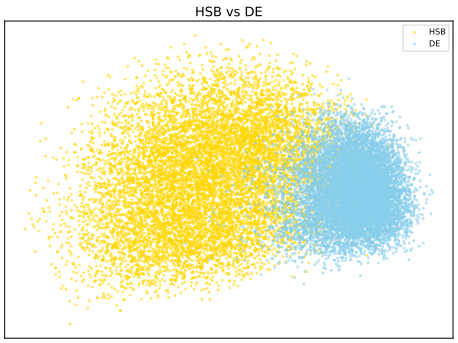
Pierre Zweigenbaum, Serge Sharoff, and Reinhard Rapp. 2018. Overview of the third BUCC shared task: Spotting parallel sentences in comparable corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

10. Language Resource References

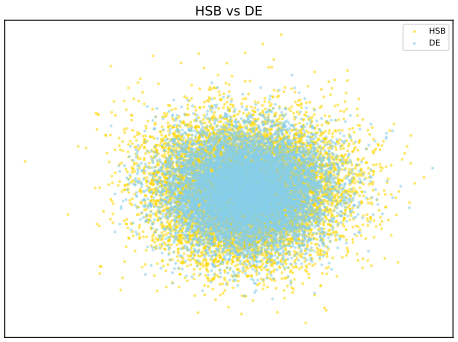
Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the Leipzig corpora collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 759–765, Istanbul, Turkey. European Language Resources Association (ELRA).

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Marion Weller-Di Marco and Alexander Fraser. 2022. Findings of the WMT 2022 shared tasks in unsupervised MT and very low resource supervised MT. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 801–805, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.



(a) LaBSE



(b) LaBSE with CBIE

Figure 3: Parallel sentences in HSB and DE with LaBSE before and after using CBIE transformation.

A. Full classification results

Tables 13 and 14 present the full classification results on the HSB–DE and HSB–DE_TR datasets for all six classifiers and five evaluation metrics.

Table 15 displays the full classification results on the cross-lingual experiments (HSB → CS and CS → HSB) with four classifiers and all five metrics.

B. CBIE transformation for LaBSE

Figure 3 displays the visualisation for parallel sentences in Upper Sorbian and German obtained with LaBSE, before and after appling CBIE.

Model	Classifier	Accuracy	Precision	Recall	F1	ROC AUC
XLM-R	LogReg	78.91	74.30	88.40	80.74	88.26
	RF	86.14	83.89	89.47	86.59	93.95
	LGBM	92.14	91.34	93.12	92.22	97.84
	SVM	89.67	87.92	91.97	89.90	96.28
	XGB	92.42	91.93	93.00	92.46	97.83
	MLP	97.43	96.92	97.98	97.45	99.65
Glot500m	LogReg	74.73	68.97	89.88	78.05	85.15
	RF	89.12	87.22	91.67	89.39	95.81
	LGBM	95.34	94.66	96.10	95.38	99.13
	SVM	96.19	95.58	96.87	96.22	99.27
	XGB	95.26	94.63	95.97	95.29	99.02
	MLP	98.21	97.41	99.05	98.22	99.75
LaBSE	LogReg	99.67	99.60	99.73	99.67	100.00
	RF	98.38	98.19	98.57	98.38	99.90
	LGBM	99.51	99.58	99.43	99.51	99.99
	SVM	99.68	99.62	99.73	99.68	99.99
	XGB	99.53	99.55	99.50	99.52	99.99
	MLP	99.68	99.63	99.72	99.68	99.99

Table 13: Full classification results of Table 5 (on the HSB–DE dataset) for the six classifiers.

Model	Classifier	Accuracy	Precision	Recall	F1	ROC AUC
XLM-R	LogReg	80.31	73.99	93.46	82.59	80.31
	RF	77.44	71.26	91.96	80.30	77.44
	LGBM	84.37	79.11	93.40	85.67	84.37
	SVM	82.58	77.07	92.75	84.19	82.58
	XGB	84.58	79.83	92.53	85.71	84.58
	MLP	87.60	83.71	93.38	88.28	87.60
Glot500m	LogReg	82.00	74.86	96.37	84.26	82.00
	RF	80.61	75.46	90.73	82.39	80.61
	LGBM	86.56	82.85	92.20	87.38	86.56
	SVM	85.27	79.47	95.12	86.59	85.27
	XGB	86.39	81.15	94.79	87.40	86.39
	MLP	89.98	87.67	93.05	90.28	89.98
LaBSE	LogReg	71.51	68.31	80.23	73.80	71.51
	RF	69.83	67.52	76.41	71.69	69.83
	LGBM	74.23	72.16	78.90	75.38	74.23
	SVM	69.75	66.52	79.51	72.44	69.75
	XGB	74.72	72.98	78.52	75.65	74.72
	MLP	78.37	76.31	82.28	79.18	78.37

Table 14: Classification results on the HSB–DE_TR dataset with the six classifiers.

Model	Classifier	Accuracy	Precision	Recall	F1	ROC AUC
XLM-R	LGBM	74.78	75.94	72.53	74.20	74.78
	SVM	75.43	75.08	76.12	75.59	75.43
	XGB	74.81	76.28	72.00	74.08	74.81
	MLP	74.27	79.64	65.20	71.70	74.27
Glot500m	LGBM	78.71	76.91	82.05	79.40	78.71
	SVM	79.73	76.81	85.17	80.77	79.73
	XGB	78.84	75.07	86.37	80.32	78.84
	MLP	78.16	76.54	81.20	78.80	78.16
LaBSE	LGBM	65.54	63.44	73.35	68.04	65.54
	SVM	65.63	69.46	55.77	61.87	65.63
	XGB	65.85	64.38	70.97	67.51	65.85
	MLP	68.62	72.33	60.30	65.77	68.62

(a) HSB → CS

Model	Classifier	Accuracy	Precision	Recall	F1	ROC AUC
XLM-R	LGBM	80.67	74.65	92.85	82.77	80.67
	SVM	78.81	72.04	94.17	81.63	78.81
	XGB	79.93	73.31	94.12	82.42	79.93
	MLP	80.64	74.86	92.26	82.66	80.64
Glot500m	LGBM	84.15	80.57	90.00	85.02	84.15
	SVM	82.12	83.71	79.77	81.69	82.12
	XGB	83.95	78.87	92.76	85.25	83.95
	MLP	84.11	79.43	92.04	85.27	84.11
LaBSE	LGBM	67.85	63.86	82.24	71.89	67.85
	SVM	65.25	61.26	82.98	70.48	65.25
	XGB	66.89	62.39	85.02	71.97	66.89
	MLP	69.73	64.87	86.06	73.97	69.73

(b) CS → HSB

Table 15: Full classification performance on the HSB → CS and CS → HSB datasets using three language models and four classifiers.