Fairness in Multilingual Sentiment Analysis: A Comparative Study of BERT, DistilBERT, and XLM-RoBERTa

Paul Clotan (paulioan.clotan@studio.unibo.it)
Valeria Izvoreanu (valeria.izvoreanu@studio.unibo.it)
Antonio Gravina (antonio.gravina@studio.unibo.it)

June 2025

1 Abstract

In this project, we explore fairness in multilingual sentiment analysis by comparing three widely-used transformer-based models: bert-base-multilingual-cased (Google) [1], DistilBERT-multilingual-cased (HuggingFace) [2], and XLM-RoBERTa-base (Meta) [3] on the Amazon Multilingual Reviews Corpus (MARC [4]). We will evaluate, compare and try to improve their performance through techniques such as language specific fine-tuning and majority voting ensembling. Our goal is identifying potential biases in models and comparing the results across the different techniques, languages and model employed. Our results are surprising, with the ensembling technique performing the worse (propagating the biases of the components).

2 Introduction

Since the introduction of the transformer architecture, AI research is mainly focused on achieving the best performances on tasks related to Natural Language Understanding (NLU). Usually, the results are evaluated through metrics like accuracy or precision, ignoring the possibility of biases and unfairness across different languages.

The potential discriminatory behavior of LLMs often comes from the presence of biases in the training corpora. Thus, we want to assess and compare the unfairness of three widely used encoder-only multilingual models (presented in Section 5.5), fine-tuned on MARC for a classification task. MARC contains reviews from Amazon users in several languages, which we separate in Negative and Positive classes (as described in Section 5.2). Our analysis focuses on assessing difference of performances of the three models (in terms of fairness metrics, see Section 5.4) across English, German and Spanish reviews. After that, we try to mitigate unfairness using techniques such as language-specific fine-tuning and ensembling.

The encoder-only models that we selected are all BERT based, which uses the attention mechanism to understand the context of the words in a bidirectional way. The pre-training of BERT, performed by Google AI Language team, was carried out with two different goals, which are Masked Language Modeling and Next Sentence Prediction. The pre-trained model can then be fine-tuned for specific tasks, like sentiment analysis. In fact, due to its impressive performances and its open-source nature, BERT is currently a SOTA language model for NLP/NLU tasks, and many architectures are derived from it, like a multilingual version on BERT, trained on Wikipedia in all the available languages, and DistilBERT, a lighter, more efficient model that is trained via

knowledge distillation. By comparing the fairness metrics of these two models we can check if the distillation process may lead to increased unfair behavior. In fact, if the teacher model exhibits unfair or biased behavior, the distilled model will likely inherit and potentially amplify those biases. A further comparison is carried out with XLM-RoBERTa, trained on a different corpus and optimized for cross-lingual understanding.

Before the actual tuning of the models, we performed data preprocessing and analysis (Section 5.3) to have a better understanding of the corpus we use and check the presence of biases in MARC.

3 Work Contribution

Antonio Gravina

- Worked on data analysis.
- Performed fine-tuning and training on BERT.
- Performed fine-tuning and training on Microsoft DeBERTa Multilingual (discarded due to computational power limitations).
- Developed and evaluated an ensemble model for English language data.

Paul-Ioan Clotan

- Designed and implemented the dataset preprocessing pipeline.
- Developed the pipeline for model training.
- Fine-tuned and trained DistilBERT.
- Developed and evaluated an ensemble model for Spanish language data.

Valeria Izvoreanu

- Led the analysis of metrics to evaluate fairness of the model.
- Fine-tuned and trained XLM-RoBERTa-base.
- Explored optimization strategies for the training pipeline to reduce computational resource usage associated with XLM-RoBERTa-base.
- Developed and evaluated an ensemble model for German language data.

4 Objectives

- Assess the fairness of three multilingual Transfomers models: DistilBERT-multilingual-cased (Hugging Face), XLM-RoBERTa-base (Meta), and Bert-base-multilingual-cased (Google) in the context of sentiment analysis.
- Focus on three widely spoken languages: English, German, and Spanish, by evaluating potential model biases in each language, and see how we can minimize the unfairness in those models.

- Evaluate disparities in error rates using the SAPMOC metrics [5], focusing on fairness beyond overall accuracy.
- Examine how model architecture might affect fairness outcomes, as well as differences between standard and distilled BERT models.
- Explore potential strategies to mitigate bias, such as language-specific finetunning, and implementing a voting-based ensemble system.

5 Methodology

5.1 Dataset

In this study case we use the Multilingual Amazon Reviews Corpus dataset. The data used in our study case was first released in this paper [6]. It contains Amazon reviews for multilingual text classification in English, German, French, Japanese, Spanish and Chinese. For our study, we will select three languages, namely Spanish, German and English. Among the data fields present in the dataset, we are interested in the following columns:

- Language: it tells us the language of the review.
- Stars: it reflect the satisfaction of the reviewer.
- Review title: Contains title of the review.
- Review body: Contains the feedback regarding the product.
- Product category: It contains the object category to which the bought product belongs to (ex: Kitchen).

5.2 Data preprocessing

This process starts by eliminating the irrelevant columns, and the languages we are not interested in. Namely, after dropping the irrelevant columns, we proceed by retaining only the reviews that are in one of the following languages: German, Spanish and English.

The next step is creating a classification label using feature engineering. This is done by using the Stars column, which contains numerical values from 1 to 5, 1 being the worst grade, and 5 being the best. Using this column, we create the target feature, which for readability purposes will be named *Sentiment*. The Sentiment is negative if the star grade of the review is either one or two, or positive if the review is four or five. We choose to drop the reviews that have a grade of 3 in order to eliminate as much uncertainty as possible from the data, as the three stars can often mean a mediocre review, which will only confuse the models.

The last step is to check the dataset for null values. 0.00008% of them are identified in the English language, for the training split, which are dropped. This percentage is considered insignificant.

At the end, we have a dataset that has reviews for English, Spanish and German, with a dedicated split for training, validation and testing.

5.3 Data analysis

We divided the training set into four different instances:

- General dataset, which is the original split of the dataset containing all the three languages.
- English dataset, containing only English reviews.
- German dataset, containing only German reviews.
- Spanish dataset, containing only Spanish reviews.

We observe in Figure 1 that the number of samples is perfectly balanced across the three languages (around 160k per each). The distribution of the labels is also balanced, with half of the samples labelled as Negative and the other half as Positive.

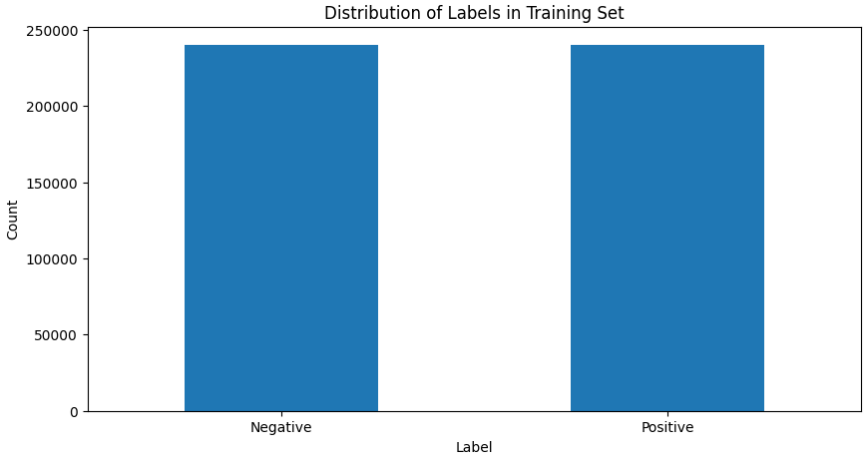

Figure 1: Distribution of the labels.

When we instead look at the distribution of the labels per product category in Figure 2, we notice that *home* and *wireless* have a significant higher frequency than the other categories, having the double of samples of the third most represented category (*sports*). Among the less frequent ones, there are *video games* and *musical instruments*. Moreover, if we compare the distribution of the reviews per product category by language, there are some interesting insights:

- The category *grocery* is more frequent in the English dataset w.r.t. the other ones. We assume that this is caused from the greater prevalence of online grocery shopping among English speakers, particularly in the UK and US. In contrast, such purchases are less common in continental Europe, where "walkable cities" often facilitate in-person shopping. This discrepancy could introduce a form of bias, which arises from cultural (but also infrastructural) differences. This phenomenon may be even more emphasized for languages spoken in geographical regions outside Europe and North America.
- The same considerations are valid for the categories *apparel*, *shoes* and *beauty*, that are more represented in the English dataset. In Spain and in Germany, for example, it is more common to go to a local store and to try on clothes/shoes, w.r.t. to the U.S. cities, where people generally need private vehicles to go to malls/supermarkets.

- There is also a significant difference of frequency of the categogry *drugstore*, way more common in the English dataset. English speaking populations might be more accustomed to purchasing drugstore items online. In contrast, in Spanish and German speaking countries there might be a stronger tradition of buying such items in physical stores. Another possible reason is that drugstores in English speaking countries may actively encourage online purchases through promotional campaigns or discounts, which could result in a higher number of reviews.

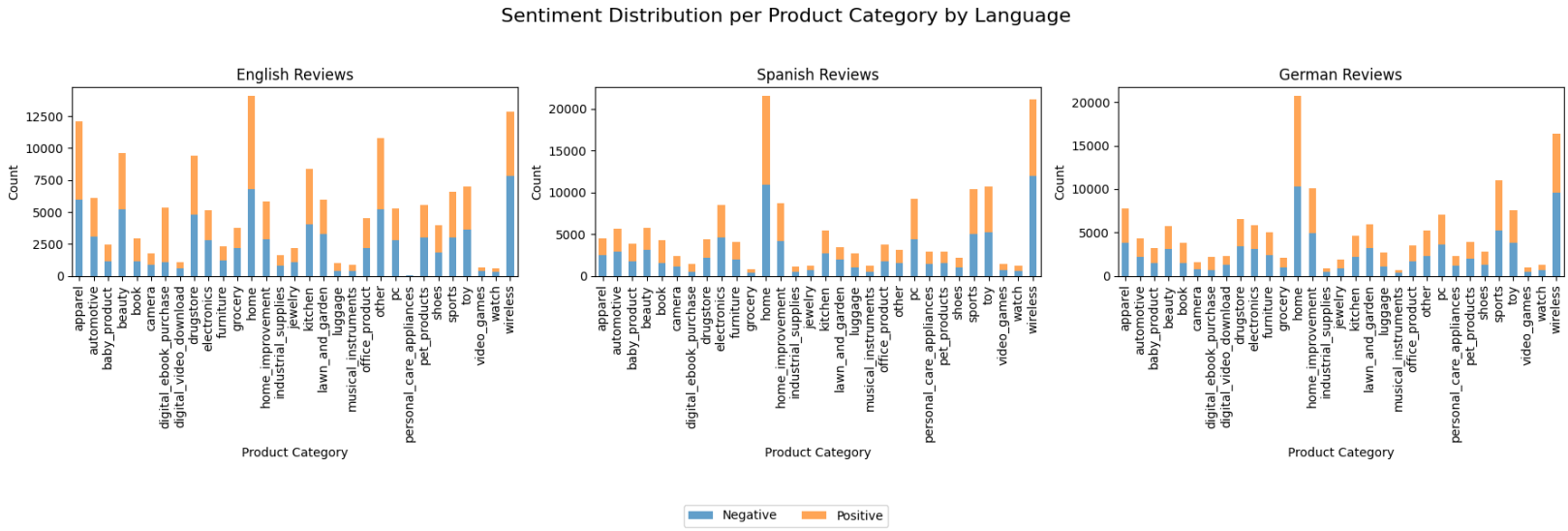

Figure 2: Distribution of the labels per product category.

By looking at the distribution of review lengths by language in Figure 3, we notice that English reviews are, in general, shorter than German and Spanish reviews. This is probably because German compound words and formal structures may result in longer expressions to describe a product or experience, and Spanish reviews may reflect a more "conversational style" (due to the abundant use of adjectives). Another consideration to take into account is that some English reviews may be written by non-native English speakers, which may not have a very big vocabulary (resulting in scarcity of usage of technical terms or adjectives). We assume that classifiers that are able to perform well on short English reviews because it's way easier for them to memorize simple patterns. On the other hand, the longer Spanish and German reviews may be harder to classify correctly due to their higher complexity.

We then check the average review length by sentiment per each language in Figure 4. For all the three languages, the average length of the reviews is higher for Negative label (around 35). This may be caused by the fact that, usually, people are more prone to describe in details the negative sides of a product. In fact, negative experiences often evoke stronger emotions than positive ones, and people may feel compelled to elaborate on their dissatisfaction to vent their frustration or to process their emotions. This is true for the Western civilization, but not for other cultures. Japanese users, for example, may use less words in average for negative reviews, because Japanese communication tends to be more implicit and subtle. So, if a model learns that longer reviews correlate with negativity (based on Western training data), it might misinterpret shorter negative reviews coming by other cultures.

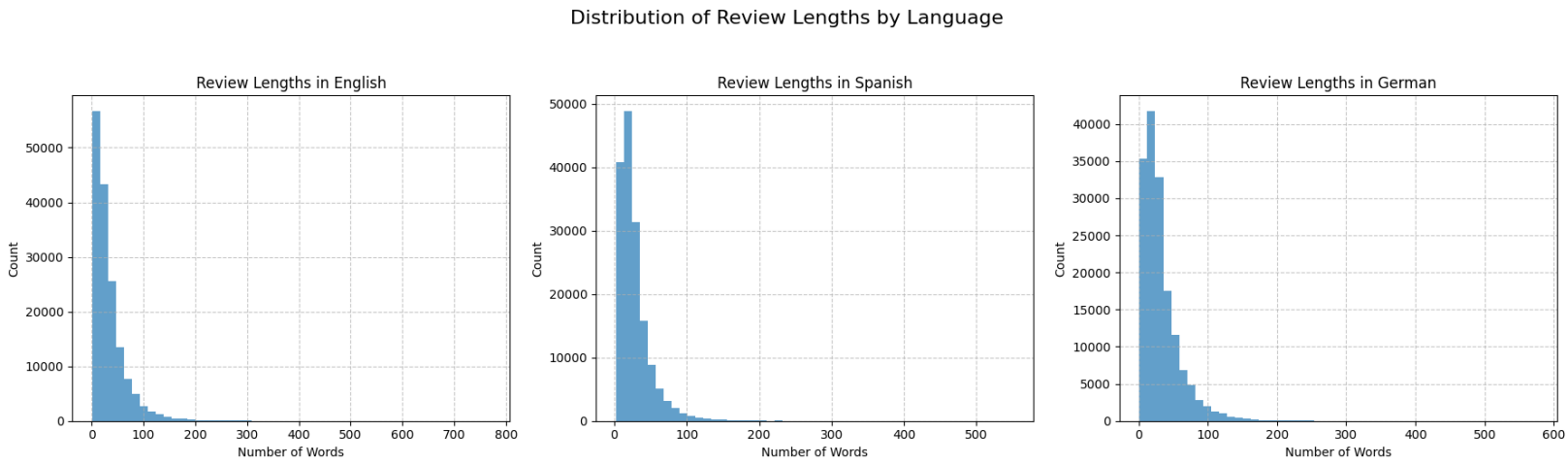

Figure 3: Distribution of the review length by language.

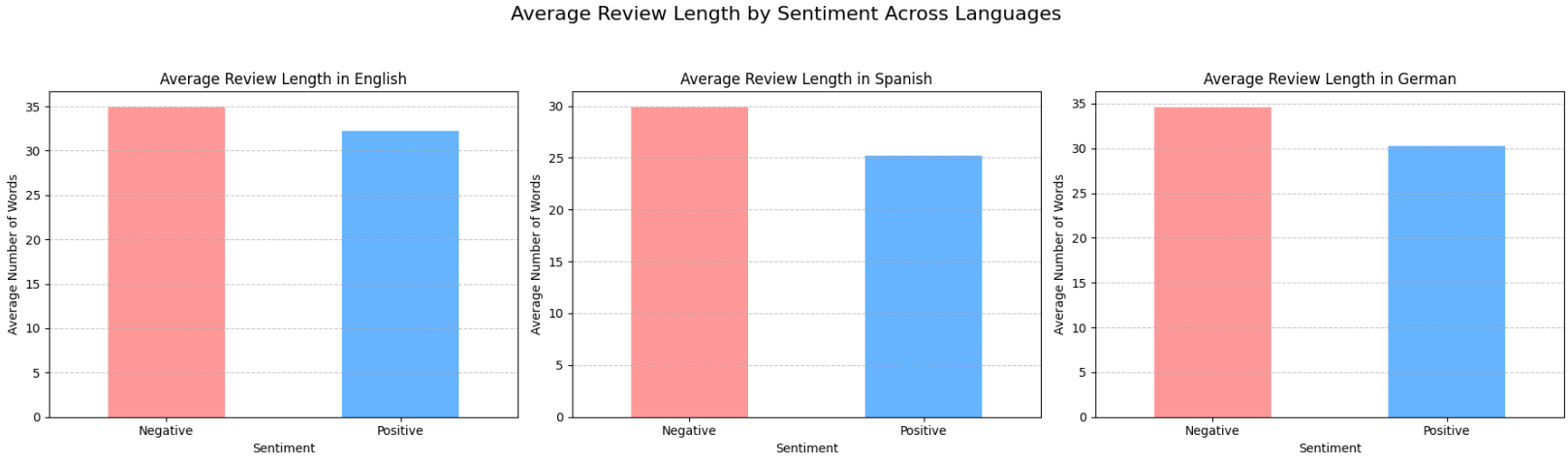

Figure 4: Average review length by sentiment.

Lastly, if we look at the most common words per each language (excluding stopwords and punctuation) dipsplayed in Figure 5, we observe that English and Spanish have overall similar common terms (i.e. "good"/"bien" and "product"/"producto"), generally neutral or positive words. For German is instead more frequent to use negative terms, like "leider" (meaning "unfortunately"). The predominance of negative terms like in German reviews can be attributed to a greater tendency among Germans to express dissatisfaction more openly.

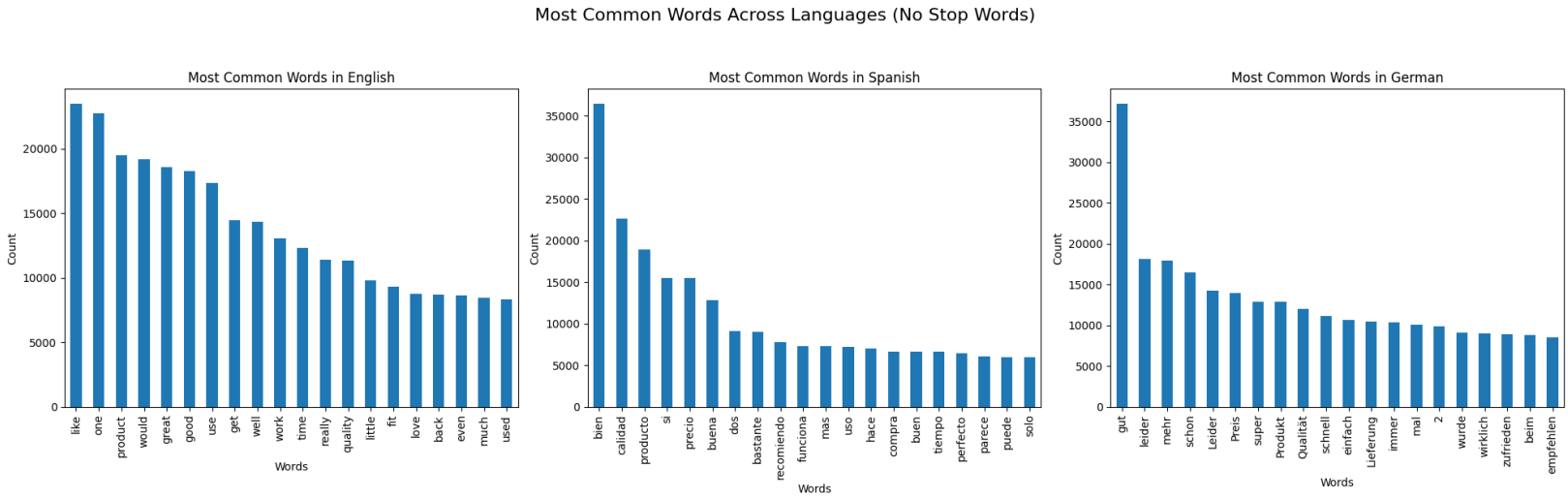

Figure 5: Most common words by language.

5.4 Metrics

In this section we describe the metrics that we will use to analyze the performance of our models and their bias towards a certain category or language. For this we will use the following metrics:

- Accuracy: It represents a widely used metric in classification tasks, good for getting an initial read on how the model is performing.
- F1 Score: it represents the harmonic mean between precision and recall, offering a more balanced view of the performance of the model, when compared to accuracy.

The next metrics will be inspired from the case of SAPMOC presented in the lectures. These metrics are computed based on the confusion matrix, and they help discern where there are biases in the data. The following abbreviations will be used:

- True Positives (TP) - A prediction is put in this category twhen the predicted label is correct.
- False Positives (FP): The model predicted positive, but the true outcome is negative.
- True Negatives (TN): The model predicted negative, and true outcome is negative.
- False Negatives (FN): The model predicted negative, but the true outcome is positive.

Next we will present the SAPMOC metrics, based on the notions explained and enumerated above:

- Statistical Parity - Each group should have an equal probability of being predicted as positive. Formula:

$$\text{Statistical Parity} = \frac{TP + FP}{TP + FP + TN + FN} \tag{1}$$

  For a binary classification task, the amount of predicted positives should be 50 percent if the dataset is balanced.

- Equal of opportunity (True Positive Rate): The members of each group, which have the same features, should be predicted in equal proportion. Formula:

$$\text{Equal of opportunity(Positive)} = \frac{TP}{TP + FN} \tag{2}$$

$$\text{Equal of opportunity(Negative)} = \frac{TN}{TN + FP} \tag{3}$$

  In practice, this metric should have the same value across all groups.

- Calibration (TN): Within each group, the ratio between positive(negative) predictions is equal with the number of actual positives(negatives). Broadly, calibration holds if the following formulas are equal:

$$\text{Calibration(Positive)} = \frac{TP}{TP + FP} \tag{4}$$

$$\text{Calibration(Negative)} = \frac{TN}{TN + FN} \tag{5}$$

  The values of the two equations should be equal across all groups.

- Conditional use error (False Rate): Error rates should be similar across groups. This translated into the following two formulas having similar values:

$$\text{Calibration(Positive)} = \frac{FP}{TP + FP} \tag{6}$$

$$\text{Calibration(Negative)} = \frac{FN}{TN + FN} \tag{7}$$

- Treatment Equality: No group should receive more favorable or disadvantageous treatment due to prediction errors. In practice, we need the following formulas to be equal across all the groups.

$$\text{Positive Side} = \frac{FP}{FN} \tag{8}$$

$$\text{Negative Side} = \frac{FN}{FP} \tag{9}$$

5.5 Models

5.5.1 BERT multilingual

BERT multilingual is an encoder-only model pretrained on a large corpus of multilingual data (104 languages). It is based on the concept of bidirectional attention, in order to understand the context both left-to-right and right-to-left. It is widely used in NLP tasks like sentence classification, as it is possible to add a classification head on top of the main architecture and fine-tune the model on a downstream task. This generally involves the freezing of the original weights, performing backpropagation only on the parameters of the classification head (as we did in this project). Its original version has 179 million parameters, but there are many lighter (and better performing) variants, obtained via transfer learning, distillation or other SOTA techniques.

5.5.2 DistilBert

The Distilbert-base-multilingual-cased was published by Microsoft and it is the distilled version of the BERT multilingual base model, that was described above. The model has 134 milions parameters, 6 layers, and a hidden dimension of 768. This model is on average up to two times faster than the classical model. Furthermore, the model was trained on more than 134 languages, which offers a good basis to start from for our task.

5.5.3 XLM roberta

The XLM-RoBERTa-base model was published by Meta AI and is a multilingual variant of the RoBERTa architecture. It has approximately 270 million parameters, 12 transformer layers, and a hidden size of 768. XLM-RoBERTa was trained on 2.5TB of filtered CommonCrawl data in 100 languages, using a SentencePiece tokenizer without language-specific tokens. In comparison to earlier multilingual models, XLM-RoBERTa demonstrates improved performance on cross-lingual testing due to its enormous training and more robust tokenization. Its multilingual competence makes it a good candidate for sentiment analysis across a broad spectrum of languages.

5.6 Experiments

5.6.1 Baseline model training

For our baseline experiment, we fine-tuned each of the three pretrained multilingual transformer models on the full train dataset (containing only the languages of interest).

We froze all the parameters of the base transformer layers, allowing only the classification head to be updated during training. This was necessary because the training time for 2 of the models exceeded more than 12 hours per epoch. This significantly reduced computational cost and training time, while still adapting the models to the sentiment classification task.

Still, as a result of this modification, the training would take more than 4 hours per epoch for the bigger models (XLM-RoBERTa-base and bert-base-multilingual-cased). That is why we used automatic mixed precision (AMP) via PyTorch's autocast context. This way the calculations were faster and it reduced memory usage by performing calculations in lower precision where feasible. We also employed a gradient scaler, which dynamically adjusts loss scaling to prevent numerical instability and vanishing gradients during backpropagation. In the end, each epoch took approximatively 1 hour and that is why we could train it for only a small number of epochs.

Each model was fine-tuned with the following hyperparameters:

- **XLM-RoBERTa-base:** 3 epochs, learning rate of 2×10^{-4}, batch size of 256.

- **DistilBERT-multilingual-cased:** 3 epochs, learning rate of 5×10^{-6}, batch size of 16.
- **bert-base-multilingual-cased:** 3 epochs, learning rate of 2×10^{-4}, batch size of 256.

To evaluate the models we looked at the F1-scores and accuracy to asses the performance. We also evaluated model fairness using group-parity criteria inspired by the SAPMOC framework.

5.6.2 Model fine-tuning per language

To address any identified bias and explore language-specific behavior, we fine-tuned three separate instances of each base model on the three languages in the training dataset, keeping the same hyperparameters.

The goal was to evaluate whether specialized fine-tuning for each language would improve performance and/or fairness. After training, we saved the model weights for each configuration and compared results with the baseline model and between the languages.

5.6.3 Ensemble Experiment

In the final experiment, we examined whether combining models could produce more fair results across languages. Each team member focused on one language and used the saved weights from the per-language fine-tuned models to classify the test data with every model.

We implemented a majority voting ensemble, where predictions were made based on the majority class output by the three individual models for each example. This method attempted to reduce individual model biases via the application of complementary strengths over models.

6 Deliverables

- **Code Repository**: A GitLab repository hosted on Apice, containing scripts for dataset preprocessing and analysis, model fine-tuning and evaluation, fairness metric computation, and bias mitigation (where applicable).
- **Model Weights**: The weights of the trained models in order to ensure the reproducibility of the experiments.
- **Final Report**: A comprehensive written report summarizing background motivation, dataset and preprocessing details, model comparison, fairness evaluation, mitigation strategies explored, and key findings.

7 Results

7.1 Baseline model training

7.1.1 BERT

Below are reported the performances of bert-base-multilingual-cased trained on all the three languages.

Language Group	Accuracy	F1 Score
German	74.4%	73.7%
English	77.0%	76.4%
Spanish	77.3%	78.1%

Table 1: Language-specific performance metrics for bert-base-multilingual-cased

We can notice that, while English and Spanish have overall similar results, German has a statistically significant difference. Since MARC is perfectly balanced in terms of language distribution, the performance gap can be explain with an under representation of German in the training corpora used by Google AI Language team for the pre-training of the model. Further analysis will be presented in Section 8. We then present the fairness metrics in the following table.

	Statistical Parity	Equality of Opportunity		Calibration		Conditional Use Error		Treatment Equality	
Model	**Positive**	**Pos**	**Neg**	**Pos**	**Neg**	**Pos**	**Neg**	**Pos**	**Neg**
German	47.3%	71.8%	75.8%	77.1%	73.2%	24.2%	26.8%	81.1%	123.4%
English	47.4%	74.4%	78.6%	73.4%	75.7%	21.4%	24.3%	79.3%	126.1%
Spanish	53.9%	81.2%	75.3%	79.8%	79.6%	24.7%	20.4%	141.5%	70.7%

Table 2: Fairness metrics for bert-base-multilingual-cased (trained on all languages) across test languages.

We can notice that, for German, the Statistical Parity shows a slight underrepresentation of positive samples, meaning that the model leans towards False Negative predictions. The same happens with English, while for Spanish there is an opposite trend. Moreover, Positive Equality of Opportunity is highly unbalanced, especially between German and Spanish, with a delta of almost 10%. There is instead a very slight unbalance for Negative Equality of Opportunity. Overall, we can observe unbalanced metrics on all the other fairness metrics, even if slight.

The different statistical parity distribution across the three languages is also reflected in Treatment Equality, with Spanish having an opposite trend w.r.t. English and German. This indicates that the model's errors disproportionately affect positive samples differently depending on the language, which could have real-world implications for fairness in multilingual sentiment analysis applications.

7.1.2 DistilBert

In this section we will analyze both the highlevel performance of DistilBert-multilingual-cased model on the selected languages, and the overall disparieties in fairness the model exhibits on each and every language.

Language Group	Accuracy	F1 Score
German	78.7%	78.3%
English	78.2%	76.9%
Spanish	78.8%	78.3%

Table 3: Language-specific performance metrics for DistilBERT-multilingual-cased (trained on all languages)

The table 3 offers a high level view of the performance of the model on each language. While the accuracy seems to tell us that the model performs well on each language, examining a somewhat better-suited metric, the F1-Score points out that the English language has more bias in it.

In the next table, we will use the SAPMOC metrics, to take a closer look at the situation.

	Statistical Parity	Equality of Opportunity		Calibration		Conditional Use Error		Treatment Equality	
Language	**Positive**	**Pos**	**Neg**	**Pos**	**Neg**	**Pos**	**Neg**	**Pos**	**Neg**
German	48.3%	76.9%	80.4%	79.7%	77.7%	20.3%	22.3%	84.9%	117.9%
English	44.1%	72.4%	84.1%	81.9%	75.2%	18.0%	24.8%	57.5%	173.9%
Spanish	48.0%	76.8%	80.7%	79.9%	77.7%	20.1%	22.3%	83.2%	120.2%

Table 4: Per-language fairness metrics for DistilBERT-multilingual-cased (trained on all languages)

We will start by taking a look at the Spanish language. The Statistical Parity shows a slight bias against the positive label (2% from ideal). Taking a look at the Equality of Opportunity, we can see a difference (3.2%) between the positive and negative sections, indicating that the model correctly predicts a true label more reliably for the negative label. For calibration, we can observe a slight discrepancy (2.2% difference), indicating a close-to-balanced model. Taking a look at the Conditional Use Error, we can see that the model has a slightly higher chance of mistaking a negative label for a positive one.

Examining the model in the German Language, we can see that the fairness performance follows the Spanish one.

The model is 53% more likely to misclassify positive English reviews as negative w.r.t. Spanish. The rest of the metrics solidify this finding, such as a 4% decrease in the Positive Statistical Parity.

To sum up, the DistilBert-multilingual-cased model has an overall underrepresentation tendency towards the Positive class, with the German and Spanish languages being less prone to this, while the English reviews suffer the most.

7.1.3 XLM-RoBERTa-base

In this section, we analyze the performance and fairness of XLM-RoBERTa-base, trained on all three target languages: German, English, and Spanish. Among the three evaluated models, XLM-RoBERTa-base demonstrated the highest overall performance.

Language Group	Accuracy	F1 Score
German	89.3%	89.2%
English	85.1%	85.5%
Spanish	87.9%	88.2%

Table 5: Language-specific performance metrics for XLM-RoBERTa-base

As shown in Table 5, the model achieves its best performance on the German dataset, with an accuracy of 89.3% and an F1 score of 89.2%. This is followed by Spanish and then English.

	Statistical Parity	Equality of Opportunity		Calibration		Conditional Use Error		Treatment Equality	
Language	**Positive**	**Pos**	**Neg**	**Pos**	**Neg**	**Pos**	**Neg**	**Pos**	**Neg**
German	49.6%	88.9%	89.7%	89.6%	88.9%	10.4%	11.1%	92.4%	108.3%
English	52.4%	87.6%	82.8%	83.6%	86.9%	16.4%	13.1%	138.2%	72.4%
Spanish	52.7%	90.6%	85.2%	86.0%	90.0%	14.1%	10.0%	156.6%	63.9%

Table 6: Per-language fairness metrics for XLM-RoBERTa-base

To further examine performance disparities, we analyze fairness using the SAPMOC framework. With a balanced dataset, Statistical Parity should ideally be 50%; German aligns most closely (49.6%), while English and Spanish deviate slightly. In terms of Equality of Opportunity, German again shows the most balanced recall across classes (88.9% positive, 89.7% negative), whereas Spanish and English favor positive predictions, with lower recall on negatives-indicating less equitable treatment.

Treatment Equality, measured as the ratio of false positive to false negative rates, shows the largest discrepancies. Spanish shows a strong imbalance with a ratio of 156.6% (positive) to 63.9% (negative), while English shows 138.2% to 72.4%. German is the most balanced, with 92.4% to 108.3%, reflecting more equitable treatment between error types.

XLM-RoBERTa-base delivers the strongest overall performance and fairness metrics reveal slight disparities between languages.

7.2 Model fine-tuning per language

7.2.1 BERT

Below are the results for multilingual BERT fine-tuned per language.

Language Group	Accuracy	F1 Score
German	73.3%	74.5%
English	76.0%	74.7%
Spanish	74.4%	73.7%

Table 7: Performance metrics by language group for bert-base-multilingual-cased fine-tuned per language

As the table above shows, the overall results are slightly worse than the baseline model. Further analysis will be carried out in the Discussion. We proceed by analyzing the fairness metrics in the table below.

	Statistical Parity	Equality of Opportunity		Calibration		Conditional Use Error		Treatment Equality	
Model	**Positive**	**Pos**	**Neg**	**Pos**	**Neg**	**Pos**	**Neg**	**Pos**	**Neg**
German	54.8%	78.2%	68.6%	71.3%	75.8%	28.7%	24.2%	143.7%	69.6%
English	44.7%	70.8%	81.3%	79.1%	73.5%	20.9%	26.5%	63.8%	156.8%
Spanish	53.9%	80.2%	72.4%	74.4%	78.6%	25.6%	21.4%	140.0%	71.4%

Table 8: Per-language fairness metrics for bert-base-multilingual-cased fine-tuned models.

While English Positive class is still underrepresented, the trend is inverted for German, and this is reflected in all the other fairness measures, with German and Spanish reviews being treated very similarly. The loss of cross-lingual knowledge transfer probably leads to these different outcomes: in fact, German and English both belong to the Germanic branch of the Indo-European language family, so without a joint context the performances will surely be different. This is not valid for Spanish, since it is a romance language (thus, the lack of cross-lingual training is less evident).

7.2.2 DistilBert

Language Group	Accuracy	F1 Score
German	77.7%	77.7%
English	78.4%	77.0%
Spanish	81.7%	81.5%

Table 9: Performance metrics by language group for DistilBERT fine-tuned and evaluated on each specific language

The table presented above, when compared with Table 3, shows that the overall performance for English and German remained the same, with under 1% increases or decreases in their F1 scores, while the Spanish finetuned version shows significant improvement, reaching an 81.5% performance, compared with the 78.3% baseline.

	Statistical Parity	Equality of Opportunity		Calibration		Conditional Use Error		Treatment Equality	
Language	**Positive**	**Pos**	**Neg**	**Pos**	**Neg**	**Pos**	**Neg**	**Pos**	**Neg**
German	49.9%	77.6%	77.8%	77.8%	77.3%	22.2%	22.7%	98.7%	101.4%
English	43.6%	72.1%	84.8%	82.6%	75.2%	17.4%	24.8%	54.2%	184.5%
Spanish	48.9%	76.8%	80.7%	80.6%	82.4%	17.6%	18.9%	88.9%	112.5%

Table 10: Per-language fairness metrics for DistilBERT-base fine-tuned and evaluated on each specific language

For Spanish, fine-tuning led to a 0.9% improvement in Statistical Parity, reduced prediction bias, and significant improvements in Conditional Use Error and Treatment Equality metrics, reflecting an 8% decrease in overrepresentation of negative labels. German showed a slight decline in F1-score and accuracy but achieved a near-perfect Statistical Parity (49.9%) and substantial improvements in Treatment Equality, reducing the misclassification rate of positive reviews. However, for English, fairness got worse, with decreases in Statistical Parity and a significant rise in Treatment Equality disparity, indicating increased misclassification of positive reviews as negative.

7.2.3 XLM-RoBERTa-base

We now examine the performance and fairness of XLM-RoBERTa-base models fine-tuned separately on German, English, and Spanish datasets.

Language Group	**Accuracy**	**F1 Score**
German	86.9%	86.8%
English	84.3%	84.5%
Spanish	84.3%	84.7%

Table 11: Performance metrics by language group for XLM-RoBERTa-base fine-tuned per language

As shown in Table 11, the German model achieves the highest accuracy, maintaining its lead from the multilingual setup, though with a slight performance drop of 2.4 percentage points compared to the jointly trained version.

	Statistical Parity	**Equality of Opportunity**		**Calibration**		**Conditional Use Error**		**Treatment Equality**	
Language	**Positive**	**Pos**	**Neg**	**Pos**	**Neg**	**Pos**	**Neg**	**Pos**	**Neg**
German	49.5%	86.4%	87.4%	87.3%	86.5%	12.7%	13.5%	92.6%	107.9%
English	50.9%	85.3%	83.4%	83.7%	84.9%	16.3%	15.0%	112.2%	89.1%
Spanish	52.8%	87.1%	81.6%	82.5%	86.3%	17.5%	13.7%	142.5%	70.2%

Table 12: Per-language fairness metrics for XLM-RoBERTa-base fine tuned models

Looking at Statistical Parity (Table 12), German again shows the most balanced distribution at 49.5%, while English is showing an improvement comparing to the baseline model getting closer to parity by 1.5%. German also maintains strong alignment in Equality of Opportunity with recall values of 86.4% (positive) and 87.4% (negative).

Calibration values across all fine-tuned models range from a minimum of 82.5% to a maximum of 87.3%, resulting in a 4.8% spread. This is an improvement over the multilingual setup, which showed a wider gap of 6.4%. The reduced spread suggests that the fine-tuned models exhibit more consistent calibration across languages and classes.

7.3 Ensemble per language compared to the base models

The following table presents the performance metrics of the ensemble model.

Language Group	Accuracy	F1 Score
German	82.6%	82.7%
English	82.2%	81.7%
Spanish	84.4%	83.3%

Table 13: Performance metrics by language group for the ensemble model

Even if the ensemble model demonstrates lower performance compared to XLM-RoBERTa-base in terms of accuracy and F1-score, these metrics are not the main focus of the design of the ensemble, since we want to achieve a fairer model. The following table presents the relevant fairness metrics.

	Statistical Parity	Equality of Opportunity		Calibration		Conditional Use Error		Treatment Equality	
Language	**Positive**	**Pos**	**Neg**	**Pos**	**Neg**	**Pos**	**Neg**	**Pos**	**Neg**
German	50.7%	83.3%	82.0%	82.2%	83.1%	17.8%	16.9%	107.8%	92.8%
English	46.9%	79.2%	85.3%	84.4%	80.4%	15.6%	19.6%	70.4%	141.9%
Spanish	43.4%	77.9%	91.1%	89.7%	80.4%	10.3%	19.6%	40.4%	247.0%

Table 14: Per-language fairness metrics for Ensemble model

The ensemble model aggregates predictions from language-specific fine-tuned models, but it shows mixed fairness results: while calibration is stable and English sees slight fairness improvements, significant gaps emerge for Spanish, with high Treatment Equality imbalance and reduced Equality of Opportunity. Performance metrics also decline w.r.t. the fine-tuned XLM-RoBERTa model. In general, the ensemble struggles to significantly improve fairness, often amplifying biases from individual models.

8 Discussion

8.1 BERT

BERT has some interesting results that tells us about possible unfairness in the corpus used for the pre-training. In fact, the performance metrics of the model fine-tuned on specific languages (presented in Table 7) are worse than the ones obtained with the baseline model (presented in Table 1). This may be caused by the loss of cross-lingual knowledge transfer that occurs when fine-tuning separately per language: the baseline model benefits from learning shared representations and patterns across languages, which can improve generalization. Instead, when fine-tuned individually, each language model has access to fewer training samples and the languages are treated independently, resulting in a decrease of the knowledge transfer capability (leading to lower performances).

German and English have similar results when treated in a joint context, with Spanish being an outlier. At the same time, German and Spanish have similar results when treated independently, with English being an outlier. This leads to believe that there is indeed an unbalance in the training corpora used for the pre-training of multilingual BERT, probably with a higher frequency of English content. Indeed, if we look at the training corpus (the whole Wikipedia), we notice that it is highly skewed towards English.

This unbalance is mitigated by cross-language transfer knowledge for languages that are similar to English (like German), but is more evident if the languages are treated independently. Note

that English and German are both part of the Germanic branch of Indo-European languages, while Spanish is part of the Neo-Latin branch.

This suggests that while language-specific fine-tuning helps to align the treatment of non-English languages, it also creates a higher treatment gap between those and English.

As we can see for the other models, this phenomenon slightly happens for DistilBERT, while is not present in XLM-RoBERTa, since it is specifically designed to optimize cross-language transfer knowledge, with a balanced language sampling and with a bigger training corpus.

8.2 DistilBert

We begin our analysis with the general DistilBERT-multilingual-cased model trained on all languages and evaluated separately on each. One of the most indicative metrics is Treatment Equality, which shows a consistent tendency of the model to classify positive reviews as negative, especially in Spanish and English. Even if the dataset is perfectly balanced across languages and sentiment labels, the model shows an overall under-representation of the Positive class.

For Spanish, Treatment Equality shows a 37% discrepancy between the False Negative and False Positive rates, showing a bias towards over-predicting the negative class. German follows a similar trend, but the discrepancy is slightly lower at 33%, suggesting a smaller (but still present) bias. English has the most severe disparity: a False Negative to False Positive ratio of 173%, compared to 120% in Spanish. This means that the model is 53% more likely to misclassify positive English reviews as negative. Other fairness metrics confirm this trend, with English also showing a 4% decrease in Positive Statistical Parity. So, while the multilingual model performs consistently across languages, it shows a clear trend to misclassify positive sentiment, especially in English.

In contrast, when the model is fine-tuned separately for each language, we observe mixed results. Spanish benefits from fine-tuning. Statistical Parity improves by 0.9%, indicating a more even distribution between positive and negative predictions. Conditional Use Error drops below 19%, down from 22.3%, and Treatment Equality improves from over 120% to 112.5%, reflecting an 8% decrease in the bias toward negative predictions.

German shows a similar improvement in fairness. Even if the overall performance slightly drops in terms of F1 score and accuracy, fairness metrics show a better improvement. Statistical Parity reaches 49.9%, compared to a baseline of 48.3%, and Treatment Equality drops from 17.4% to just 1.4%, suggesting that the model is now almost equally likely to classify positive and negative reviews correctly.

English, however, presents a different outcome. Fine-tuning the model on English leads to a decline in fairness. Treatment Equality reaches 184.5% (meaning an 84.5% deviation from the ideal case), and Positive Statistical Parity drops by 0.7%. This shows an evident downgrade in the model's ability to represent the Positive class fairly, reinforcing that fine-tuning, while beneficial for some languages, can amplify existing biases if not properly managed.

In summary, the general multilingual model shows a stable but biased behavior across languages, with a consistent over-representation of the Negative class, with the worst case being the case of the English reviews. Fine-tuning, on the other hand, helps reduce this bias in Spanish and leads to near-perfect fairness in German, where disparities across all metrics become less significant. English shows a clear decline in fairness following fine-tuning, with increased bias against the Positive class. These findings suggest that fine-tuning can be an effective strategy for correcting imbalance, but it is not universally beneficial. Each language requires careful evaluation to ensure that improvements in performance do not come at the cost of fairness.

8.3 XLM-RoBERTa-base

The performance and fairness results across languages (see Table 5) show that XLM-RoBERTa-base performs better than BERT-base-multilingual-cased and DistilBERT. When fine-tuned jointly on English, German, and Spanish Amazon reviews, it achieves the highest scores overall, with especially strong results in German (89.3% accuracy, 89.2% F1-score). Across all three languages, XLM-RoBERTa demonstrates the most equitable performance in terms of fairness measures showing the least amount of bias.

The better performance demonstrated by XLM-RoBERTa is a result of the combination of its architectural design along with its extensive training data. XLM-RoBERTa was trained on 2.5TB of filtered CommonCrawl data for 100 languages which represents a significant enhancement over BERT's training on 104 languages from Wikipedia text. The extensive and diverse corpus of XLM-RoBERTa provides superior cross-lingual generalization with enhanced contextual representation which enables the model to process syntactic and lexical variations more effectively. The compressed nature of DistilBERT from BERT results in insufficient representation depth which causes the model to show lower accuracy and F1-scores across all languages compared to XLM-RoBERTa.

The model seems to be pretty fair overall. In terms of Statistical Parity each language is very close to 50%, showing that the model is just as likely to predict positive and negative values for all languages (and given the balanced distribution of the data, this is the desired outcome). Equality of Opportunity shows relatively high and balanced scores across both positive and negative classes, particularly for German (88.9%, 89.7%) and Spanish (90.6%, 85.2%), implying that the model performs similarly across all the languages. Calibration scores are similarly consistent, given that we want all the values to be the same across all groups and sentiments, the delta between the maximum and minimum value is of only 6%. Treatment Equality metrics show greater disparities, especially for English (138.2%, 72.4%) and Spanish (156.6%, 63.9%), revealing big imbalances in false positive/negative tradeoffs between groups, which could cause some fairness concerns.

Interestingly, the effectiveness of XLM-RoBERTa experiences a slight decline in performance when each language receives individual fine-tuning. The results displayed in Table 11 demonstrate that accuracy and F1-scores decrease by a small amount (e.g., German: 86.9% accuracy, 86.8% F1). When it comes to bias, there aren't any great modifications, since the model was originally quite fair, but we can see that the calibration gap decreased slightly (by 2%) and the Treatment of Equality difference between positive positive and negative values decreased by 20% for Spanish, the one with the biggest bias in the baseline model.

8.4 Ensemble model

The ensemble model, created by aggregating predictions from the three language-specific fine-tuned models (BERT-base, DistilBERT and XLM-RoBERTa-base), was designed to combine the strengths of each individual model and reduce variance in performance. However, the fairness metrics in Table 14 reveal mixed outcomes. Calibration values remain relatively stable across languages and classes (ranging from 80.4% to 89.7%), and Conditional Use Error rates for German and English stay within a moderate range. Still, the ensemble introduces notable disparities in other fairness dimensions, particularly for Spanish.

Especially, Treatment Equality for Spanish is highly imbalanced, with a false prediction ratio of 40.4% (positive) to 247.0% (negative), representing the most severe skew observed across all evaluated models. Similarly, Equality of Opportunity for Spanish shows a gap between classes, with a strong negative-class performance (91.1%) but a substantially lower rate on the positive class (77.9%).

Overall, the ensemble model does not consistently outperform individual fine-tuned models in terms of fairness, for example, it introduces or exacerbates disparities for Spanish. A likely explanation is that the ensemble aggregates not only the strengths but also the biases of its components. Since the underlying models share similar architectures and training data distributions, they may exhibit overlapping, uncorrected biases. When these are combined, the ensemble may amplify unfair treatment of certain groups: Spanish in this case being the most affected.

9 Conclusion

In conclusion, fine-tuning multilingual BERT can be effective, but it requires careful consideration, particularly when addressing disparities in treatment across languages. The baseline BERT model demonstrates a tendency toward false negative predictions for English and German tweets, while showing the opposite trend (false positives) for Spanish tweets. Language-specific fine-tuning does not significantly improve fairness, as Spanish and German become more aligned in their behavior (leaning towards false positives), with English remaining the outlier.

On the other hand, the general DistilBERT model shows the potential of language-specific fine-tuning to enhance fairness in sentiment detection with encoder-only transformers. While the general model exhibits a consistent bias against the Positive class, most notably for English, fine-tuning yields encouraging yet mixed results. Spanish and German achieve noticeable fairness improvements, with German nearly achieving perfect parity across fairness metrics. However, English fairness worsens after fine-tuning, underscoring the complex nature of fairness. After these considerations, the distillation process actually helped to mitigate some of the biases present in the original BERT multilingual model.

Among the evaluated models, XLM-RoBERTa stands out as both the most accurate and the fairest across languages. Compared to BERT, which struggles with performance, and DistilBERT, which trades representation capacity for speed, XLM-RoBERTa delivers consistent fairness alongside better performance. These findings accentuate the importance of model selection in mitigating bias in multilingual sentiment analysis.

Finally, we explored the potential of model ensembling to enhance fairness. Contrary to expectations, the ensemble approach amplified biases rather than mitigating them, as the two BERT-based models are too similar with each other, thus their biases are overlapped (leading to even more unfairness). We believe that the ensemble technique has potential for reducing unfairness, but we may need a more diverse and a bigger number of models so that the sum of the components results in better fairness than the individual instances.

10 Future work

Future work should examine the fairness of pre-training corpus, the one used for BERT and DistilBERT is highly skewed towards English, with more than 7M Wikipedia articles in that language (as of May 2025).

Another potential future direction is a further exploration of the distillation process, assessing if it can actually reduce bias w.r.t. the teacher model, or if our study case was just an outlier.

Lastly, further analysis could focus on designing more diverse ensembles to maximize their potential in reducing bias.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[2] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[3] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116, 2019.

[4] Amazon Web Services. Amazon multilingual reviews corpus (marc). `https://huggingface.co/datasets/amazon_reviews_multi`, 2020.

[5] Francesca Lagioia, Riccardo Rovatti, and Giovanni Sartor. Algorithmic fairness through group parities? the case of compas-sapmoc. *AI & Society*, 38(2):459–478, 2023.

[6] Phillip Keung, Yichao Lu, Etienne Marcheret, and Matias Susik. The multilingual amazon reviews corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4563–4568, 2020.