Semantic Meaning or Script Shape? A Comparative Study of Cross-Lingual Transfer in mBERT and PIXEL

Zhenming Li, Kazutaka Shimada

Proceedings of the 39th Pacific Asia Conference on
Language, Information and Computation (PACLIC 39)

Emmanuele Chersoni, Jong-Bok Kim (eds.)

2025

© 2025. Zhenming Li, Kazutaka Shimada. Semantic Meaning or Script Shape? A Comparative Study of Cross-Lingual Transfer in mBERT and PIXEL. In Emmanuele Chersoni, Jong-Bok Kim (eds.), *Proceedings of the 39th Pacific Asia Conference on Language, Information and Computation* (PACLIC 39), 561-569. Institute for the Study of Language and Information, Kyung Hee University. This work is licensed under the Creative Commons Attribution 4.0 International License.

Semantic Meaning or Script Shape? A Comparative Study of Cross-Lingual Transfer in mBERT and PIXEL

Zhenming Li
Kyushu Institute of Technology
Fukuoka, Japan
li.zhenming714@mail.kyutech.jp

Kazutaka Shimada
Kyushu Institute of Technology
Fukuoka, Japan
shimada@ai.kyutech.ac.jp

Abstract

Multilingual BERT (mBERT) has been extensively investigated for cross-lingual transfer learning (CLTL), achieving strong performance owing to its rich semantic representations. Recent studies have demonstrated that visually grounded models, such as PIXEL, can also be applied to CLTL by leveraging character-level glyph information. In this work, we present a comparative study of mBERT and PIXEL in a zero-shot cross-lingual transfer setting across 5 languages. Our results show that mBERT consistently outperforms PIXEL in overall accuracy, underscoring the effectiveness of semantic representations for CLTL. Nevertheless, we find that PIXEL exhibits competitive performance for visually similar language pairs and maintains robustness when semantic information is limited, suggesting the usefulness of visual information in cross-lingual transfer scenarios.

1 Introduction

Cross-lingual transfer learning (CLTL) aims to exploit knowledge acquired in a source language to improve performance in a target language. multilingual BERT (mBERT) (Devlin et al., 2019), has been widely explored in CLTL tasks and shows great performance. mBERT adopts a multilingual subword tokenization strategy and learns shared semantic representations from large-scale multilingual corpora, thereby enabling robust cross-lingual generalization. It is well established that semantic-based models, such as mBERT, can serve as effective backbones for CLTL tasks.

Complementary to these token-based approaches, recent work in Visual NLP explores the potential of processing written language through its visual form, rather than as sequences of discrete tokens. A notable example is PIXEL (Rust et al., 2022), a tokenizer-free language model that processes text as images. PIXEL renders text into fixed-size grayscale images, partitions them into non-overlapping patches, and processes these patches with a Vision Transformer (ViT) backbone trained using a masked patch prediction objective. By bypassing language-specific tokenization, PIXEL inherently supports script-agnostic processing and facilitates cross-lingual transfer without relying on a shared subword vocabulary. Empirical results across multiple multilingual benchmarks (Rust et al., 2022) show that PIXEL achieves competitive performance, demonstrating robustness to diverse scripts, orthographic variation, and visual distortions. These findings suggest that visual-based models such as PIXEL can also serve as effective backbones for CLTL tasks.

Given that both semantic-based models and vision-based models are capable of supporting transfer learning, in this work, we ask a simple question: Is cross-lingual transfer learning more effective when models focus on semantic meaning, like mBERT, or when they focus on the visual form of text, such as the shapes of scripts and glyphs, like PIXEL? To answer this question, we compare semantic-based and shape-based transfer in a CLTL setting. In the setting, the source task is sentiment classification in five languages: Chinese, Japanese, English, French, and Spanish, while the target task is Chinese offensive language classification. We conduct experiments with both mBERT and PIXEL, followed by a comparative analysis of their performance across model types and writing scripts.

Our key findings are summarized as follows:

1. mBERT outperforms PIXEL in overall accuracy, indicating the superior effectiveness of semantic information in transfer learning.

2. PIXEL transfers effectively for visually similar language pairs and performs robustly un-

der limited semantic information by exploiting glyph cues.

2 Related Work

2.1 Cross-Lingual Transfer learning

Transfer learning (Pan and Yang, 2010) is a widely adopted machine-learning methodology in which a model trained on a source task or domain is reused and adapted to improve performance on a related target task or domain. Cross-lingual transfer learning (CLTL) constitutes a transfer learning paradigm that emphasizes the transfer of knowledge between two distinct languages. This approach has been recognized as an effective framework for tasks such as offensive language detection. Ranasinghe and Zampieri (2020) and Montariol et al. (2022) employed zero-shot cross-lingual transfer in their respective studies, demonstrating its capability to facilitate knowledge transfer to previously unseen languages. However, Nozza (2021) highlights that such zero-shot transfer may incur performance degradations and be susceptible to cultural and lexical pitfalls, particularly when transferring from English to other languages without access to labeled data in the target language. In contrast, De la Peña Sarracén et al. (2023), Röttger et al. (2022), and Caselli and Plaza-del Arco (2025) investigated the few-shot setting, wherein the model is provided with a limited number of labeled examples in the target language to enhance adaptation and mitigate transfer-related deficiencies.

The relationship between the source and target languages has been extensively examined in the context of CLTL. Blaschke et al. (2025) analyzed 263 languages across three NLP tasks and conclude that choosing the similarity measure based on the task is important, with lexical similarity being most predictive for lexicon-heavy tasks. Lim et al. (2024) examined multiple source–target setups and found that multi-source transfer performs best when at least one typologically close language is combined with several diverse sources, while random selection can be suboptimal. Lin et al. (2024) introduced a model-embedding-based similarity metric and showed it predicts cross-lingual transfer performance better than typology-based metrics, especially in low-resource settings. Prior findings indicate that higher similarity between the source and target languages can enhance the effectiveness of cross-lingual transfer.

2.2 Vision Transformer

As we introduced in section 1, PIXEL (Rust et al., 2022) is one of the representatives of Vision Transformer (ViT)-based models. Original PIXEL only pretrained on English corpora, PIXEL-M4 (Kesen et al., 2025) extended the approach through multilingual pretraining on four visually and linguistically diverse languages: English, Hindi, Ukrainian, and Simplified Chinese, yielding substantially improved performance and cross-script transfer in non-Latin scripts compared to its monolingual counterparts. The Vision Transformer (ViT), originally proposed by Dosovitskiy et al. (2020), has inspired a wide range of subsequent works, including ViLT (Kim et al., 2021), ALBEF (Li et al., 2021), and PaLI (Chen et al., 2022).

Vision Transformer (ViT) concepts have been increasingly applied in multimodal NLP tasks. Kim et al. (2021) proposed ViLT, a convolution-free vision-and-language model that achieves competitive performance on Visual Question Answering, image–text retrieval, and visual reasoning, while being substantially more efficient than prior visual models. Ganz et al. (2024) presented QA-ViT, a Question-Aware Vision Transformer that embeds question-specific awareness directly within the vision encoder, allowing dynamic visual feature adaptation to queries and achieving consistent improvements across various multimodal reasoning tasks. Chochlakis et al. (2022) introduced VAuLT, extending ViLT with BERT to enhance semantic representations in multimodal sentiment analysis, improving sentiment prediction accuracy.

3 Zero-shot CLTL from Sentiment to Offensiveness

In this study, we explain the details of zero-shot cross-lingual transfer learning from sentiment analysis to offensive language detection across multiple languages, as illustrated in Figure 1.

First of all, our experimental design follows the inductive transfer learning paradigm defined by Pan and Yang (2010), wherein the source and target tasks differ. We focus on the transfer from sentiment classification (source task) to offensive language detection (target task). In sentiment classification, each sentence is labeled as either “positive” or “negative”. In offensive language detection, sentences are labeled as either “non-offensive” or “offensive”. We adopt a label alignment assumption, mapping negative sentiment to the offensive

Dataset	language	offensive	non offensive	total
COLD	Chinese	18041	19439	37480
Dataset	**language**	**negative**	**positive**	**total**
WeiboSenti	Chinese	10000	10000	20000
WRIME	Japanese	11604	10834	22438
Sentiment140	English	9980	9968	19948
French_multi	French	10000	10000	20000
Spanish_multi	Spanish	9994	9976	19970

Table 1: Data statistics of datasets in the experiment

class and positive sentiment to the non-offensive class. This mapping is consistent with prior work leveraging sentiment features for offensive or sarcasm language detection (Islam, 2024; Husain and Uzuner, 2021).

Under this alignment, we fine-tune multilingual models on sentiment datasets and evaluate them directly on offensive language detection without exposure to offensive data during training, constituting a zero-shot transfer setting. We employ sentiment datasets in 5 source languages: Chinese, Japanese, English, French, and Spanish. Evaluation is performed exclusively on a Chinese offensive language test set.

For each experiment, we fine-tune a model using a single source language, without combining datasets across languages. In addition to zero-shot transfer, we establish a supervised baseline by fine-tuning the models on Chinese offensive language training data and evaluating them on the same Chinese offensive language test set.

We experiment with two multilingual pre-trained models: mBERT and multilingual PIXEL (PIXEL-M4). Both models are pre-trained on corpora including Chinese and English; however, PIXEL-M4 does not include Japanese, French, or Spanish in its pretraining, whereas mBERT does. Hyperparameter configurations for fine-tuning both models are reported in Table 5 and Table 6 in the Appendix A. Fine-tuning parameters are kept consistent across all source languages and the baseline.

4 Data Collection

We now describe the datasets used in our experiments. Our setup requires both offensive language data for target evaluation and multilingual sentiment datasets for source languages.

For the Chinese offensive language dataset, we utilize the COLD dataset (Deng et al., 2022), a

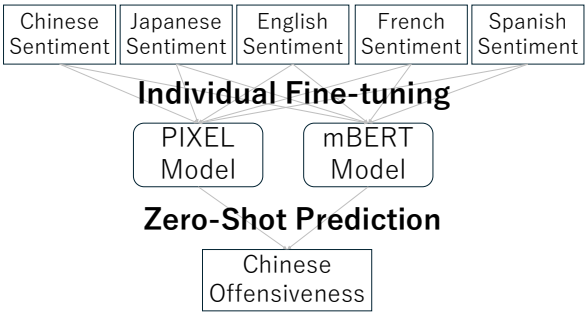

Figure 1: Zero-shot CLTL from Sentiment to Offensiveness. We fine-tune the PIXEL and mBERT with sentiment data in different languages individually and conduct zero-shot prediction on Chinese offensiveness.

publicly available corpus comprising 37,480 social media comments annotated with binary offensive labels. Following the original data partitioning scheme proposed by the authors, we divide the corpus into training, development, and test subsets. For the baseline configuration, the mBERT and PIXEL models were trained on the training set, with hyperparameters optimized using the development set of the offensive language dataset, while the test set is reserved exclusively for final evaluation. The test set comprises 5,323 instances from the offensive language dataset and is consistently employed across both the baseline and all transfer learning settings.

Then we introduce the sentiment classification datasets for five languages. The Chinese sentiment data is drawn from the work of (Wan et al., 2020), named WeiboSenti. For Japanese, we adopt the WRIME (Kajiwara et al., 2021) dataset, which provides labeled sentiment annotations for Japanese texts. English sentiment data is obtained from the widely used Sentiment140 (Go et al., 2009) dataset, developed by a research team at Stanford University. For French and Spanish, we aggregate samples from multiple publicly available sources, guided by documentation provided by

Language	PIXEL-M4	mBERT
Baseline	0.7367	0.7950
Chinese	0.4885(▼0.2482)	0.5784(▼0.2166)
English	0.4708(▼0.2659)	0.5533(▼0.2417)
Japanese	0.4582(▼0.2785)	0.5525(▼0.2425)
French	0.3900(▼0.3467)	0.6084(▼0.1866)
Spanish	0.3472(▼0.3895)	0.4880(▼0.3070)

Table 2: F1-Score (Macro Average for two classes) for PIXEL-M4 and mBERT on zero-shot sentiment transfer. The symbol ▼ denotes a performance decrease relative to the Baseline.The baseline is trained with the target data set, namely the offensive language data. In other words, this essentially corresponds to the upper bound for this target task against zero-shot approaches from sentiment analysis.

Brand24/mms (Łukasz Augustyniak et al., 2023) and named them French_multi and Spanish_multi. The French data is drawn from the datasets of (Narr et al., 2012; Keung et al., 2020). The Spanish data is sourced from datasets of (Cruz et al., 2008; Keung et al., 2020; Keith Norambuena et al., 2019; Patwa et al., 2020; Mozetič et al., 2016).

To ensure fair cross-lingual comparison, we control the size of each sentiment dataset to approximately 20,000 examples per language. For Chinese and English, where the sentiment data originates from a single source, we randomly sample subsets from the original datasets. For French and Spanish, where data is drawn from multiple sources, we allocate samples proportionally to achieve a total of 20,000 instances per language. For Japanese, we use the entire original dataset, as its size is already close to the target. During data collection, we balance the label distribution within each dataset to achieve an approximately 1:1 ratio between positive and negative sentiment classes, thereby mitigating potential bias from class imbalance in training. After data collection, we apply a simple preprocessing pipeline to remove hashtags, emojis, and URLs. Some instances are discarded during this process, and the final dataset statistics used in our experiments are reported in Table 1.

We follow the original data splits provided by the authors for COLD, using their train, evaluation, and test sets. For the sentiment analysis datasets (WeiboSenti, WRIME, Sentiment140, French_multi, and Spanish_multi), we partition the data into training (80%) and evaluation (20%) subsets, as no test data of sentiment is required.

5 Experiment Results and Analysis

The experimental results are presented in Table 2. We conduct 3 runs under the same experimental setting and report the average performance across 3 runs as the final result. We report only the macro-averaged F1 score over the two classes (offensive and non-offensive) in Table 2 for analysis. Additional details, including training statistics and other classification metrics, are provided in Table 7 in Appendix B and Table 8 in Appendix C. The Baseline setting corresponds to fine-tuning the models on the COLD training set and evaluating them on the COLD test set. Names of the languages correspond to the source sentiment datasets in Table 1.

Across all transfer scenarios, we observe a decrease in performance relative to the baseline. This performance degradation can be attributed to task mismatch or language differences between the source and target. For example, a conceptual and label mismatch leads to inevitable information loss: 0.2482 for PIXEL-M4 and 0.2166 for mBERT, even in the same language transfer setting from Chinese sentiment to Chinese offensiveness.

Comparing the two models, we first observe that the baseline performance of mBERT surpasses that of PIXEL. Furthermore, mBERT consistently exhibits smaller performance decrease than PIXEL-M4. For instance, in the English case, the decrease is 0.2417 for mBERT compared to 0.2659 for PIXEL-M4, and this pattern holds for all languages. This observation suggests that semantic information may play a more robust and critical role in cross-lingual transfer than visual information.

Although PIXEL-M4 underperformed in terms of F1-score, this outcome does not necessarily imply that visual information is useless in transfer learning. From a cross-lingual perspective, languages may share substantial visual similarity in their orthographic forms. For example, Chinese and Japanese share a large inventory of Chinese characters, many of which are identical or visu-

Dataset	original	replaced with radical	label
COLD	真恶心啊那个男的 (That guy is disgusting)	具亚心口二丨力勺	offensive
WeiboSenti	人口就是多国庆出门 (Many people go out on National Day)	人口京日夕口大凵门	positive
WRIME	友達コロナ疑惑とか焦るー (My friend is worried about coronavirus)	又辶コロナ疋心とか灬るー	negative

Table 3: Examples of replacement with radicals in Chinese and Japanese datasets. Inside the () is the meaning of the sentences.

Dataset	PIXEL-M4	mBERT
COLD	0.7367	0.7950
COLD_radicals	0.5360 (▼ 0.2007)	0.4789 (▼ 0.3161)
WeiboSenti	0.4885	0.5784
WeiboSenti_radicals	0.4670 (▼ 0.0215)	0.4368 (▼ 0.1416)
WRIME	0.4582	0.5525
WRIME_radicals	0.5002 (▲ 0.0420)	0.5287 (▼ 0.0238)

Table 4: F1-Score (Macro Average for two classes) for Ablation study: replacement with radicals. The symbol ▼ denotes a performance decrease relative to the no replacement. The symbol ▲ denotes performance increase.

ally similar. Likewise, English, French, and Spanish employ closely related Latin scripts with only minor orthographic variations. Focusing on the PIXEL-M4 model alone, we find evidence that visually similar languages can still yield better results. PIXEL-M4 was pretrained on Chinese and English; accordingly, we compare transfer performance from these two languages and observe that transfer from Chinese sentiment achieves higher accuracy than from English (0.4885 vs. 0.4708). For languages not seen in PIXEL-M4 pretraining: Japanese, French, and Spanish, Japanese yields the highest performance (0.4582), followed by French (0.3900) and Spanish (0.3472). These results indicate that, for PIXEL-M4, transferring from visually similar languages (e.g., Chinese, Japanese) to Chinese offensiveness classification tends to be more effective.

In summary, while mBERT appears more robust overall likely due to its stronger semantic representation, results of PIXEL-M4 demonstrate that visual similarity remains a contributive factor in cross-lingual transfer. We further conduct a paired t-test and a Wilcoxon signed-rank test to assess the significance of the performance difference between the two models. For the paired t-test, the results are $t = -4.8359$ and $p = 0.0047$, while for the Wilcoxon signed-rank test, the results are $W = 0.0$ and $p = 0.0312$. Both tests indicate a statistically significant difference at the 0.05 level, thereby demonstrating the robustness and reliability of the observed results.

6 Ablation Study on Visual Similarity

As discussed in the previous section, visual similarity can be a contributing factor in cross-lingual transfer, though it is not decisive as semantic representation. We now investigate an extreme scenario in which the semantic content of sentences is removed, to assess whether visual information alone can still facilitate transfer learning.

To this end, we conduct an ablation study by replacing all Chinese characters with their corresponding radicals. A radical in Chinese is a subcharacter component that often contributes to the meaning or pronunciation of the full character, but is not sufficient on its own to convey the original word’ s semantic meaning. Importantly, radicals preserve substantial visual information, such as stroke patterns, visual structure, and positional arrangement, thereby maintaining a high degree of script-level similarity to the original characters even when semantic content is removed.

We apply this replacement procedure to all 3 datasets containing Chinese characters, namely COLD, WeiboSenti, and WRIME. It is worth noting that a single Chinese character (kanji in Japanese) may contain multiple components that can be interpreted as radicals. In our experiments, we employ the RadicalFinder module from the

cjkradlib [1] Python library and select the first candidate radical for a Chinese character suggested by the tool. For COLD, we replace only the training set while keeping the original test set unchanged. For WeiboSenti and WRIME, we replace the entire datasets, as no test data is required from these datasets. We replace only the texts, while preserving their original sentiment and offensiveness labels. Some examples of the replacement are illustrated in Table 3. The modified sentences become semantically uninterpretable while preserving visual similarity to the original languages. These radical scripts can be regarded as a pseudo-language devoid of semantic content, and thus treated as the source languages in the zero-shot CLTL framework described in Section 3. For COLD, We fine-tune the models on the radical-based training data and evaluate them on the original test set. For WeiboSenti and WRIME, we apply the same procedure as in Section 3, fine-tuning on radical-based scripts and evaluating on the original COLD test set.

The experimental results are presented in Table 4. Performance is measured using the macro-averaged F1-score over the two classes (offensive and non-offensive), consistent with Table 2. The modified pseudo-scripts data are named _radicals after their original datasets. We observe that replacing characters with radicals generally leads to performance degradation across all three datasets for both models. For instance, on COLD, PIXEL-M4 experiences a decrease of 0.2007, while mBERT suffers a greater decrease of 0.3161. All settings show performance decreases, except for WRIME with PIXEL-M4, where we observe a slight improvement (+0.0420). These findings reinforce the conclusion that semantic meaning is critical for both PIXEL-M4 and mBERT. Although PIXEL-M4 is effective in capturing script-level features, it remains susceptible to semantic information loss, which in turn degrades performance.

Interestingly, when comparing models within each dataset, we find the opposite trend from the previous section: PIXEL-M4 consistently exhibits less performance decrease than mBERT in the case of radical replacement. For example, in WeiboSenti_raidcals, the decrease of PIXEL-M4 is only 0.0215, whereas the decrease of mBERT is 0.1416. In the WRIME_radicals, the performance of PIXEL-M4 even improves with the radical replacement, which warrants further investigation. These results suggest that when semantic meaning is absent, visual similarity can still be leveraged for transfer learning, and visual-based models such as PIXEL are more robust in such cases. Consequently, visual similarity between source and target languages may serve as an important criterion when selecting language pairs for cross-lingual transfer, particularly in scenarios where semantic information is scarce or unavailable.

[1] https://pypi.org/project/cjkradlib/

7 Conclusion

In this paper, we examine the impact of transfer learning on two multilingual models: the semantic-oriented mBERT and the visual-oriented PIXEL. We adopt a zero-shot cross-lingual transfer setting across five languages to compare their performance. Experimental results indicate that semantic information constitutes a reliable and effective basis for transfer learning, outperforming purely visual cues. However, a closer analysis of PIXEL reveals that it facilitates transfer particularly well between visually similar source and target language pairs, suggesting that visual information remains a non-negligible factor in cross-lingual transfer. Furthermore, our ablation study shows that even in the absence of semantic information, PIXEL can achieve robust transfer performance. This finding highlights the potential of visual information as a viable alternative for transfer learning in scenarios where semantic information is limited. Our current work focuses exclusively on standard quantitative classification metrics such as the F1 score. For future work, we aim to conduct a detailed error analysis to identify qualitative linguistic patterns (e.g, failure cases and script-specific phenomena), thereby providing deeper insights into the role and impact of visual information.

Limitations

This study has several limitations. Most notably, the sentiment-to-offensiveness transfer setting relies on the simplifying assumption that offensive language is conceptually equivalent to negative sentiment. While there is empirical overlap between the two, this assumption may not capture the full complexity of offensive language, which can include sarcastic, provocative, or contextually ambiguous expressions that are not necessarily associated with negative sentiment. Future work should aim to develop stronger transfer settings, poten-

tially involving intermediate tasks, richer label taxonomies, or adversarial adaptation strategies, to more accurately bridge the gap between sentiment and offensiveness.

Moreover, in the current study, we focus exclusively on the case of sentiment-to-offensiveness transfer across five languages. To enable a more comprehensive analysis, future work should extend the experiments beyond this setting to encompass additional transfer scenarios and a broader range of languages.

Acknowledgments

This work was supported by JST, the establishment of university fellowships towards the creation of science technology innovation, Grant Number JPMJFS2133.

References

Verena Blaschke, Masha Fedzechkina, and Maartje Ter Hoeve. 2025. Analyzing the effect of linguistic similarity on cross-lingual transfer: Tasks and experimental setups matter. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8653–8684, Vienna, Austria. Association for Computational Linguistics.

Tommaso Caselli and Flor Miriam Plaza-del Arco. 2025. Learning from disagreement: Entropy-guided few-shot selection for toxic language detection. In *Proceedings of The 9th Workshop on Online Abuse and Harms (WOAH)*, pages 53–66, Vienna, Austria. Association for Computational Linguistics.

Xi Chen, Xiao Wang, Soravit Changpinyo, Anthony J Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, and 1 others. 2022. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*.

Georgios Chochlakis, Tejas Srinivasan, Jesse Thomason, and Shrikanth Narayanan. 2022. Vault: Augmenting the vision-and-language transformer for sentiment classification on social media. *arXiv preprint arXiv:2208.09021*.

Fermin L Cruz, Jose A Troyano, Fernando Enriquez, and Javier Ortega. 2008. Experiments in sentiment classification of movie reviews in spanish. *Procesamiento del Lenguaje Natural*, 41:73–80.

Gretel De la Peña Sarracén, Paolo Rosso, Robert Litschko, Goran Glavaš, and Simone Ponzetto. 2023. Vicinal risk minimization for few-shot cross-lingual transfer in abusive language detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4069–4085, Singapore. Association for Computational Linguistics.

Jiawen Deng, Jingyan Zhou, Hao Sun, Chujie Zheng, Fei Mi, Helen Meng, and Minlie Huang. 2022. COLD: A benchmark for Chinese offensive language detection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11580–11599, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929.

Roy Ganz, Yair Kittenplon, Aviad Aberdam, Elad Ben Avraham, Oren Nuriel, Shai Mazor, and Ron Litman. 2024. Question aware vision transformer for multimodal reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13861–13871.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 1(12):2009.

Fatemah Husain and Ozlem Uzuner. 2021. Leveraging offensive language for sarcasm and sentiment detection in arabic. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 364–369.

Khondoker Ittehadul Islam. 2024. Leveraging sentiment for offensive text classification. *arXiv preprint arXiv:2412.17825*.

Tomoyuki Kajiwara, Chenhui Chu, Noriko Takemura, Yuta Nakashima, and Hajime Nagahara. 2021. WRIME: A new dataset for emotional intensity estimation with subjective and objective annotations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2095–2104, Online. Association for Computational Linguistics.

Brian Keith Norambuena, Exequiel Lettura, and Claudio Villegas. 2019. Sentiment analysis and opinion mining applied to scientific paper reviews. *Intelligent Data Analysis*, 23:191–214.

Ilker Kesen, Jonas F. Lotz, Ingo Ziegler, Phillip Rust, and Desmond Elliott. 2025. Multilingual pretraining for pixel language models. *ArXiv*, abs/2505.21265.

Phillip Keung, Yichao Lu, György Szarvas, and Noah A. Smith. 2020. The multilingual Amazon reviews corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4563–4568, Online. Association for Computational Linguistics.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, pages 5583–5594. PMLR.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705.

Seong Hoon Lim, Taejun Yun, Jinhyeon Kim, Jihun Choi, and Taeuk Kim. 2024. Analysis of multi-source language training in cross-lingual transfer. *Preprint*, arXiv:2402.13562.

Peiqin Lin, Chengzhi Hu, Zheyu Zhang, Andre Martins, and Hinrich Schuetze. 2024. mPLM-sim: Better cross-lingual similarity and transfer in multilingual pretrained language models. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 276–310, St. Julian's, Malta. Association for Computational Linguistics.

Syrielle Montariol, Arij Riabi, and Djamé Seddah. 2022. Multilingual auxiliary tasks training: Bridging the gap between languages for zero-shot transfer of hate speech detection models. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 347–363, Online only. Association for Computational Linguistics.

Igor Mozetič, Miha Grčar, and Jasmina Smailović. 2016. Multilingual twitter sentiment classification: The role of human annotators. *PLOS ONE*, 11(5):1–26.

Sascha Narr, Michael Hülfenhaus, and Sahin Albayrak. 2012. Language-independent twitter sentiment analysis. In *Workshop on Knowledge Discovery, Data Mining and Machine Learning (KDML-2012)*.

Debora Nozza. 2021. Exposing the limits of zero-shot cross-lingual hate speech detection. 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 907–914, Online. Association for Computational Linguistics.

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22:1345–1359.

Parth Patwa, Gustavo Aguilar, Sudipta Kar, Suraj Pandey, Srinivas PYKL, Björn Gambäck, Tanmoy Chakraborty, Thamar Solorio, and Amitava Das. 2020. SemEval-2020 task 9: Overview of sentiment analysis of code-mixed tweets. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 774–790, Barcelona (online). International Committee for Computational Linguistics.

Tharindu Ranasinghe and Marcos Zampieri. 2020. Multilingual offensive language identification with cross-lingual embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5838–5844, Online. Association for Computational Linguistics.

Paul Röttger, Debora Nozza, Federico Bianchi, and Dirk Hovy. 2022. Data-efficient strategies for expanding hate speech detection into under-resourced languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5674–5691, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Phillip Rust, Jonas F. Lotz, Emanuele Bugliarello, Elizabeth Salesky, Miryam de Lhoneux, and Desmond Elliott. 2022. Language modelling with pixels. *ArXiv*, abs/2207.06991.

Shuo Wan, Bohan Li, Anman Zhang, Wenhuan Wang, and Donghai Guan. 2020. S2ap: Sequential sentiweibo analysis platform. In *Database Systems for Advanced Applications: 25th International Conference, DASFAA 2020, Jeju, South Korea, September 24–27, 2020, Proceedings, Part III*, page 745–749, Berlin, Heidelberg. Springer-Verlag.

Łukasz Augustyniak, Szymon Woźniak, Marcin Gruza, Piotr Gramacki, Krzysztof Rajda, Mikołaj Morzy, and Tomasz Kajdanowicz. 2023. Massively multilingual corpus of sentiment datasets and multi-faceted sentiment classification benchmark. *Preprint*, arXiv:2306.07902.

A Fine-tuning Parameters

Parameters	Values
Rendering backend	PyGame
Classification head pooling	Mean
Optimizer	AdamW
Adam β	(0.9, 0.999)
Adam ϵ	1e-8
Weight decay	0
Learning rate	3e-5
Learning rate warmup steps	100
Learning rate schedule	Linear decay
Max sequence length	529
Batch size	64
Max steps	15000
Early Stopping	Active
Eval interval	100 steps
Dropout probability	0.1

Table 5: Fine-tuning parameters for PIXEL models.

Parameters	Values
Classification head pooling	CLS embedding
Optimizer	AdamW
Adam β	(0.9, 0.999)
Adam ϵ	1e-8
Weight decay	0
Learning rate	3e-5
Learning rate warmup steps	100
Learning rate schedule	Linear decay
Max sequence length	256
Batch size	64
Max steps	15000
Early Stopping	Active
Eval interval	100 steps
Dropout probability	0.1

Table 6: Fine-tuning parameters for mBERT models.

B Best Evaluation F1 during Training

Datasets	mBERT	PIXEL-M4
WeiboSenti	0.8845	0.7851
WRIME	0.8527	0.7304
Sentiment140	0.8039	0.6657
French_multi	0.9231	0.7960
Spanish_multi	0.7782	0.6878

Table 7: Best F1 score recorded when evaluating the training effect using sentiment evaluation data. The checkpoint corresponding to the best F1 score is saved at the end of training.

C Full Metrics for Experiment Results

Dataset	Model	Precision (Mean)	Precision (Std)	Recall (Mean)	Recall (Std)	F1-score (Mean)	F1-score (Std)
COLD	mbert	0.7961	0.0030	0.8090	0.0034	0.7950	0.0046
COLD	pixel	0.7395	0.0051	0.7501	0.0055	0.7367	0.0062
WeiboSenti	mbert	0.6651	0.0045	0.6375	0.0137	0.5784	0.0259
WeiboSenti	pixel	0.5972	0.0067	0.5683	0.0082	0.4885	0.0181
Sentiment140	mbert	0.5621	0.0112	0.5589	0.0177	0.5533	0.0157
Sentiment140	pixel	0.5109	0.0119	0.5105	0.0108	0.4708	0.0698
WRIME	mbert	0.6591	0.0046	0.6221	0.0053	0.5525	0.0109
WRIME	pixel	0.4986	0.0154	0.5002	0.0065	0.4582	0.0357
French_multi	mbert	0.6214	0.0099	0.6241	0.0095	0.6084	0.0032
French_multi	pixel	0.4551	0.0306	0.4966	0.0010	0.3900	0.0097
Spanish_multi	mbert	0.6841	0.0035	0.5972	0.0249	0.4880	0.0497
Spanish_multi	pixel	0.5909	0.0306	0.5178	0.0145	0.3472	0.0557

Table 8: Macro Avg Metrics: Mean and Standard Deviation(Std) of Precision, Recall, and F1 score over 3 Runs of the CLTL experiment.