

Quantifying Cross-Lingual Sentiment Prediction Difference in Bilingual Pop-Culture Reviews via Dual Monolingual Transformers and Jensen-Shannon Divergence

Azfani Naurotul Jannah¹, Triando Hamonangan Saragih², Irwan Budiman³, Muliadi⁴,
Radityo Adi Nugroho⁵

^{1,2,3,4,5}Faculty of Mathematics and Natural Science, Department of Computer Science, Lambung Mangkurat University, Kalimantan, Indonesia

Article Info

Article history:

Received July 16, 2026

Revised August 03, 2026

Accepted August 20, 2026

Keywords:

Cross-lingual sentiment analysis

Machine translation

Chinese-RoBERTa

IndoBERT

Jensen-Shannon Divergence

ABSTRACT

Machine translation has been widely used to support cross-lingual sentiment analysis when labeled data in the target language are limited. However, differences in linguistic representation between the source text and its translation may affect the probabilistic outputs of sentiment classification models. This study investigated differences in sentiment predictions between original Mandarin reviews and their Indonesian translations using two independently fine-tuned monolingual transformer models. Chinese-RoBERTa-wwm-ext was applied to the original Mandarin reviews, whereas IndoBERT was used to classify the Indonesian translations. Jensen-Shannon Divergence was employed to compare the sentiment probability distributions generated by the two models. The results showed that Chinese-RoBERTa achieved an accuracy of 74%, whereas IndoBERT achieved 69% on the pseudo-labeled evaluation dataset. Furthermore, 73.81% of the review pairs retained consistent sentiment predictions, while 26.19% exhibited prediction shifts, with Polarity Amplification being the most frequently observed category and most transitions occurring between adjacent sentiment classes. The probability-distribution analysis also revealed substantial differences in prediction confidence for some review pairs, even when the predicted sentiment labels remained identical. These findings demonstrated that comparing probability distributions provided complementary information beyond label-based evaluation for analyzing prediction differences between independently trained monolingual sentiment models on bilingual review pairs.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Triando Hamonangan Saragih,
Faculty of Mathematics and Natural Science,
Department of Computer Science,
Lambung Mangkurat University,
Kalimantan, Indonesia
Email: triando.saragih@ulm.ac.id

1. INTRODUCTION

Sentiment analysis has become one of the primary approaches in Natural Language Processing (NLP) for extracting users' opinions and perceptions from textual data generated across various digital platforms [1], [2]. The rapid growth of multilingual digital content has expanded the application of sentiment analysis beyond monolingual settings, making cross-lingual sentiment analysis increasingly important for organizations and researchers seeking to understand public opinion across diverse languages [3]. In practice, however, collecting large-scale sentiment datasets for every language is often costly and time-consuming. Consequently, Machine Translation (MT) has become a practical solution for translating source-language texts into a target language before

performing sentiment classification, enabling existing monolingual sentiment models to be utilized without requiring extensive manual annotation in every language [3], [4].

Although modern Neural Machine Translation (NMT) systems have significantly improved translation fluency, preserving the semantic and affective information of the source text remains a challenge. Language-specific expressions, culture-dependent terminology, implicit meanings, and contextual nuances may not always be conveyed in exactly the same manner after translation [5]. Such linguistic differences may cause downstream sentiment classification models to produce different probability distributions or prediction outcomes for aligned source and translated texts, even when both texts remain understandable to human readers [6]. Therefore, evaluating only the final sentiment label may overlook more subtle differences in prediction confidence and class-level probability distributions.

Recent studies on cross-lingual sentiment analysis have largely relied on multilingual transformer models, such as mBERT and XLM-R, because they enable knowledge transfer across multiple languages within a unified representation space [3]. Nevertheless, previous studies have shown that the effectiveness of multilingual representations is influenced by linguistic diversity, cultural expressions, and pragmatic context, particularly when transferring sentiment across typologically different languages [7], [8]. In contrast, the present study employs independently fine-tuned monolingual transformer models for Mandarin and Indonesian. Chinese-RoBERTa-wwm-ext is used to process the original Mandarin reviews, whereas IndoBERT is used to process their Indonesian translations. This design allows each text version to be analyzed using a model specifically pretrained for its respective language. However, because the two models differ in their pretrained representations, tokenization mechanisms, and language-specific decision boundaries, differences between their outputs cannot be attributed exclusively to machine translation [9].

Previous studies have generally evaluated machine translation in cross-lingual text and sentiment classification using translation-quality metrics or downstream classification metrics, such as accuracy, precision, recall, and F1-score [10], [11]. Artetxe et al. compared translate-test and translate-train strategies and demonstrated that machine-translation-based cross-lingual classification is commonly assessed through downstream task performance [12]. Similarly, Al Amer et al. evaluated the effects of machine translation using accuracy and weighted F1-score [12], [13], while Shode et al. assessed multilingual movie-review sentiment classification primarily through accuracy and the preservation of sentiment-related words [14]. Although these approaches provide valuable evidence regarding overall classification effectiveness, their reliance on final labels and aggregate performance metrics offers limited insight into differences in prediction confidence across sentiment classes. Moreover, Qian et al. emphasized that conventional translation metrics do not adequately account for emotion, slang, irony, and other affective characteristics of user-generated content [15]. Therefore, relatively limited attention has been given to frameworks that jointly examine sentiment-label transitions, probability-distribution divergence, and qualitative differences between aligned source and translated reviews. This methodological limitation constitutes the research gap addressed in the present study.

To address the research gap identified above, this study proposes a hybrid analytical framework for examining cross-lingual sentiment prediction differences in translated reviews. Two independently fine-tuned monolingual transformer models are employed, namely Chinese-RoBERTa-wwm-ext for the original Mandarin reviews and IndoBERT for their Indonesian translations. Instead of relying solely on final sentiment labels or aggregate classification metrics, the proposed framework integrates three complementary analyses: (1) sentiment classification performance evaluation, (2) Jensen–Shannon Divergence (JSD) to quantify differences between the sentiment probability distributions generated by the two models, and (3) qualitative examination of representative review pairs exhibiting high probabilistic divergence. Furthermore, sentiment-label transitions are categorized into Consistent, Polarity Amplification, Polarity Dilution, and Polarity Reversal to provide a more comprehensive characterization of cross-lingual prediction differences.

Accordingly, this study has four objectives: (1) to evaluate the sentiment classification performance of Chinese-RoBERTa-wwm-ext and IndoBERT using a pseudo-labeled evaluation dataset; (2) to quantify probabilistic divergence between aligned Mandarin and Indonesian review pairs using Jensen–Shannon Divergence; (3) to analyze sentiment-label transition patterns between

the original and translated reviews; and (4) to qualitatively investigate representative high-divergence review pairs in order to identify linguistic characteristics associated with polarity amplification, polarity dilution, polarity reversal, and other cross-lingual prediction discrepancies. The findings are expected to provide a more comprehensive understanding of cross-lingual sentiment prediction differences in translated reviews while acknowledging that the observed divergence may reflect both translation-related linguistic variation and differences between the two independently trained monolingual models.

2. RESEARCH METHOD

This study employs a comparative experimental approach based on bilingual parallel text data to compare the sentiment predictions produced by independently fine-tuned monolingual models for aligned Mandarin–Indonesian reviews. The computational workflow was systematically organized into several stages, including data collection, text preprocessing, data splitting, model fine-tuning, performance evaluation of the monolingual models, comparison of sentiment probability distributions using Jensen–Shannon Divergence (JSD), and qualitative analysis of representative review pairs with high JSD values. The chronological workflow of the entire experiment is presented in the form of a research flowchart in Figure 1.

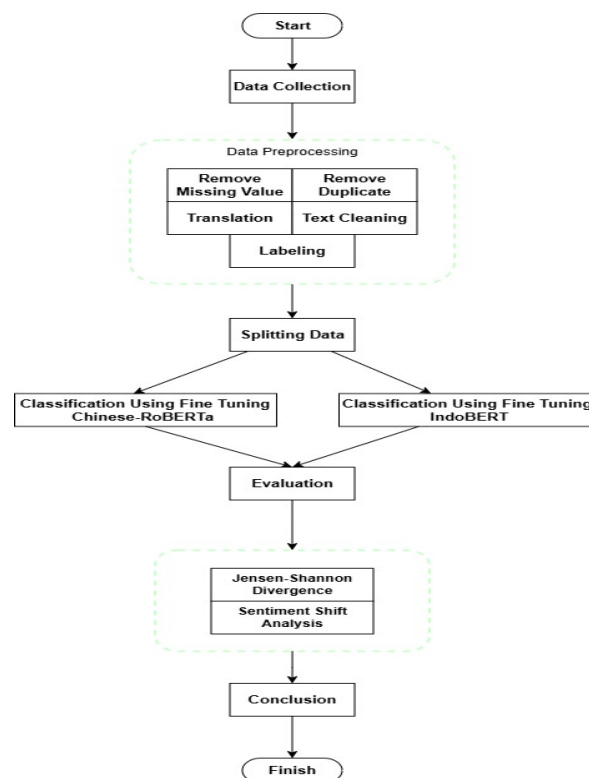


Figure 1. Flow of Research

2.1. Data Collection

The dataset used in this study consists of textual user reviews (*user-generated content*) collected via web crawling on the Tencent Video digital platform during May 2026. Data extraction was performed using a web scraper extension to collect popular review data in a structured manner. The extracted reviews were consolidated into a single spreadsheet, forming the initial dataset used throughout the subsequent preprocessing and experimental stages. The dataset was compiled from six popular Chinese animated series (*donghua*), namely *Sword of Coming*, *Renegade Immortal*, *Perfect World*, *Shrouding the Heavens*, *Lord of the Mysteries*, and *Throne of Seal*. These titles were selected because they represent active pop-culture communities with abundant user-generated reviews containing diverse linguistic characteristics, including internet slang, community-specific terminology, and context-dependent expressions. Through this digital

data harvesting procedure, this study successfully collected a total raw corpus of 4,340 lines of pure review text documents. The distribution of the collected reviews across the six selected donghua titles is summarized in Table 1, while an example of the raw review records obtained through the web crawling process is presented in Table 2.

Table 1. Six donghua

No	Title	Count
1.	Sword of Coming	457
2.	Renegade Immortal	796
3.	Perfect World	987
4.	Shrouding the Heavens	923
5.	Lord of the Mysteries	256
6.	Throne of Seal	921
Total		4340

Table 2. Sample data from scraping

name	date	review	like_count	reply_count
巴	2025-06-17·山东	千万年后，叶天帝想起了他的奔驰车，来到地球后，保安拦住了他：你的车我研究了一万年了，吾以车证道，名为懂车帝	7396	289
imma	2024-07-10·云南	第65话06:20建模没把裤子建模上	2099	326
★♀风之痕♀ ★	2024-01-10·浙江	没有对比，何来伤害？	1176	334

2.2. Data Preprocessing

The data preprocessing stage was designed to transform the collected Mandarin review dataset into a bilingual parallel corpus suitable for sentiment classification. Data preprocessing began by cleaning the Chinese text while preserving Hanzi characters. The preprocessing steps included removing URLs, mentions, and hashtags, eliminating excessive whitespace, and filtering out irrelevant characters [16]. Furthermore, records with empty or duplicate reviews were removed to ensure that only valid and unique instances were included in the subsequent analysis. This preprocessing step was intended to enhance dataset quality, minimize data redundancy, and facilitate the subsequent pseudo-labeling and translation processes.

After the Mandarin texts were cleaned, the data was labeled using a pseudo-labeling approach, in which a pre-trained transformer model automatically generated labels for previously unlabeled data [17]. In this stage, the IDEA-CCNL/Erlangshen-Roberta-110M-Sentiment model was employed because it is a pre-trained transformer specifically developed for Mandarin sentiment analysis and has been fine-tuned on multiple Chinese sentiment datasets [18]. The model is capable of capturing contextual semantic representations in Mandarin text and was therefore employed to generate initial sentiment labels [19]. The model's predictions were then used as initial labels consisting of three sentiment categories: positive, neutral, and negative.

After the pseudo-labeling process was completed, the labeled Mandarin reviews were translated into Indonesian using the Google Translate web service, which is based on Neural Machine Translation [20], [21]. The translation was performed in May 2026 using the Google Translate Documents feature by uploading a spreadsheet containing the complete collection of Mandarin reviews, with Mandarin specified as the source language and Indonesian as the target language. Because Google Translate is a continuously updated cloud-based service, the exact underlying translation model version could not be identified or controlled. The pseudo-labeling process was intentionally performed before translation so that sentiment labels were assigned to the original Mandarin reviews, thereby avoiding potential inconsistencies introduced by machine translation during the labeling stage. No manual post-editing or correction was performed after translation, and all translated reviews were used directly in the subsequent preprocessing stage. The translated texts were subsequently preprocessed before the next stage. This step involved removing punctuation generated during the translation process and applying case folding to convert all text to lowercase [22].

As a quality assurance procedure rather than a manual annotation process, 500 review pairs were randomly sampled from the dataset and manually examined by the first author before model

development. The review involved comparing the Indonesian translations generated by Google Translate with translations produced by several publicly available machine translation services (e.g., DeepL and Bing Translator). This comparison was intended to facilitate the interpretation of the overall semantic meaning of each review, thereby enabling an assessment of whether the automatically assigned sentiment labels remained consistent with the underlying sentiment expressed in the original text. Whenever discrepancies were identified between the pseudo-generated sentiment labels and the semantic content of the reviews, the corresponding labels were manually corrected. Conversely, labels that were found to be consistent with the review content were retained without modification. This quality assurance procedure was applied only to the randomly selected sample, while the remaining data retained their original pseudo-generated sentiment labels.

Whenever a pseudo-generated label was found to be inconsistent with the sentiment expressed in the review, the corrected label replaced the original pseudo-generated label in the main dataset. After completion of the quality assurance procedure, the updated dataset was partitioned into the training, validation, and testing subsets using the stratified random splitting procedure described in Section 2.3. Consequently, the manually corrected review pairs were not reserved as a separate human-verified evaluation set but were randomly distributed across the three subsets according to the predefined 80:10:10 split. Accordingly, no manually verified evaluation subset was constructed, and the corrected review pairs followed the same stratified random splitting procedure as the remaining instances.

Label assignment during the manual quality assurance process was conducted based on the sentiment tendency expressed in each review. Reviews were classified as positive if they contained expressions of appreciation or praise toward the film, its characters, storyline, animation quality, or other aspects, as well as indications of enthusiasm, recommendations, or encouragement for others to watch the film. Reviews were classified as negative if they contained criticism, dissatisfaction, complaints, offensive language, or other expressions reflecting negative evaluations of the film or its elements. Meanwhile, reviews were classified as neutral if they conveyed factual information without expressing a clear evaluative judgment, such as information about release dates, episode counts, or descriptive summaries of the storyline or film content without indicating either a positive or negative sentiment.

The manual quality assurance process was performed solely as a quality control measure on a subset of the pseudo-labeled data rather than as a manual annotation procedure for constructing a ground-truth dataset. Consequently, this process did not involve independent annotators, formal annotation guidelines, or the assessment of inter-annotator agreement. Although the manually corrected labels were incorporated into the final dataset before data splitting, most review pairs retained their automatically generated labels. Therefore, systematic labeling errors originating from the pseudo-labeling model may still remain in the training, validation, and testing subsets. Accordingly, the evaluation dataset should be interpreted as a pseudo-labeled benchmark rather than a fully human-annotated benchmark. An example of the bilingual dataset after preprocessing is presented in Table 3.

Table 3. Dataset after preprocessing

Mandarin clean	Indonesia clean	label name	label num
还泥腿子 搞笑 一群蝼蚁	berkaki lumpur itu konyol sekumpulan semut	negative	0
06集阿良帅炸裂 怎一个猛字了得	episode 6 ah liang benar benar luar biasa dia sungguh hebat	positive	2
这是真的吗	apakah ini benar	neutral	1

2.3. Splitting Data

The data splitting process was performed randomly using a stratification method to maintain a balanced distribution of sentiment classes across each data subset [23]. To ensure reproducibility of the data partitioning, the `train_test_split()` function was implemented with `random_state = 42`, thereby fixing the random seed used for generating the training, validation, and testing subsets. Following standard splitting methodology in neural network architecture, this

parallel bilingual dataset was strictly split using an 80:10:10 ratio. These proportions are specifically allocated as follows: 80% as the training set for optimizing the model's internal weights, 10% as the validation set to identify overfitting during the iterative process, and the remaining 10% as an isolated testing set to measure the model's generalization ability at the end of the experimental phase [24], [25].

2.4. Chinese-RoBERTa-wwm-ext

Model training was conducted using two separate pre-trained monolingual transformer-based architectures that functioned independently as expert models in their respective language spaces. This study employs the Chinese-RoBERTa-wwm-ext model to perform sentiment classification on Mandarin-language review texts. Chinese-RoBERTa-wwm-ext is a pre-trained language model specifically designed for Chinese natural language processing (NLP) tasks by integrating the RoBERTa architecture with the Whole Word Masking (WWM) strategy [19]. The Chinese-RoBERTa-wwm-ext architecture follows the RoBERTa Transformer encoder, which consists of an embedding layer and multiple Transformer encoder layers. Each encoder layer incorporates a multi-head self-attention mechanism, a position-wise feed-forward neural network, residual connections, and layer normalization to generate contextualized semantic representations for each input token [26]. Unlike the conventional BERT model, which masks text at the subword level, WWM masks entire words during pre-training, making it more suitable for the linguistic characteristics of the Chinese language [27]. Furthermore, the model has been pre-trained on a large-scale and diverse Chinese corpus, enabling it to achieve strong generalization performance across various NLP tasks, including sentiment analysis [19]. During fine-tuning and inference, the model generates contextualized semantic representations of input tokens, which are subsequently utilized to predict the sentiment class of each review. Figure 2 illustrates the architecture of the Chinese-RoBERTa-wwm-ext model employed in this study.

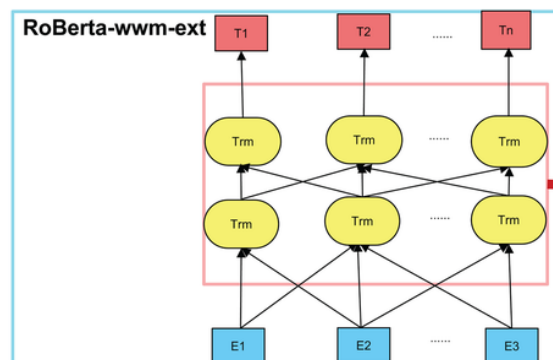


Figure 2. RoBERTa-wwm-ext Architecture

The pre-trained Chinese-RoBERTa-wwm-ext model was fine-tuned end-to-end for sentiment classification using the labeled Mandarin review dataset. Model training employed the AdamW optimizer with a learning rate of 2×10^{-5} , a batch size of 16, a maximum sequence length of 128 tokens [28]. The training process was conducted for 5 epochs with a warmup ratio of 0.1 and a weight decay of 0.01. Before selecting the final training configuration, several preliminary experiments were conducted to determine an appropriate number of training epochs. Three epoch settings (4, 5, and 6) were evaluated while maintaining a fixed learning rate of 2×10^{-5} , a batch size of 16, a warmup ratio of 0.1, and a weight decay of 0.01. Each hyperparameter configuration was trained and evaluated once. Epoch 5 was selected as the final training configuration because it achieved the best overall trade-off between the performance of the Chinese-RoBERTa-wwm-ext and IndoBERT models, maintaining the highest performance for Chinese-RoBERTa while avoiding the performance degradation observed in IndoBERT when trained for six epochs. Therefore, five training epochs were used for all subsequent experiments. Model training was implemented using the Hugging Face Trainer API while maintaining identical model architectures, hyperparameter settings, and data partitions across all experiments.

2.5. Classification using IndoBERT Algorithms

This study also employs the IndoBERT model to perform sentiment classification on the translated Indonesian-language dataset. IndoBERT is a pre-trained language model based on the Transformer architecture, specifically developed for various Natural Language Processing (NLP) tasks in the Indonesian language. The model was pre-trained on a large-scale Indonesian corpus, enabling it to generate contextual semantic representations that effectively capture the linguistic characteristics of Indonesian [29]. IndoBERT adopts the BERT-Base architecture, consisting of 12 Transformer encoder layers, 12 self-attention heads, and a hidden size of 768 dimensions [30].

In this study, the IndoBERT model was fine-tuned on a sentiment-labeled dataset of translated reviews. During the fine-tuning process, the contextual representation of the [CLS] token from the final hidden layer was employed as the aggregate representation of the input sentence and passed to a fully connected classification layer, followed by a Softmax activation function to generate the probability distribution over the sentiment classes [29]. Model optimization employed the AdamW optimizer with the same hyperparameter configuration as the Chinese-RoBERTa-wwm-ext model, including a learning rate of 2×10^{-5} , a batch size of 16, 5 training epochs, a warmup ratio of 0.1, and a weight decay of 0.01. To maintain consistency between both monolingual models, IndoBERT was fine-tuned using the same training configuration and hyperparameter settings as the Chinese-RoBERTa-wwm-ext model. Figure 3 presents the architecture of the IndoBERT model used for sentiment classification of the translated Indonesian reviews.

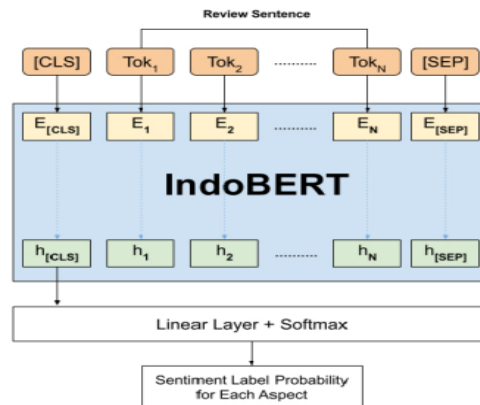


Figure 3. IndoBERT Architecture

2.6. Evaluation Metrics

The performance evaluation of the sentiment classification model in this study uses four key metrics: *accuracy*, *precision*, *recall*, and *F1-score*. These four metrics are commonly used in the evaluation of text classification models because they provide a comprehensive picture of a model's performance, especially on datasets with an imbalanced class distribution [31].

1. Accuracy

Accuracy measures the proportion of correct predictions relative to the total number of data points tested. This metric indicates the model's overall accuracy [32].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad \dots (1)$$

2. Precision

Precision measures the accuracy of positive predictions by comparing the number of true positives to the total number of positive predictions generated by the model [33].

$$Precision = \frac{TP}{TP + FP} \quad \dots (2)$$

3. Recall

Recall measures a model's ability to identify all data points that belong to the positive class [33].

$$Recall = \frac{TP}{TP + FN} \quad \dots (3)$$

4. F1-Score

The F1-score is the harmonic mean of precision and recall and provides a balanced measure of classification performance, particularly when the class distribution is uneven [34].

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad \dots (4)$$

2.7. Jensen-Shannon Divergence (JSD)

Jensen-Shannon Divergence (JSD) is a statistical measure used to calculate the difference between two probability distributions, derived from the Kullback–Leibler Divergence (KLD). Compared to KLD, JSD is symmetric and always has a bounded value, making it suitable for comparing probability distributions in various natural language processing and text analysis applications. [35].

In this study, Jensen–Shannon Divergence (JSD) was employed to quantify the difference between the probability distributions produced by the Chinese-RoBERTa-wwm-ext and IndoBERT sentiment classification models for each aligned Mandarin–Indonesian review pair. Each probability distribution consists of the predicted probabilities of the three sentiment classes (negative, neutral, and positive) generated by the respective monolingual models. JSD was computed directly from the softmax probability distributions produced by the independently fine-tuned Chinese-RoBERTa-wwm-ext and IndoBERT models without applying additional probability calibration. The JSD between two probability distributions P and Q is defined as follows [35]:

$$JSD(P, Q) = \frac{1}{2} D_{KL}(P||M) + \frac{1}{2} D_{KL}(Q||M) \quad \dots (5)$$

$$M = \frac{1}{2} (P + Q) \quad \dots (6)$$

D_{KL} is the Kullback–Leibler divergence and M is the mean of P and Q . The JSD value ranges from 0 to 1, with a value of 0 indicating that the two distributions are identical and values approaching 1 indicating a greater degree of difference between the distributions [35].

Following the JSD computation, a sentiment shift analysis was performed to identify changes in sentiment polarity between the original Mandarin reviews and their Indonesian translations. To facilitate the qualitative analysis, an operational threshold of $JSD > 0.5$ was adopted to identify review pairs exhibiting relatively high divergence. This threshold was adopted as an operational criterion for exploratory qualitative analysis rather than as a universal indicator of semantic change. A total of 39 review pairs satisfied this criterion and were subsequently examined qualitatively. The selected review pairs were subsequently analyzed by comparing the sentiment labels predicted by the Chinese-RoBERTa-wwm-ext and IndoBERT models. The prediction results were categorized into sentiment preservation, where both models predicted the same sentiment label, and sentiment shift, where different sentiment labels were predicted. Sentiment shifts were further classified into polarity amplification, polarity dilution, and polarity reversal according to the direction of the label changes between the two model predictions.

3. RESULTS AND DISCUSSION

3.1. Preprocessing and Splitting Data

The initial dataset consisted of 4,340 reviews. During the preprocessing stage, 146 duplicate entries and one empty review were removed, resulting in a final dataset of 4,193 review pairs for subsequent analysis. Duplicate records were removed to eliminate redundant information that could introduce bias during model training, while samples with missing textual content were

excluded to ensure that each instance contained sufficient textual information for processing by the sentiment classification model.

The label distribution shows that the positive class is the most prevalent, comprising 1,975 instances, followed by the negative class with 1,243 instances and the neutral class with 975 instances. This uneven distribution across the sentiment classes indicates the presence of a moderate class imbalance, although the imbalance is not substantial. The label distribution indicates the presence of class imbalance, with the positive class containing substantially more samples than the negative and neutral classes. To mitigate the impact of this imbalance during model training, this study employs a class-weighting strategy using the balanced option throughout the fine-tuning process. This strategy automatically assigned weights of 1.1247, 1.4333, and 0.7076 to the negative, neutral, and positive classes, respectively, thereby imposing greater penalties for misclassifying minority-class instances during optimization.

Table 4. Sentiment label distribution

label	Dataset count
Positive	1975
Negative	1243
Neutral	975

Table 4 presents the distribution of the quality-assured pseudo-labels after the manual quality assurance procedure described in Section 2.2. These labels were subsequently used for model training and evaluation. After the sentiment labeling process was completed, the dataset was partitioned into training, validation, and testing sets using an 80:10:10 ratio. This partitioning produced 3,354 training samples, 419 validation samples, and 420 testing samples. The purpose of this data split was to ensure that model training, validation, and performance evaluation were conducted on mutually exclusive subsets, thereby providing a more objective assessment of the model's generalization performance.

3.2. Performance Validation of Monolingual Models

A preliminary evaluation was conducted to verify that both monolingual models were able to learn the sentiment patterns represented by the quality-assured pseudo-labeled datasets before proceeding to the cross-linguistic analysis. As presented in Table 5, Chinese-RoBERTa achieved an accuracy of 74% on the original Mandarin reviews, while IndoBERT achieved an accuracy of 69% on their Indonesian translations. These evaluation results reflect each model's agreement with the quality-assured pseudo-labels used in this study rather than classification accuracy against an independently human-annotated ground-truth dataset. Accordingly, the reported performance metrics should be interpreted as agreement with the quality-assured pseudo-labeled benchmark rather than as absolute classification accuracy against a fully human-verified evaluation set. Overall, both models demonstrated sufficient agreement with the quality-assured pseudo-labels to support the subsequent comparative analysis of sentiment probability distributions using Jensen–Shannon Divergence (JSD).

Table 5. Comparison of Monolingual Model Performance

Model	Accuracy	Macro Precision	Macro Recall	Macro F1-Score	Weighted F1-Score
Chinese-RoBERTa	0.74	0.71	0.71	0.71	0.74
IndoBERT	0.69	0.65	0.64	0.64	0.68

Table 6. Performance by Sentiment Category

Model	Class	Precision	Recall	F1-Score
Chinese-RoBERTa	Negative	0.76	0.74	0.75
	Neutral	0.52	0.57	0.54
	Positive	0.84	0.82	0.83
IndoBERT	Negative	0.65	0.63	0.64
	Neutral	0.52	0.43	0.47
	Positive	0.77	0.84	0.80

A class-wise evaluation was also conducted, as presented in Table 6. For both models, the positive class achieved the highest F1-score, reaching 0.83 for Chinese-RoBERTa and 0.80 for IndoBERT. In contrast, the neutral class obtained the lowest F1-score, with values of 0.54 and 0.47, respectively. Meanwhile, the negative class achieved F1-scores of 0.75 for Chinese-RoBERTa and 0.64 for IndoBERT.

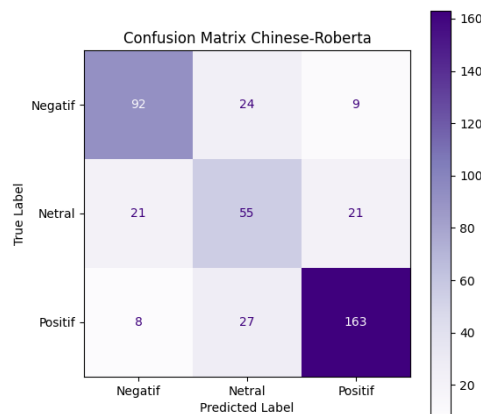


Figure 4. Confusion Matrix Chinese-RoBERTa

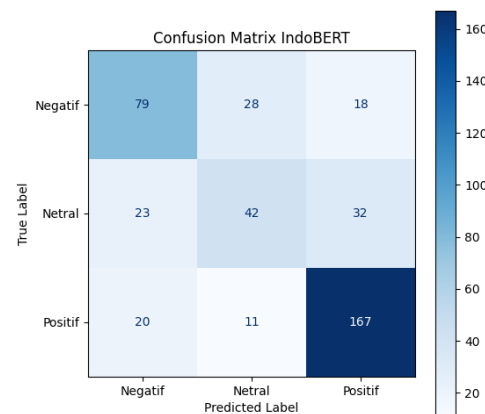


Figure 5. Confusion Matrix IndoBERT

Figures 4 and 5 illustrate the confusion matrices for both monolingual models. In both models, the neutral class exhibits the lowest classification performance, with many neutral reviews being misclassified as either the positive or negative class. For Chinese-RoBERTa, the misclassifications of neutral reviews are relatively balanced between the two sentiment classes. In contrast, IndoBERT predicts a larger proportion of neutral reviews as positive while correctly identifying fewer neutral instances. These observations indicate differences in the classification patterns produced by the two monolingual models, particularly for reviews with ambiguous sentiment. However, confusion matrices only describe changes in the predicted class and do not capture differences in the underlying probability distributions. Therefore, Jensen–Shannon Divergence (JSD) is employed in the subsequent analysis to quantify differences in the complete sentiment probability distributions generated by the two models.

3.3. Jensen-Shannon Divergence Analysis

Following the evaluation of the individual monolingual models, the subsequent analysis quantified differences between the sentiment probability distributions produced by the two independently fine-tuned models using Jensen–Shannon Divergence (JSD). Specifically, the probability distributions predicted by Chinese-RoBERTa for the original Mandarin reviews were compared with those predicted by IndoBERT for the corresponding Indonesian translations.

Table 7. Descriptive Statistics: Jensen-Shannon Divergence

Statistic	Value
Mean	0.2502
Median	0.2125
Min	0.0157
Max	0.7871
Std	0.1555

A statistical summary of the JSD values is presented in Table 7. The results show a mean JSD of 0.2502 and a median of 0.2125. The minimum JSD value of 0.0157 indicates that some review pairs exhibited nearly identical sentiment probability distributions, whereas the maximum value of 0.7871 reflects substantial differences between the probability distributions predicted by Chinese-RoBERTa and IndoBERT. Furthermore, the standard deviation of 0.1555 suggests considerable variability in JSD values across the analyzed review pairs. The distribution of JSD values for all review pairs is presented in Figure 6. Most review pairs exhibit relatively low JSD values, while only a small proportion show higher JSD values. This pattern indicates that the two

monolingual models produced relatively similar sentiment probability distributions for most aligned Mandarin–Indonesian review pairs, while larger differences in prediction probabilities were observed for only a limited subset of the data.

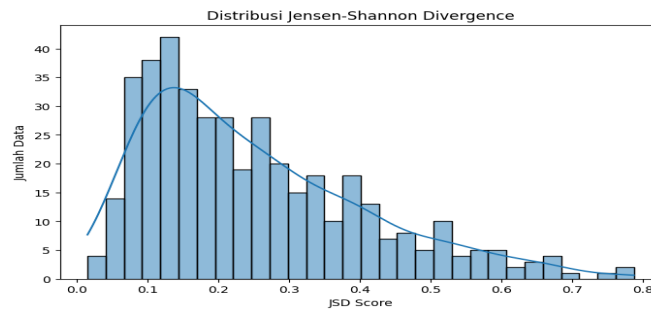


Figure 6. Jensen-Shannon Divergence Distribution

In addition to the probability distribution analysis, this study also examined differences in the sentiment categories predicted by the two monolingual models for each aligned Mandarin–Indonesian review pair. A comparison of the predicted sentiment categories is presented as a transition matrix in Figure 7. The results show that the majority of review pairs received the same sentiment prediction from both models. Nevertheless, a subset of review pairs was assigned different sentiment categories, indicating inconsistencies between the predictions produced by the two monolingual models.

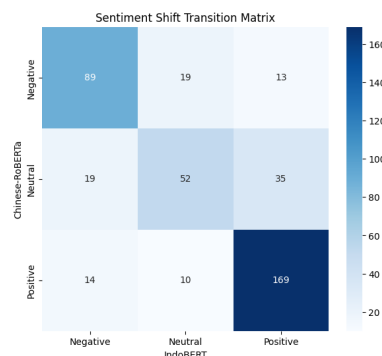


Figure 7. Sentiment Shift Transition Matrix

A more detailed summary of the prediction transition categories is presented in Table 8. A total of 310 aligned review pairs received the same sentiment prediction from both monolingual models, whereas 110 review pairs were assigned different sentiment categories. Among the prediction differences, Polarity Amplification was the most frequently observed category, accounting for 54 review pairs, followed by Polarity Dilution (29 review pairs) and Polarity Reversal (27 review pairs). These categories describe the direction of differences between the sentiment predictions produced by the two monolingual models for each aligned Mandarin–Indonesian review pair.

Table 8. Distribution of Sentiment Shifts.

Shift Category	Count	Percentage
Consistent	310	73.80%
Polarity Amplification	54	12.85%
Polarity Dilution	29	6.90%
Polarity Reversal	27	6.43%

As a complement to the quantitative analysis, this study also presents several review pairs with the highest Jensen–Shannon Divergence (JSD) values, as shown in Table 9. These examples were selected to illustrate representative cases of the prediction transition categories identified

throughout the analysis, including Polarity Amplification, Polarity Dilution, and Polarity Reversal. The JSD values of these review pairs range from 0.53 to 0.79, indicating substantial differences in the predicted sentiment probability distributions compared with the remaining review pairs.

Table 9. Representative High-JSD Review Pairs

Mandarin Text	Indonesian Translation	Mandarin Prediction	Indonesian Prediction	JSD	Category
第二季再见	sampai jumpa di musim kedua	Neutral	Positive	0.7871	Polarity Amplification
柳神是什么样的存在	liu shen itu makhluk seperti apa	Neutral	Negative	0.5717	Polarity Amplification
纯净版侧颜杀	versi murni dari profil samping yang mematikan	Positive	Neutral	0.7386	Polarity Dilution
仙逆剧组 超级大鸡腿	para pemain against the gods mendapatkan paha ayam yang super besar	Negative	Positive	0.6662	Polarity Reversal
紫月新建模还是太权威了 姬紫月	model baru ziyue masih terlalu berwibawa jiziyue	Positive	Negative	0.6515	Polarity Reversal
不良人听令 去看置子传	para penjahat dengarkan perintahku pergilah menonton chronicles of gangzi	Negative	Neutral	0.5338	Polarity Dilution
想紫月了	aku rindu ziyue	Neutral	Positive	0.5657	Polarity Amplification
柳神 月婵 欣赏一下吧	willowgoddess moonchan selamat menikmati	Neutral	Positive	0.7791	Polarity Amplification
完美三部曲	trilogi sempurna	Neutral	Positive	0.6883	Polarity Amplification
制作仙逆的那些大佬 我求你们能不能把原著给做出来 按原版剧情做出来 ...全文	kepada para pencipta xian ni saya mohon bisakah kalian mengadaptasi novel aslinya mengadaptasi cerita aslinya secara persis teks lengkap	Negative	Positive	0.6671	Polarity Reversal
神印王座六对CP组合	enam pasangan di the throne of seal	Positive	Neutral	0.6669	Polarity Dilution
各位我想问一下第二季13集对应小说多少章啊	halo semuanya saya ingin bertanya berapa bab novel yang sesuai dengan episode 13 musim 2	Neutral	Positive	0.5554	Polarity Amplification

3.4. Discussion

3.4.1. Discussion of Monolingual Model Performance

The results presented in Tables 5 and 6 indicate that the Chinese-RoBERTa model consistently outperformed IndoBERT across all evaluation metrics. Chinese-RoBERTa achieved an accuracy of 0.74, whereas IndoBERT attained an accuracy of 0.69. A similar performance gap was observed in Macro Precision, Macro Recall, Macro F1-score, and Weighted F1-score. These findings indicate that the two models exhibited different levels of agreement with the quality-assured pseudo-labels used as the evaluation reference. In the single experimental run conducted in this study, Chinese-RoBERTa achieved higher evaluation scores than IndoBERT across all reported metrics.

The observed performance difference is unlikely to be attributable to a single factor. In this study, several aspects may have contributed to the evaluation outcomes. First, the evaluation was conducted using pseudo-labels generated by the Chinese sentiment analysis model, making the labels inherently more aligned with the linguistic representation of the source language, although the quality of the pseudo-labels was partially verified through a manual quality assurance process. Second, Chinese-RoBERTa and IndoBERT are two monolingual language models pretrained on different corpora and employing distinct tokenization strategies and language representations, resulting in inherently different classification behaviors. Third, the data processed by IndoBERT consisted of machine-translated reviews, whose linguistic characteristics differ from those of the original Mandarin texts. Therefore, the observed performance gap should not be attributed solely to the machine translation process but rather interpreted as the result of the interaction among multiple factors within the design of this study.

The observed performance difference is consistent with previous cross-lingual sentiment analysis studies, which reported that independently pretrained monolingual transformer models often exhibit different classification performance because they learn language-specific lexical, syntactic, and semantic representations from different corpora. Consequently, differences in prediction performance should be interpreted as reflecting not only language translation but also the inherent characteristics of the underlying language models [6], [36].

It should be noted that the reported performance comparison is based on a single fine-tuning run for each model under fixed experimental settings. Consequently, statistical uncertainty associated with random initialization and training variability was not estimated. Furthermore, the observed five-percentage-point difference in accuracy between Chinese-RoBERTa and IndoBERT was not subjected to a formal statistical significance test. Therefore, the reported performance difference should be interpreted as a descriptive observation within the context of this study rather than as evidence of statistically significant superiority of one model over the other.

The superior performance on the positive class can likely be attributed to two factors. First, as shown in Table 4, the positive class contains the largest number of samples, providing the models with more training instances during the fine-tuning process. This larger sample size enables the models to learn more stable decision boundaries than for classes with fewer training examples. Second, positive sentiment in user reviews is typically expressed more explicitly through words and phrases conveying praise or satisfaction, making it easier for both models to identify such expressions consistently. In contrast, the neutral class exhibited the lowest performance, particularly for IndoBERT, which achieved a recall of only 0.43. This finding indicates that reviews with ambiguous sentiment remain the most challenging category to classify consistently. This finding is consistent with previous sentiment classification studies, which have reported that the neutral class generally represents the most challenging category because neutral expressions often contain weaker affective cues and greater semantic ambiguity than clearly positive or negative opinions [37].

These findings are also reflected in the confusion matrices presented in Figures 4 and 5. For Chinese-RoBERTa, most predictions are concentrated along the main diagonal, indicating generally correct classifications, although misclassifications remain relatively frequent for the neutral class. In contrast, the IndoBERT confusion matrix reveals that a larger proportion of neutral reviews are misclassified as positive, while fewer neutral reviews are correctly identified. This pattern suggests that the prediction discrepancies between the two models primarily arise from reviews containing ambiguous sentiment. Nevertheless, confusion matrices only illustrate changes in the final predicted labels and do not capture how the models' confidence levels vary across sentiment classes. Two reviews may receive the same predicted label even though the probability distributions produced by the two models differ substantially. Therefore, the Jensen–Shannon Divergence (JSD) analysis presented in the following section complements the label-based evaluation by comparing the complete probability distributions generated by the two monolingual models.

3.4.2. Analysis of Jensen–Shannon Divergence and Sentiment Shift

The JSD statistics presented in Table 7 show a mean value of 0.2502 and a median of 0.2125, indicating that the sentiment probability distributions generated by Chinese-RoBERTa and IndoBERT are generally similar across most Mandarin–Indonesian review pairs. The relatively low mean JSD indicates that, despite being processed by two independently trained monolingual models operating on different language versions, most aligned review pairs produced comparable probability distributions. This suggests that large prediction discrepancies are concentrated in only a limited subset of reviews rather than being uniformly distributed across the dataset. This interpretation is further supported by the JSD distribution shown in Figure 6, which exhibits a right-skewed pattern, where the majority of review pairs have low JSD values, while only a small proportion display high JSD values. These findings suggest that the two models generally produce comparable sentiment probability distributions for most translated review pairs. Nevertheless, a non-zero JSD value indicates the presence of differences in the predicted probability distributions,

suggesting that the confidence associated with the predictions produced by the two models is not necessarily identical, even when both models ultimately assign the same sentiment label. The difference in confidence levels may arise because the probability distributions being compared are generated by two independently developed monolingual models that process review pairs in different languages. Although both models may produce the same sentiment label, each constructs its probability distribution based on distinct linguistic representations, pre-training corpora, and fine-tuning parameters. Consequently, agreement in the predicted sentiment label does not necessarily imply similarity in the underlying probability distributions. Unlike conventional evaluation based solely on predicted labels, this probabilistic analysis reveals differences in prediction confidence that remain hidden even when both models assign the same sentiment category.

The transition matrix presented in Figure 7 shows that the majority of review pairs were assigned the same sentiment category by both monolingual models, as indicated by the dominance of values along the main diagonal of the transition matrix. Among the review pairs exhibiting prediction discrepancies, most transitions occurred between the neutral class and its adjacent sentiment classes, particularly from neutral to positive. In contrast, direct transitions between the positive and negative classes were relatively uncommon. This pattern suggests that disagreements between the two models primarily involved reviews expressing weaker or more ambiguous sentiment, rather than reviews with clearly defined positive or negative polarity. This observation is consistent with the confusion matrices presented in Figures 4 and 5, where most prediction errors also involve the neutral class rather than direct transitions between positive and negative sentiment. One possible explanation for this pattern is that each review pair is processed by two independently developed monolingual models that rely on different linguistic representations. When the probabilities assigned to two adjacent sentiment categories are relatively similar, even minor differences in the probability distributions produced by the two models may cause the highest-probability prediction to shift to a neighboring class. In contrast, a direct transition to the opposite sentiment polarity requires a substantially larger change in the underlying probability distribution, making such transitions comparatively less frequent. Therefore, the transition pattern observed in Figure 7 suggests that discrepancies between the two models are generally incremental rather than representing extreme changes in sentiment polarity. These findings also provide important context for interpreting the JSD results. Although most review pairs receive the same sentiment label or exhibit only a shift to an adjacent sentiment category, the underlying probability distributions may still differ. Accordingly, the transition matrix illustrates the direction of changes in the predicted sentiment categories, whereas JSD complements this analysis by quantifying the extent to which the probability distributions produced by the two models differ for each review pair.

The sentiment shift distribution presented in Table 8 shows that Polarity Amplification was the most prevalent category, followed by Polarity Dilution, whereas Polarity Reversal occurred least frequently. This distribution suggests that when prediction discrepancies occur, they are more likely to involve transitions between adjacent sentiment categories than direct shifts to the opposite sentiment polarity. The dominance of the Polarity Amplification category indicates that most observed prediction discrepancies do not correspond to a complete polarity change but rather reflect shifts in the degree of sentiment orientation. In contrast, the relatively low frequency of Polarity Reversal suggests that predictions with completely opposite sentiment polarities are uncommon among the analyzed review pairs. Overall, the distribution of prediction shift categories indicates that discrepancies between the two models generally remain within adjacent sentiment categories rather than involving extreme polarity changes. These findings provide additional insight into the JSD analysis. While the JSD values quantify the magnitude of the differences between the probability distributions produced by the two models, the distribution of prediction shift categories in Table 8 illustrates how these differences are manifested as changes in the predicted sentiment categories. Therefore, the two analyses complement each other by providing both probabilistic and categorical perspectives on the prediction agreement between Chinese-RoBERTa and IndoBERT for Mandarin reviews and their Indonesian translations.

Table 9 presents several representative review pairs with high JSD values, selected from the group of reviews with JSD scores greater than 0.5, as defined in the research methodology.

These examples cover multiple prediction shift categories, including Polarity Amplification, Polarity Dilution, and Polarity Reversal, thereby illustrating the different forms of prediction discrepancies that may arise when the two monolingual models produce substantially different probability distributions. Based on the examples presented in Table 9, review pairs with high JSD values are not confined to a particular type of expression. Some reviews consist of very short phrases, such as 第二季再见 and 想紫月了, whereas others are longer or contain character names and terms specific to the series, such as 柳神, 姬紫月, and 仙逆. This diversity suggests that differences in the probability distributions produced by the two models can arise across various forms of reviews, ranging from simple expressions to those containing context-dependent references.

Interestingly, several review pairs with high JSD values remain generally understandable after translation, yet they produce substantially different probability distributions and, in some cases, different predicted sentiment categories across the two models. This observation suggests that a high JSD value is not necessarily associated with an explicitly erroneous translation. However, since this study does not include a systematic evaluation of translation quality or a comprehensive linguistic analysis of all review pairs with high JSD values, the examples presented in Table 9 are not intended to identify the direct causes of the observed prediction discrepancies. Rather, they serve as qualitative illustrations that facilitate the interpretation of the JSD analysis by highlighting review pairs exhibiting the largest differences in probability distributions.

Although JSD values provide information about the degree of difference between the probability distributions produced by the two monolingual models, their interpretation should be approached with caution. In this study, the probability distributions were compared directly without applying probability calibration techniques, such as temperature scaling. Consequently, part of the observed variation in JSD values may reflect differences in the probability calibration characteristics of the individual models, in addition to linguistic differences between the Mandarin reviews and their Indonesian translations. Therefore, the JSD values reported in this study should be interpreted as a measure of the divergence between the probabilistic outputs of two independently trained monolingual models rather than as a direct measure of machine translation quality or the preservation of semantic meaning.

4. CONCLUSION

This study proposed a framework that combines two independently fine-tuned monolingual transformer models with Jensen–Shannon Divergence (JSD) to compare the sentiment probability distributions generated for aligned Mandarin–Indonesian review pairs. The experimental results obtained from the single experimental run conducted in this study showed that Chinese-RoBERTa achieved higher evaluation scores than IndoBERT on the quality-assured pseudo-labeled evaluation dataset, while the JSD analysis indicated that the sentiment probability distributions generated by the two models were generally similar for most aligned review pairs. Although most review pairs received identical sentiment predictions from both models, differences in prediction probabilities were still observed, and approximately one quarter of the review pairs exhibited different predicted sentiment categories. Among these prediction differences, Polarity Amplification occurred more frequently than Polarity Dilution and Polarity Reversal.

Overall, these findings of this study suggest that comparing probability distributions provides additional information beyond conventional label-based evaluation. While two monolingual models may assign the same sentiment label to an aligned review pair, their prediction confidence may differ substantially. Therefore, probability-distribution analysis can serve as a complementary evaluation approach for examining prediction differences between monolingual sentiment models operating on bilingual parallel texts. The interpretation of these differences, however, should consider that the reported JSD values reflect the outputs of two independently fine-tuned monolingual models and may also be influenced by factors such as model calibration, the use of pseudo-labeled training data, and the single-run experimental design adopted in this study.

Future research may extend this framework by incorporating probability calibration techniques before probability-distribution comparison, conducting repeated fine-tuning experiments with multiple random seeds to assess statistical uncertainty, evaluating additional machine translation systems and multilingual language models, and employing human-annotated benchmark datasets to further investigate the factors contributing to differences in sentiment probability distributions across languages.

ACKNOWLEDGEMENTS

The authors would like to thank the Department of Computer Science, Faculty of Mathematics and Natural Sciences, Lambung Mangkurat University, for the academic support and facilities provided during the completion of this research.

CONFLICT OF INTEREST

The authors declare that they have no conflicts of interest related to this study.

DATA AVAILABILITY

The data supporting the findings of this study are available from the corresponding author upon reasonable request.

AUTHOR CONTRIBUTIONS

Azfani Naurotul Jannah: Conceptualization; Data Curation; Investigation; Methodology; Software; Formal Analysis; Visualization; Writing-Original Draft. **Triando Hamonangan Saragih:** Conceptualization; Methodology; Supervision; Validation; Writing-Review & Editing. **Irwan Budiman:** Methodology; Supervision; Validation; Writing-Review & Editing. **Muliadi:** Validation; Writing-Review & Editing. **Radityo Adi Nugroho:** Validation; Writing-Review & Editing.

REFERENCES

- [1] Y. Mao, Q. Liu, and Y. Zhang, "Sentiment analysis methods, applications, and challenges: A systematic literature review," Apr. 01, 2024, *King Saud bin Abdulaziz University*. doi: 10.1016/j.jksuci.2024.102048.
- [2] K. L. Tan, C. P. Lee, and K. M. Lim, "A Survey of Sentiment Analysis: Approaches, Datasets, and Future Research," Apr. 01, 2023, *MDPI*. doi: 10.3390/app13074550.
- [3] Y. Xu, H. Cao, W. Du, and W. Wang, "A Survey of Cross-lingual Sentiment Analysis: Methodologies, Models and Evaluations," Sep. 01, 2022, *Springer*. doi: 10.1007/s41019-022-00187-3.
- [4] M. Krasitskii, G. Sidorov, O. Kolesnikova, L. Chanona-Hernandez, and A. Gelbukh, "Multilingual sentiment analysis of summarized texts: a cross-language study of text shortening effects," *PeerJ Comput. Sci.*, vol. 12, 2026, doi: 10.7717/peerj-cs.3406.
- [5] P. Naveen and P. Trojovský, "Overview and challenges of machine translation for contextually appropriate translations," *iScience*, vol. 27, no. 10, p. 110878, Oct. 2024, doi: 10.1016/j.isci.2024.110878.
- [6] P. Přibáň, J. Šmíd, J. Steinberger, and A. Mištera, "A comparative study of cross-lingual sentiment analysis," *Expert Syst. Appl.*, vol. 247, p. 123247, Aug. 2024, doi: 10.1016/j.eswa.2024.123247.
- [7] M. S. U. Miah, M. M. Kabir, T. Bin Sarwar, M. Safran, S. Alfarhood, and M. F. Mridha, "A multimodal approach to cross-lingual sentiment analysis with ensemble of transformer and LLM," *Sci. Rep.*, vol. 14, no. 1, p. 9603, Apr. 2024, doi: 10.1038/s41598-024-60210-7.
- [8] E. M. Mercha and H. Benbrahim, "Machine learning and deep learning for sentiment analysis across languages: A survey," *Neurocomputing*, vol. 531, pp. 195–216, Apr. 2023, doi: 10.1016/j.neucom.2023.02.015.
- [9] A. Velankar, H. Patil, and R. Joshi, "Mono vs Multilingual BERT for Hate Speech Detection and Text Classification: A Case Study in Marathi," Apr. 2022, doi: 10.1007/978-3-031-20650-4_10.

- [10] M. Karpinska, N. Raj, K. Thai, Y. Song, A. Gupta, and M. Iyyer, "DEMETER: Diagnosing Evaluation Metrics for Translation," Oct. 2022, [Online]. Available: <http://arxiv.org/abs/2210.13746>
- [11] C. Zhao *et al.*, "A Systematic Review of Cross-Lingual Sentiment Analysis: Tasks, Strategies, and Prospects," *ACM Comput. Surv.*, vol. 56, no. 7, Oct. 2024, doi: 10.1145/3645106.
- [12] M. Artetxe, V. Goswami, S. Bhosale, A. Fan, and L. Zettlemoyer, "Revisiting Machine Translation for Cross-lingual Classification." [Online]. Available: https://spacy.io/models/xx#xx_sent_ud_sm
- [13] S. Al Amer, M. Lee, and P. Smith, "Cross-lingual Classification of Crisis-related Tweets Using Machine Translation," in *International Conference Recent Advances in Natural Language Processing, RANLP*, Incoma Ltd, 2023, pp. 22–31. doi: 10.26615/978-954-452-092-2_003.
- [14] I. Shode, D. Ifeoluwa Adelani, J. Peng, and A. Feldman, "NollySenti: Leveraging Transfer Learning and Machine Translation for Nigerian Movie Sentiment Classification," *Short Papers*. [Online]. Available: <https://github.com/IyanuSh/NollySenti>
- [15] S. Qian, C. Or, F. do Carmo, and D. Kanojia, "Evaluating Machine Translation for Emotion-loaded User Generated Content (TransEval4Emo-UGC)." [Online]. Available: <https://weibo.com/>
- [16] Y. Li, T. Duan, and L. Zhu, "Public Attitudes and Sentiments toward Common Prosperity in China: A Text Mining Analysis Based on Social Media," *Applied Sciences (Switzerland)*, vol. 14, no. 10, May 2024, doi: 10.3390/app14104295.
- [17] K. Kuligowska and B. Kowalczyk, "Pseudo-labeling with transformers for improving Question Answering systems," in *Procedia Computer Science*, Elsevier B.V., 2021, pp. 1162–1169. doi: 10.1016/j.procs.2021.08.119.
- [18] J. Zhang *et al.*, "Fengshenbang 1.0: Being the Foundation of Chinese Cognitive Intelligence," Mar. 2023, [Online]. Available: <http://arxiv.org/abs/2209.02970>
- [19] Y. Cui, W. Che, T. Liu, B. Qin, S. Wang, and G. Hu, "Revisiting Pre-Trained Models for Chinese Natural Language Processing," Nov. 2020, doi: 10.18653/v1/2020.findings-emnlp.58.
- [20] Z. Tan *et al.*, "Neural machine translation: A review of methods, resources, and tools," Jan. 01, 2020, *Elsevier B.V.* doi: 10.1016/j.aiopen.2020.11.001.
- [21] W. Jooste, R. Haque, and A. Way, "Philipp Koehn: Neural Machine Translation," *Machine Translation*, vol. 35, no. 2, pp. 289–299, Jun. 2021, doi: 10.1007/s10590-021-09277-x.
- [22] A. S. Hanin and M. Maryam, "Sentiment Analysis of Twitter Towards the Free Lunch Program Using the C4.5 Algorithm," *International Journal of Advances in Data and Information Systems*, vol. 6, no. 1, pp. 31–45, Apr. 2025, doi: 10.59395/ijadis.v6i1.1357.
- [23] V. R. Joseph, "Optimal ratio for data splitting," *Stat. Anal. Data Min.*, vol. 15, no. 4, pp. 531–538, Aug. 2022, doi: 10.1002/sam.11583.
- [24] J. H. Syu, M. Fojcik, R. Cupek, and J. C. W. Lin, "HTTPS: Heterogeneous Transfer learning for split Prediction System evaluated on healthcare data," *Information Fusion*, vol. 113, p. 102617, Jan. 2025, doi: 10.1016/J.INFFUS.2024.102617.
- [25] Q. H. Nguyen *et al.*, "Influence of data splitting on performance of machine learning models in prediction of shear strength of soil," *Math. Probl. Eng.*, vol. 2021, 2021, doi: 10.1155/2021/4832864.
- [26] Y. Liu *et al.*, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," Jul. 2019, [Online]. Available: <http://arxiv.org/abs/1907.11692>
- [27] F. Qiu, H. Xiao, K. Yu, and Q. Qiu, "MP-CNER: A Course Named Entity Recognition Model Based on Multi-Dimensional Position Features," *Electronics (Switzerland)*, vol. 15, no. 11, Jun. 2026, doi: 10.3390/electronics15112396.
- [28] I. Loshchilov and F. Hutter, "DECOUPLED WEIGHT DECAY REGULARIZATION." [Online]. Available: <https://github.com/loshchil/AdamW-and-SGDW>

- [29] R. Anadra, H. Wijayanto, and K. Sadik, "Sentiment Analysis of Tokopedia Customer Reviews Using BiLSTM and IndoBERT with Comparative Analysis of Preprocessing and Labeling Methods," *International Journal of Advances in Data and Information Systems*, vol. 6, no. 3, pp. 773–788, Dec. 2025, doi: 10.59395/ijadis.v6i3.1458.
- [30] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," Online. [Online]. Available: <https://huggingface.co/>
- [31] D. Jurafsky and J. H. Martin, "Speech and Language Processing An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models Third Edition draft."
- [32] M. Owusu-Adjei, J. Ben Hayfron-Acquah, T. Frimpong, and G. Abdul-Salaam, "A systematic review of prediction accuracy as an evaluation measure for determining machine learning model performance in healthcare systems," Jun. 04, 2023. doi: 10.1101/2023.06.01.23290837.
- [33] D. M. W. Powers and Ailab, "EVALUATION: FROM PRECISION, RECALL AND F-MEASURE TO ROC, INFORMEDNESS, MARKEDNESS & CORRELATION."
- [34] D. J. Hand, P. Christen, and N. Kirielle, "F*: an interpretable transformation of the F-measure," *Mach. Learn.*, vol. 110, no. 3, pp. 451–456, Mar. 2021, doi: 10.1007/s10994-021-05964-1.
- [35] B. Shade and E. G. Altmann, "Quantifying the Dissimilarity of Texts," *Information (Switzerland)*, vol. 14, no. 5, May 2023, doi: 10.3390/info14050271.
- [36] X. Zhu, S. Gardiner, T. Roldán, and D. Rossouw, "The Model Arena for Cross-lingual Sentiment Analysis: A Comparative Study in the Era of Large Language Models," 2024. [Online]. Available: <https://platform.openai.com/docs/models/>
- [37] J.-Q. Wang, "CLAS-Net: A study on cross-lingual intelligent sentiment analysis model fusing semantic alignment," *PLoS One*, vol. 21, no. 2, p. e0342342, Feb. 2026, doi: 10.1371/journal.pone.0342342.