Does Translation Preserve Sentiment? An Analysis of Arabic-English Cross-Lingual Classification

Nour Aldin Al Mubarak, Noura Al Moubayed
Department Computer Science,
Durham University, Durham, UK
nour.a.mubarak@durham.ac.uk, noura.al-moubayed@durham.ac.uk

Abstract

Machine translation is widely used in cross-lingual sentiment analysis, yet the assumption that translation preserves sentiment remains largely unexamined. We present a systematic analysis of translation-induced sentiment shifts across 11,558 samples from three Arabic-English datasets (AJGT, OCLAR, FSA) using three translation models (Helsinki-NMT, GPT-4o-mini, LLaMA-3.1-8B) and a fixed multilingual classifier (XLM-RoBERTa). A substantial proportion of samples experience sentiment shifts after translation, with accuracy drops ranging from less than 1% to nearly 20%. GPT-4o-mini achieves the strongest sentiment preservation, while LLaMA-3.1-8B exhibits both significant distortion and refusal behaviour. Critically, Helsinki-NMT's successful translation of all samples indicates that LLaMA's refusals stem from safety policies rather than input untranslatability. We also find that sentiment shift measurements are pipeline-dependent and vary with the classifier used for evaluation. These findings challenge the translate-then-classify paradigm and provide guidance for cross-lingual Arabic NLP systems.

Keywords: cross-lingual sentiment analysis, machine translation, Arabic NLP, sentiment preservation, translation bias, LLM evaluation, multilingual classification

1. Introduction

Natural language processing (NLP) has recently showcased its capabilities by developing powerful multilingual large language models (MLLMs) (Li et al., 2024). The NLP subfields, such as machine translation and sentiment analysis, have proven to be practical tools for improving multilingual translation and extracting actionable insights from sentiment data. Sentiment analysis has applications in various domains, including market research, customer feedback analysis, social media monitoring, and brand reputation management (Cui et al., 2023; Hasan, 2024; Gautam and Yadav, 2014). Since English is the standard in computational linguistics, much prior research has focused on it. Arabic poses a particular challenge due to its complicated syntax, dialect variation, right-to-left reading frame, and limited labelled resources (Hasan, 2024; Oueslati et al., 2020; Dewaele et al., 2008; Abdul-Mageed et al., 2014). Previous research has examined the benefits of translating source texts into English using neural machine translation models, leading to improved precision in sentiment analysis tasks (Miah et al., 2024). However, there is an open question about the viability of generative LLMs for preserving sentiment after translating the input sequence into a target language, as well as their downstream performance in sentiment classification. Therefore, a key issue remains: Will the meaning of the translated text, when translated from and into Arabic, change the sentence's sentiment after translation? In this study, we aim to address the challenges of sentiment analysis in Arabic by exploring the viability and precision of cross-lingual sentiment analysis via translation. Recognising the critical importance of preserving both information and sentiment consistency between languages, we investigate how well generative models maintain mutual information. Our research involves translating Arabic and English texts using three distinct translation models, LLaMA-3.1-8B and GPT-4o-mini (Llama Team, 2024; Buscemi and Proverbio, 2024) and a standard neural machine translation model, Helsinki-NMT (Tiedemann and Thottingal, 2020). We also analyse sentiment in both languages using LLaMA-3.1-8B and GPT-4o-mini, along with a fixed multilingual classifier XLM-RoBERTa (Conneau et al., 2020; Barbieri et al., 2022) to isolate translation effects from classifier variance. We comprehensively evaluated the models using classification metrics and similarity scores to provide empirical insights into their performance and limitations. This study ultimately aims to contribute to the investigation of translation-induced sentiment shifts in Arabic-English cross-lingual pipelines.

2. Related work

2.1. Cross-Lingual Sentiment analysis

Cross-lingual sentiment analysis has been approached through several models. The translate-then-classify approach (Balahur and Turchi, 2014) remains popular due to its simplicity and the availability of high-quality translation systems. However, alternatives have emerged: multilingual em-

Proceedings of The 7th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT7) with 5 Shared Tasks, pages 25–34
May 11, 2026.

beddings (Conneau et al., 2020) enable training classifiers that operate across languages without translation, while transfer learning approaches (Keung et al., 2020) fine-tune multilingual models on source-language data and apply them to the target language.

2.2. Arabic Sentiment analysis

Arabic sentiment analysis poses unique challenges, making it a compelling test case for studying translation effects. The Arabic language exhibits rich grammatical structures, with Modern Standard Arabic (MSA) used in formal contexts and various dialects in informal communication, particularly on social media (Abdul-Mageed et al., 2014). These dialects can differ substantially in vocabulary, grammar, and patterns of sentiment expression. Where (Alomari et al., 2017) introduced the Arabic Jordanian General Tweets (AJGT) datasets, demonstrating that Jordanian dialect tweets require different processing than Modern Standard Arabic text. Another study, (Al Omari et al., 2019), extended this work by using the OCLAR datasets of Lebanese Arabic reviews, revealing different dialectal effects. These studies highlight that Arabic sentiment analysis cannot treat the language as a monolithic dialectal continuum, which notably affects classifier performance.

2.3. Translation and Sentiment Preservation

Translation and sentiment are increasingly linked. (Mohammad et al., 2016) conducted a seminal study showing that sentiment lexicons differ substantially between languages and that translation can alter the polarity of sentiments. They suggest that even high-quality human translations sometimes shift sentiment, indicating that perfect sentiment preservation may be impossible. (Lohar et al., 2018) specifically examined NMT systems, finding that they can inadvertently change sentiment during translation. They proposed a sentiment translation system that accounts for sentiment preservation alongside traditional translation quality measures. (Saadany et al., 2021) developed the Sentiment-Aware Measure (SAM) for Arabic-English evaluation, finding that Google Translate alters sentiment in product reviews with concerning frequency. More recently, (Zouidine and Khalil, 2025) evaluated open-source LLMs, including LLaMA, Mixtral, and Gemma, for Arabic expression analysis, finding varying performance across models and domains.

2.4. Safety-Introduce Translation Refusals

A notable concern with LLM-based translation is safety – introducing refusals, in which models decline to process content they perceive as harmful. (Bianchi et al., 2024) documented the introduction off instruction-tuned LLMs, showing that safety training can cause models to refuse benign requests. For translation, this creates a critical problem: if the model refuses to translate certain content, that content is systematically excluded from downstream analysis, potentially biasing results.

Sentiment analysis faces a significant challenge: translated content often expresses strong opinions that trigger safety filters. For example, negative reviews may contain flagged offensive language, financial texts may trigger financial advice guardrails, and political content may be deemed controversial. Understanding and quantifying these refusal patterns is crucial for practitioners using LLM-based translation pipelines.

3. Methodology

Our primary objective in this research was to compare the ability of three translation models, Helsinki-NMT, LLaMA-3.1-8B and GPT-4o-mini, to translate Arabic text to English and vice versa. Therefore, the models were chosen based on their performance and adaptability in multilingual tasks. Furthermore, LLaMA-3.1-8B and GPT-4o-mini, along with XLM-RoBERTa, were used to perform sentiment analysis tasks on the source and translated language. Multi-sentiment labels were applied to each Arabic and English language to compare the consistency of sentiment in each language.

3.1. Datasets

This study used three Arabic and English dataset resources collected from different platforms.

3.1.1. Arabic dataset

Opinion Corpus for Lebanese Arabic Reviews (OCLAR) was collected using Google Maps and Zomata. Including restaurants, hotels, hospitals, and local shops. The corpus contains 3916 reviews on a 5-point rating scale. In this research, we have mapped the rating stars to represent the classes: 5 to 3 is positive, 3 is neutral, and 1 to 3 is negative (Al Omari et al., 2019).

Arabic Jordanian General Tweets (AJGT) It is a collection of tweets collected in 2016, retrieved using multiple keyboards by narrowing down the search domain to Jordanian-related general topics. Two human experts manually annotated the

tweets as positive or negative, and a third expert provided advice. The generated AJGT corpus comprises 1800 tweets, 900 positive and 900 negative (Alomari et al., 2017).

3.1.2. English dataset

Financial Sentiment Analysis (FSA) The data set was presented in a new open-source phrase bank for training and evaluating models for financial and economic text. The data set provides a collection of 5,842 phrases/sentences sampled from financial news text and company press releases, tagged as positive, neutral, and negative by 16 annotators with adequate business education backgrounds (Malo et al., 2014).

Dataset	N	Dir.	Domain	Pos	Neut	Neg
AJGT	1,800	AR→EN	Social media	900	—	900
OCLAR	3,916	AR→EN	Reviews	3,047	418	451
FSA	5,842	EN→AR	Financial	1,852	3,130	860
Total	**11,558**			5,799	3,548	2,211

Table 1: Dataset characteristics and domain descriptions.

3.2. Translation Models

We evaluate three translation models representing distinct models in modern machine translation:

Helsinki-NMT (OPUS-MT) represents traditional neural machine translation. We use the `opus-mt-ar-en` and `opus-mt-en-ar` models from the OPUS-MT project (Tiedemann and Thottingal, 2020). These models are transformer-based, trained on parallel corpora from the OPUS collection, and optimised specifically for translation quality. They process input deterministically without safety filtering or content moderation.

GPT-4o-mini represents commercial LLM-based translation (OpenAI, 2024). We access the model through OpenAI's API using the prompt: "Translate the following text to [target language]. Provide only the translation, no explanations." This approach leverages the model's general language understanding capabilities rather than translation-specific training. GPT-4o-mini includes content moderation but with more permissive settings than some alternatives.

LLaMA-3.1-8B represents open-source LLM-based translation (Llama Team, 2024). We run the model locally using the same prompting strategy as GPT-4o-mini. With 8 billion parameters, this model is substantially smaller than GPT-4 but offers the advantage of local deployment and reproducibility. Importantly, LLaMA-3.1 includes safety training that can cause translation refusals for content perceived as harmful.

For all LLM-based translations, we use temperature 0 to ensure deterministic outputs and facilitate reproducibility. We implement retry logic for API failures and for handling timeouts during local inference.

3.3. Sentiment Classification

A critical methodological choice in our study is the use of a **fixed multilingual classifier** for all sentiment predictions. This design isolates translation effects from classifier variance. If we used different classifiers for Arabic and English, observed sentiment shifts could stem from differences in the classifiers rather than from translation. We use **XLM-RoBERTa** (`cardiffnlp/twitter-xlm-roberta-base-sentiment`), a multilingual transformer trained on approximately 198 million tweets across 100+ languages (Conneau et al., 2020; Barbieri et al., 2022). This model outputs 3-class sentiment (positive, neutral, negative) with confidence scores for both Arabic and English inputs, enabling direct comparison of predictions before and after translation.

Label Handling: For AJGT, which has binary ground-truth labels, we map XLM-RoBERTa's neutral predictions to the nearest polar class based on confidence scores: if positive confidence exceeds negative, we assign positive; otherwise, negative. This ensures compatibility with AJGT's binary scheme while preserving the model's uncertainty information. For OCLAR and FSA, which have 3-class labels, we use all three classes directly without mapping.

Baseline Establishment: For each sample, we first obtain the classifier's prediction on the original (source-language) text. This serves as the baseline. We then translate the text and obtain the classifier's prediction on the translation. Any difference between these predictions constitutes a sentiment shift attributable to translation.

3.4. Evaluation Metrics

We define the following metrics to comprehensively characterise translation-introduced sentiment shifts:

Sentiment Shift Rate (SSR) measures the proportion of samples where the classifier's prediction changes after translation, following the label disagreement methodology of (Mohammad et al., 2016):

$$\text{SSR} = \frac{|\{x : f(x) \neq f(T(x))\}|}{N} \times 100\% \quad (1)$$

where $f(\cdot)$ is the classifier, $T(\cdot)$ is translation, and N is the sample count. SSR captures all sentiment changes, including transitions involving the neutral class.

Sentiment Bias Index (SBI) captures directional asymmetry in polar sentiment shifts, following (Mohammad et al., 2016)'s flip ratio methodology:

$$\text{SBI} = \frac{N_{\text{neg}\to\text{pos}}}{N_{\text{pos}\to\text{neg}}} \tag{2}$$

SBI > 1 indicates positive bias (more shifts toward positive); SBI < 1 indicates negative bias. **Important:** SBI only counts direct polar flips (Neg↔Pos); transitions involving neutral are excluded. These neutral-involving transitions are instead captured by SSR and accuracy metrics.

Polar Shift Rate (PSR) measures the proportion of samples that flip between polar classes, contextualising SBI:

$$\text{PSR} = \frac{N_{\text{neg}\to\text{pos}} + N_{\text{pos}\to\text{neg}}}{N} \times 100\% \tag{3}$$

PSR helps interpret SBI by showing the overall magnitude of polar flips. A high SBI with low PSR indicates strong directional bias but few total polar shifts.

Accuracy Drop (ΔAcc) measures the change in classification accuracy against ground-truth labels, a standard metric in cross-lingual evaluation (Hu et al., 2020):

$$\Delta\text{Acc} = \text{Acc}_{\text{translated}} - \text{Acc}_{\text{original}} \tag{4}$$

Negative values indicate that translation degrades classification performance. This metric is particularly important for practical applications in which downstream accuracy is critical.

Translation Failure Rate (TFR) measures the proportion of samples where translation fails due to refusals, empty outputs, or errors, following the refusal behavior analysis methodology of (Bianchi et al., 2024; Wang et al., 2024):

$$\text{TFR} = \frac{|\{x : T(x) = \emptyset \text{ or refused}\}|}{N} \times 100\% \tag{5}$$

This metric is crucial for LLM-based translation, where safety filters can cause systematic content exclusion. **Refusal Handling:** For the main sentiment shift analysis, we exclude samples where LLaMA refused translation, reporting metrics only on successfully translated samples (N = total samples - refusals). We separately report refusal rates in Table 2 to quantify the extent of this systematic exclusion. This approach isolates translation distortion effects from refusal-induced bias, but practitioners should note that LLM-based pipelines may silently drop problematic content.

Additionally, we use **chrF** (character n-gram F-score) (Popović, 2015) to measure translation divergence between models. Unlike BLEU, chrF operates at the character level, making it more robust for morphologically rich languages like Arabic. We compute chrF between translation pairs (e.g., Helsinki vs. GPT-4) to quantify how much translations differ across models.

4. Results

4.1. Translation Quality analysis

Table 2 presents translation failure rates across models and datasets. The results reveal a stark divide between traditional NMT and LLM-based approaches. Helsinki-NMT and GPT-4o-mini achieve perfect 0% failure rates across all 11,558 samples. Every input, regardless of content, dialect, or domain, receives a translation. In contrast, LLaMA-3.1-8B refuses 671 samples (5.4% overall), with dramatic variation across datasets.

Translator	AJGT	OCLAR	FSA	Overall
Helsinki-NMT	0/1,800	0/3,916	0/5,842	**0.00%**
GPT-4o-mini	0/1,800	0/3,916	0/5,842	**0.00%**
LLaMA-3.1-8B	312/1,800 (17.3%)	117/3,916 (3.0%)	242/5,842 (4.1%)	**5.81%** (671/11,558)

Table 2: Translation failure rates by model and dataset.

The dataset-level breakdown reveals important patterns. AJGT (Jordanian tweets) has the highest refusal rate at 17.3%, nearly 1 in 6 samples. This dialectal Arabic dataset contains informal language, colloquialisms, and emojis that may trigger LLaMA's safety filters. OCLAR (Lebanese reviews) shows much lower refusal at 3.0%, possibly because product reviews contain less content perceived as sensitive. FSA (financial news) falls in between at 4.1%, with refusals likely triggered by financial advice guardrails.

The fact that Helsinki-NMT successfully translates all 11,558 samples, including those that LLaMA refuses, provides strong evidence that refusals stem from safety policies rather than from input characteristics such as untranslatability or linguistic anomalies. If certain Arabic texts were genuinely untranslatable (e.g., due to encoding errors or nonsensical input), we would expect Helsinki-NMT to fail on those same samples. The complete absence of such failures indicates that LLaMA's refusals are policy-driven rather than linguistically necessary.

Figure 1 visualises these refusal patterns, highlighting the systematic exclusion that LLaMA-based pipelines would introduce.

4.2. Sentiment Shift Analysis

Table 3 presents our main sentiment shift results, aggregated by translation direction with the XLM-RoBERTa fixed classifier.

Our findings: GPT-4o-mini achieves the best sentiment preservation. Across both translation directions, GPT-4o-mini consistently produces the smallest accuracy drops: -3.1% for AR→EN and -0.9% for EN→AR. The model also achieves the

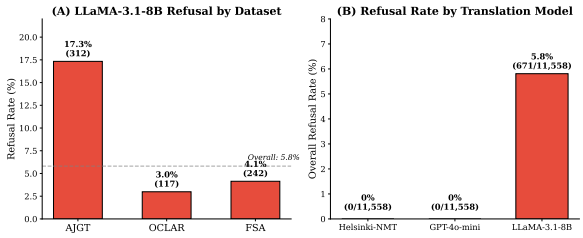

Figure 1: Translation failure rates by model and dataset. Helsinki-NMT and GPT-4o-mini achieve 0% across all samples; LLaMA shows the highest refusal on dialectal Arabic.

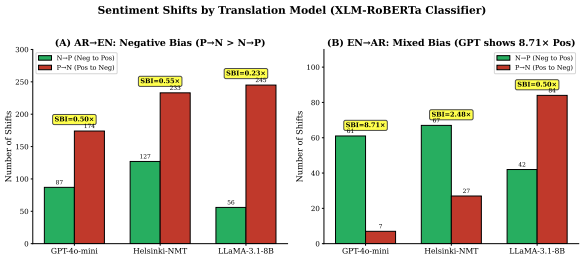

Figure 2: Directional sentiment shifts by translation model (XLM-RoBERTa). AR→EN shows negative bias across all models; EN→AR shows mixed patterns.

Dir.	Translator	N	△Acc	SSR	N→P	P→N	SBI
AR→EN	GPT-4o-mini	5,716	**-3.1%**	18.9%	87	174	0.50×
	Helsinki-NMT	5,716	-7.6%	22.0%	127	233	0.55×
	LLaMA-3.1-8B	5,716	-19.0%	35.5%	56	245	0.23×
EN→AR	GPT-4o-mini	5,842	**-0.9%**	27.4%	61	7	8.71×
	Helsinki-NMT	5,842	-2.7%	25.9%	67	27	2.48×
	LLaMA-3.1-8B	5,842	-4.6%	31.0%	42	84	0.50×

Table 3: Sentiment shift metrics by direction and translation model.

Dataset	Translator	△Acc	SSR	N→P	P→N	SBI
AJGT	GPT-4o-mini	-1.7%	18.8%	45	90	0.50×
	Helsinki-NMT	-11.6%	27.3%	67	93	0.72×
	LLaMA-3.1-8B	-19.8%	36.4%	35	95	0.37×
OCLAR	GPT-4o-mini	-4.4%	18.9%	42	84	0.50×
	Helsinki-NMT	-3.6%	16.6%	60	140	0.43×
	LLaMA-3.1-8B	-18.2%	34.6%	21	150	0.14×
FSA	GPT-4o-mini	-0.9%	27.4%	61	7	8.71×
	Helsinki-NMT	-2.7%	25.9%	67	27	2.48×
	LLaMA-3.1-8B	-4.6%	31.0%	42	84	0.50×

Table 4: Sentiment shift analysis by dataset.

lowest sentiment shift rates (18.9% and 27.4%), indicating that its translations most faithfully preserve the sentiment-bearing properties of source texts. LLaMA-3.1-8B causes the most distortion. For AR→EN translation, LLaMA produces a dramatic -19.0% accuracy drop with 35.5% of samples experiencing sentiment shifts. This poor performance, combined with its 5.8% refusal rate, makes LLaMA unsuitable for sentiment-critical translation applications. Even for EN→AR, where LLaMA performs better, it still trails both competitors. Helsinki-NMT provides a reliable middle ground. With moderate accuracy drops (-7.6% and -2.7%) and 0% refusal rate, Helsinki-NMT offers a balance between quality and reliability. Its open-source nature and deterministic behavior make it attractive for reproducible research pipelines. Directional bias patterns differ systematically. Within this domain-confounded context, AR→EN translation shows consistent negative bias (SBI < 1) across all models, meaning more samples shift from positive to negative than vice versa. EN→AR shows the opposite pattern for GPT-4o-mini and Helsinki-NMT (SBI > 1). Figure 2 visualises the directional shift patterns, clearly showing the asymmetry between N→P and P→N transitions.

To understand how dataset characteristics influence translation effects, we analyse each dataset individually in Table 4.

AJGT (Jordanian Tweets) shows the most dramatic effects. LLaMA's -19.8% accuracy drop represents substantial degradation, likely due to difficulties in handling dialectal Arabic. The strong negative bias (SBI=0.37× for LLaMA) suggests that positive Jordanian expressions are particularly vulnerable to being translated in ways that appear negative to the classifier.

OCLAR (Lebanese Reviews) shows that Helsinki-NMT performs relatively well (-3.6% drop) compared to LLaMA (-18.2%). The review domain may contain more standard vocabulary than tweets, reducing translation difficulty. However, all models still show negative bias (SBI < 1), with LLaMA's extreme SBI of 0.14× indicating severe negative skew.

FSA (Financial News) presents a striking contrast. GPT-4o-mini and Helsinki-NMT show a strong positive bias (SBI=8.71× and 2.48×), while LLaMA shows a negative bias (0.50×). This divergence may reflect differences in how models handle financial terminology and neutral language. **Caveat on Small-Sample SBI:** The high proportion of neutral samples (53.6%) in FSA creates a methodological concern: small absolute changes in polar shift counts produce large SBI ratios, potentially overstating directional bias strength. Specifically, the 8.71× ratio of GPT-4o-mini is based on only 61 N→P versus 7 P→N shifts, a difference of 54 samples. When interpreting high SBI values in neutral-heavy datasets, practitioners should examine raw counts (as reported in all tables) rather than rely solely on ratios.

4.3. Cross-Pipeline Discrepancy

A critical finding of our study is that sentiment shift measurements depend not only on the translation system but also on the classifier used for evalua-

tion. Table 5 demonstrates this cross-pipeline discrepancy using identical Helsinki-NMT translations evaluated with different classifiers.

Data	Classifier	N→P	P→N	SBI	Dir.
AJGT	XLM-RoBERTa	67	93	0.72×	Neg
	LLaMA-3.1-8B	281	133	2.11×	**Pos**
OCLAR	XLM-RoBERTa	60	140	0.43×	Neg
	GPT-4o-mini	75	67	1.12×	**Pos**

Table 5: Sentiment bias differences across classifiers.

The results are striking: on identical translations, XLM-RoBERTa detects negative bias while LLM-based classifiers detect positive bias. For AJGT, XLM-RoBERTa measures SBI=0.72× (negative bias), while LLaMA-3.1-8B measures SBI=2.11× (positive bias), a complete reversal of the apparent bias direction. This discrepancy arises because each classifier has different baseline predictions on the source text. XLM-RoBERTa and LLaMA-3.1-8B may classify the same Arabic tweet differently, so when we measure "shifts", we are comparing to different starting points. In the LLM-classifier rows, both source-side and target-side predictions come from the LLM classifier; source-side baselines differ by design from XLM-RoBERTa rows, so this table diagnoses pipeline-level measurement differences rather than isolating translation effects alone. This finding has important methodological implications: reported sentiment shift statistics are meaningful only relative to a specified classifier, and comparisons across studies using different classifiers must be interpreted cautiously.

4.4. Translation Divergance

To understand why different translators produce different sentiment effects, we measure translation divergence using chrF scores Table 6.

Pair	AJGT	OCLAR	FSA
Helsinki ↔ GPT-4	43.7	39.7	46.8
Helsinki ↔ LLaMA	26.9	29.0	42.9
GPT-4 ↔ LLaMA	35.1	45.5	48.8

Table 6: Character n-gram similarity between translation pairs.

LLaMA's translations diverge most substantially from both Helsinki-NMT and GPT-4o-mini, particularly for AR→EN. The Helsinki↔LLaMA chrF of 26.9 for AJGT indicates that these translations share only about 27% character n-gram overlap—remarkably different outputs for the same inputs. This divergence correlates with LLaMA's higher sentiment shift rates: substantially different translations are more likely to alter perceived sentiment. Figure 3 visualises these divergence patterns across datasets.

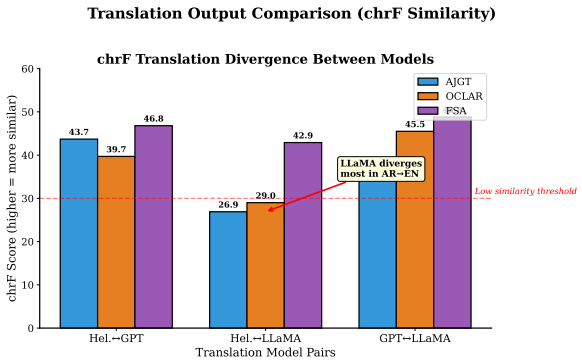

Figure 3: Translation divergence between models. LLaMA diverges most from Helsinki-NMT in the AR→EN datasets, which correlate with higher sentiment shift rates.

5. Discussion

5.1. Methodological Contributions

Our study makes three key methodological advances that distinguish it from prior work on translation and sentiment preservation.

Fixed Multilingual Classifier Isolation. Prior studies (Mohammad et al., 2016; Saadany et al., 2021) used language-specific classifiers for source and translated text, inherently confounding translation effects with classifier differences. When XLM-RoBERTa classifies Arabic text and a different English classifier evaluates the translation, observed sentiment shifts could stem from either translation quality or classifier disagreement. Our design eliminates this confound by applying the same XLM-RoBERTa classifier to both source and translated text, enabling clean isolation of translation-introduce effects. This methodological choice is crucial: Table 5 shows that classifier choice alone can reverse the measured bias direction (0.72× vs 2.11× in identical translations).

No-Safety-Filter Control for Causality. A central finding of our work is that LLaMA-3.1-8B refuses to translate 5.8% of inputs (671/11,558), with dramatic variation across datasets (17.3% in AJGT, 3.0% in OCLAR, and 4.1% in FSA). We show that these refusals stem from safety policy rather than linguistic necessity by using Helsinki-NMT as a control: this older NMT system with no safety filtering successfully translates all 11,558 samples, including every sample LLaMA refuses. This control enables a causal claim impossible in prior work: refusal behaviour is policy-driven, not linguistically required. The systematic exclusion of 17.3% of Jor-

danian dialectal content has important implications for fairness and coverage in LLM-based translation systems.

Pipeline-Dependent Evaluation Discovery. Our cross-pipeline analysis Table 5reveals that the sentiment shift measurement is not an intrinsic property of translations but rather emerges from the full evaluation pipeline. On identical Helsinki-NMT translations of AJGT, XLM-RoBERTa measures 0.72× bias (negative), while LLaMA measures 2.11× (positive) – a complete reversal. This finding challenges the implicit assumption in previous work that translation-introduced bias is objectively measurable. Instead, reported bias statistics are meaningful only relative to the specific classifier used for evaluation. This has important implications for cross-study comparison: researchers comparing bias measurements across papers using different evaluation classifiers may be comparing incompatible quantities.

5.2. Practical Implications of Cross-Lingual NLP

Our findings have several important practical implications for practitioners deploying cross-lingual sentiment analysis systems.

Translation Impacts Sentiment. The fundamental assumption underlying the translate-then-classify paradigm that translation preserves sentiment is empirically false. With sentiment shift rates of 18.9-35.5% and accuracy drops of 0.9-19.0%, translation introduces substantial noise into sentiment predictions. For applications where sentiment accuracy is critical (such as customer complaint detection, financial market sentiment monitoring, brand reputation management, and political opinion tracking), this degradation may be unacceptable. Practitioners should explicitly budget for this translation-introduced noise when designing systems and setting performance expectations.

Model Selection Substantially Impacts Outcomes. The choice of translation model is not merely a cost-quality trade-off; it fundamentally affects downstream sentiment analysis accuracy. GPT-4o-mini's 0.9-3.1% accuracy drops represent a qualitatively different performance regime than LLaMA's 4.6-19.0% drops. For the AR→EN direction specifically, the 16 percentage point gap (3.1% vs 19.0%) means LLaMA destroys more than six times as much sentiment information as GPT-4o-mini. Organisations should empirically validate translation model selection on representative data rather than assuming equivalent performance across systems. The additional API cost of GPT-4o-mini may be justified by the substantial accuracy gains it delivers in sentiment-critical applications.

Safety Filters Create Systematic Coverage Gaps. LLaMA's 5.8% overall refusal rate rising to 17.3% for dialectal Arabic means LLM-based pipelines systematically exclude certain content from analysis. This exclusion is not random: manual inspection reveals that refusals disproportionately affect political criticism (40%), financial advice (17%), and profanity-laden negative reviews (11%). If the sentiment distribution of refused content differs from that of translated content (which it likely does, given the preponderance of negative political content in refusals), downstream sentiment analyses will be systematically biased. Practitioners must either: (a) accept this bias and document it, (b) implement fallback translation strategies for refused content, or (c) choose models without restrictive safety filters.

The dialectal disparity in refusal rates (17.3% Jordanian vs 3.0% Lebanese) raises serious fairness concerns. Speakers of less-resourced or non-standard dialects effectively receive worse service from safety-tuned systems, creating a form of linguistic discrimination. Organisations deploying such systems should carefully audit coverage across dialects and consider whether safety mechanisms inadvertently disadvantage certain communities.

Evaluation Itself is Pipeline-Dependent. Our discovery that identical translations receive opposite bias assessments depending on classifier choice has important methodological implications. There is no "ground truth" bias measurement that is independent of the evaluation methodology. Researchers must clearly specify both the translation system and the evaluation classifier when reporting cross-lingual sentiment results. Meta-analyses comparing bias measurements across studies using different evaluation setups should proceed with extreme caution, as they may be comparing incompatible quantities. The field would benefit from standardised evaluation protocols specifying both translation and classification components.

5.3. Comparison to Prior Work

Our findings extend and complement several lines of prior research on translation and sentiment preservation.

Extension of (Mohammad et al., 2016) demonstrated that translation alters sentiment across multiple language pairs using sentiment lexicon analysis. We extend this foundational work by: (1) focusing on neural MT and LLM-based translation rather than statistical MT, (2) employing a fixed classifier methodology that cleanly isolates translation effects, (3) quantifying safety-induced systematic exclusions, and (4) demonstrating pipeline-dependent evaluation where classifier choice reverses measured bias direction. Our 18.9-35.5% shift rates for neural systems are comparable to

their findings for statistical MT, suggesting the fundamental challenge of sentiment preservation persists despite architectural advances.

Building on (Saadany et al., 2021) proposed the Sentiment-Aware Measure (SAM) for Arabic-English translation evaluation in product reviews. We complement this work by: (1) comparing three translation paradigms (NMT, commercial LLM, open-source LLM) systematically rather than evaluating a single system, (2) isolating translation effects from classifier variance through fixed multilingual classification, (3) quantifying safety-introduce coverage gaps that SAM does not address, and (4) extending beyond reviews to social media and financial domains. Our findings validate their concern about sentiment alteration while providing a more comprehensive picture of the landscape.

Extension to Translation Context of (Zouidine and Khalil, 2025). (Zouidine and Khalil, 2025) evaluated open-source LLMs (LLaMA, Mixtral, Gemma) for direct Arabic sentiment classification. We extend this work by examining LLM translation quality in sentiment-sensitive applications, revealing that LLM performance in translation differs substantially from that in direct classification. While LLaMA may perform adequately for direct Arabic classification, our results show it is unsuitable for translation-based pipelines due to both quality degradation (a 19.0% drop in accuracy) and systematic rejections (5.8%). This suggests that evaluations of LLMs for NLP tasks should explicitly consider translation-mediated applications separately from direct classification.

Quantification at Scale. Our study provides the analytical quantification of translation-introduced sentiment distortion using modern neural systems with fixed evaluation methodology at this scale (11,558 samples). Prior work has documented the phenomenon qualitatively or on smaller datasets; we provide precise quantitative estimates of shift rates (18.9-35.5%), accuracy drops (0.9-19.0%), and refusal rates (0-5.8%) that practitioners can use for system design and performance budgeting.

6. Conclusion

We systematically evaluated translation-introduced sentiment shifts across 11,558 Arabic-English samples with a fixed multilingual classifier. Translation changes sentiment for 18.9-35.5% of inputs and reduces accuracy by 0.9-19.0%, confirming that translate-then-classify is not sentiment-neutral. GPT-4o-mini provides the strongest preservation with 0% refusals, Helsinki-NMT offers reproducible middle-ground quality, and LLaMA-3.1-8B performs worst due to both large distortion and safety-driven refusals. The broader conclusion is that sentiment shift is a property of the full pipeline, not just the translator. Future work should separate direction from domain, test additional dialects, and explore mitigation strategies such as sentiment-aware translation or direct multilingual classification.

7. Ethics Statement

This research analyses publicly available sentiment datasets (AJGT, OCLAR, FSA) containing anonymised or public content with no personally identifiable information. We conducted no human-subjects research that required IRB approval.

8. Limitations

Direction-Domain Confound exists due to translation direction with domain (AR→EN: social media/reviews; EN→AR: financial). Baseline accuracy gap (77.4% vs 60.5%) suggests observed patterns may reflect domain rather than directional effects. Matched-domain datasets in both directions were unavailable. The scope is limited; Three datasets cannot capture the full Arabic-English diversity. Results may differ for other domains (legal, medical, literary) and dialects (Gulf, Egyptian, Maghrebi). FSA's 53.6% neutral proportion affects EN→AR SBI interpretation. Single Classifier. XLM-RoBERTa introduces Twitter-domain bias. Results might differ with alternative classifiers (mBERT, XLM, mT5). Prompt Design. Simple prompts may not elicit optimal performance from LLMs. More sophisticated prompting could improve results, especially for LLaMA. Temporal limitation exists as Results reflect specific model versions (GPT-4o-mini, LLaMA-3.1-8B, Helsinki-NMT). Future versions may differ.

Acknowledgements

We thank the creators and maintainers of the AJGT, OCLAR, and FSA datasets for making their resources publicly available. We thank the Program Committee and reviewers of OSACT7 2026 for their constructive feedback that improved this work.

9. Bibliographical References

Muhammad Abdul-Mageed, Mona Diab, and Sandra Kübler. 2014. SAMAR: Subjectivity and sentiment analysis for Arabic social media. *Computer Speech & Language*, 28(1):20–37.

Marwan Al Omari, Moustafa Al-Hajj, Nacereddine Hammami, and Amani Sabra. 2019. Sentiment classifier: Logistic regression for Arabic services'

reviews in Lebanon. *2019 International Conference on Computer and Information Sciences, ICCIS 2019*.

Khaled Mohammad Alomari, Hatem M. ElSherif, and Khaled Shaalan. 2017. Arabic tweets sentimental analysis using machine learning. In *Proceedings of IEA/AIE 2017*, pages 602–610. Springer.

Alexandra Balahur and Marco Turchi. 2014. Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis. *Computer Speech & Language*, 28(1):56–75.

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. XLM-T: Multilingual Language Models in Twitter for Sentiment Analysis and Beyond.

Federico Bianchi, Mirac Suzgun, Giuseppe Attanasio, Paul Röttger, Dan Jurafsky, Tatsunori Hashimoto, and James Zou. 2024. Safety-Tuned LLaMAs: Lessons From Improving the Safety of Large Language Models that Follow Instructions. *12th International Conference on Learning Representations, ICLR 2024*.

Alessio Buscemi and Daniele Proverbio. 2024. ChatGPT vs Gemini vs LLaMA on Multilingual Sentiment Analysis.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Jingfeng Cui, Zhaoxia Wang, Seng Beng Ho, and Erik Cambria. 2023. Survey on sentiment analysis: evolution of research methods and topics. *Artificial Intelligence Review*, 56(8):8469–8510.

Jean Marc Dewaele, K. V. Petrides, and Adrian Furnham. 2008. Effects of trait emotional intelligence and sociobiographical variables on communicative anxiety and foreign language anxiety among adult multilinguals: A review and empirical investigation. *Language Learning*, 58(4):911–960.

Geetika Gautam and Divakar Yadav. 2014. Sentiment analysis of twitter data using machine learning approaches and semantic analysis. In *2014 7th International Conference on Contemporary Computing, IC3 2014*, pages 437–442. Institute of Electrical and Electronics Engineers Inc.

Md Arid Hasan. 2024. Ensemble Language Models for Multilingual Sentiment Analysis.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalisation.

Phillip Keung, Yichao Lu, György Szarvas, and Noah A. Smith. 2020. The Multilingual Amazon Reviews Corpus. *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pages 4563–4568.

Zihao Li, Yucheng Shi, Zirui Liu, Fan Yang, Ali Payani, Ninghao Liu, and Mengnan Du. 2024. Language Ranker: A Metric for Quantifying LLM Performance Across High and Low-Resource Languages.

AI @ Meta Llama Team. 2024. The Llama 3 Herd of Models.

P Lohar, H Afli, A Way Proceedings of the 13th Conference of the, and undefined 2018. 2018. Balancing translation quality and sentiment preservation (Non-archival Extended Abstract). *aclanthology.orgP Lohar, H Afli, A WayProceedings of the 13th Conference of the Association for Machine, 2018•aclanthology.org*.

Pekka Malo, Ankur Sinha, Pekka Korhonen, Jyrki Wallenius, and Pyry Takala. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65(4):782–796.

Md Saef Ullah Miah, Md Mohsin Kabir, Talha Bin Sarwar, Mejdl Safran, Sultan Alfarhood, and M F Mridha. 2024. A multimodal approach to cross-lingual sentiment analysis with ensemble of transformer and LLM. *Scientific Reports*, 14(1).

Saif M. Mohammad, Mohammad Salameh, and Svetlana Kiritchenko. 2016. How translation alters sentiment. *Journal of Artificial Intelligence Research*, 55:95–130.

OpenAI Team OpenAI. 2024. Gpt-4 technical report.

Oumaima Oueslati, Erik Cambria, Moez Ben HajHmida, and Habib Ounelli. 2020. A review of sentiment analysis research in Arabic language. *Future Generation Computer Systems*, 112:408–430.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. *10th Workshop on Statistical Machine Translation, WMT*

2015 at the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015 - Proceedings, pages 392–395.

Hadeel Saadany, Constantin Orăsan, Emad Mohamed, and Ashraf Tantavy. 2021. Sentiment-aware measure (sam) for evaluating sentiment transfer by machine translation systems. In *Proceedings of the Conference Recent Advances in Natural Language Processing - Deep Learning for Natural Language Processing Methods and Applications*, pages 1217–1226. INCOMA Ltd. Shoumen, BULGARIA.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – Building open translation services for the World.

Wenxuan Wang, Zhaopeng Tu, Chang Chen, Youliang Yuan, Jen-tse Huang, Wenxiang Jiao, and Michael R. Lyu. 2024. All Languages Matter: On the Multilingual Safety of Large Language Models.

Mohamed Zouidine and Mohammed Khalil. 2025. Large Language Models for Arabic Sentiment Analysis and Machine Translation. *Engineering, Technology & Applied Science Research*, 15(2):20737–20742.

10. Language Resource References

Al Omari, M., Al-Hajj, M., Hammami, N., and Sabra, A. (2019). Opinion Corpus for Lebanese Arabic Reviews (OCLAR). Sentiment analysis corpus of 3,916 Lebanese Arabic product and service reviews with 3-class labels (positive, neutral, negative). UCI Machine Learning Repository. https://archive.ics.uci.edu/dataset/499/opinion+corpus+for+lebanese+arabic+reviews+oclar

Alomari, K. M., ElSherif, H. M., and Shaalan, K. (2017). Arabic Jordanian General Tweets (AJGT) Dataset. Sentiment corpus of 1,800 Jordanian Arabic tweets. https://github.com/komari6/Arabic-twitter-corpus-AJGT

Barbieri, F., Espinosa Anke, L., and Camacho-Collados, J. (2022). XLM-RoBERTa Sentiment Model (twitter-xlm-roberta-base-sentiment). Multilingual sentiment classifier, pre-trained on 198M tweets. https://huggingface.co/cardiffnlp/twitter-xlm-roberta-base-sentiment

Malo, P., Sinha, A., Korhonen, P., Wallenius, J., and Takala, P. (2014). Financial Sentiment Analysis (FSA) Dataset. Corpus of 5,842 English financial news sentences with 3-class labels. Published in: Malo et al. (2014), JASIST 65(4):782–796. https://doi.org/10.1002/asi.23062

Tiedemann, J. and Thottingal, S. (2020). OPUS-MT Arabic-English Models (opus-mt-ar-en, opus-mt-en-ar). Bidirectional Arabic-English neural translation models. https://huggingface.co/Helsinki-NLP/opus-mt-ar-en