Love, Anger, Sadness, or Just a Rating? Gender Assignment in Human and LLM Translations of Amazon Reviews

Shushen Manakhimova[1,2], Ekaterina Lapshinova-Koltunski[1], Khushi Pitroda[1], Maja Popović[3]

[1]University of Hildesheim, Germany [2]DFKI, Berlin, Germany
[3]IU University, Berlin, Germany
{manakhimova, lapshinovakoltun, pitroda}@uni-hildesheim.de maja.popovic@iu.org

Abstract

We investigate whether product ratings and source-text emotions are associated with first-person gender choices in English-to-Russian and English-to-Croatian translations by professional translators, students, and three LLMs (GPT-4o mini, Llama, and Mistral). We score love, sadness, and anger in the source reviews with a RoBERTa classifier fine-tuned on GoEmotions and fit logistic mixed-effects models of feminine versus masculine forms, accounting for product category, translator group, and repeated translations of the same review. Reviews with high ratings and reviews written in more positive language receive feminine forms more often in both languages. All emotion estimates point in the directions gender stereotypes predict, love and sadness toward feminine forms and anger toward masculine ones, but only love in Russian is statistically significant. We find evidence that emotion–gender associations vary across translator groups in Russian, but no clear evidence of such variation in Croatian.

1 Introduction

Psycholinguistic work shows that when a text offers no morphological or lexical cue to a referent's gender, readers fall back on stereotypes (Gygax et al., 2008). How such defaults surface in language technologies is the subject of a substantial body of work on gender bias (Gamboa et al., 2025; Savoldi et al., 2025). The cues studied are mostly occupation nouns and other role names, built into benchmarks where the referent's gender is either recoverable from context (Stanovsky et al., 2019) or deliberately ambiguous (Prates et al., 2020; Rarrick et al., 2023). A smaller strand addresses first-person speakers whose gender the source leaves unmarked while the target grammar must choose (Vanmassenhove et al., 2018; Habash et al., 2019; Bentivogli et al., 2020). These studies show that systems default to masculine forms and that explicit cues such as a speaker tag or the speaker's voice can shift this default. Still, they leave open whether gender realization also depends on more distributed properties of a text.

We study this in English Amazon reviews translated into Russian and Croatian (Lapshinova-Koltunski et al., 2022). The reviews are written in the first person, so the English source leaves the reviewer's gender unmarked, while Russian and Croatian mark it on past-tense verbs and predicative adjectives; for example, *I bought* becomes *kupio* or *kupila* in Croatian. In this corpus, translators and systems default to masculine forms, and feminine forms vary with the product category (Popović and Lapshinova-Koltunski, 2024). We ask what else in the source text the assigned gender is associated with.

We start with how the author evaluates the product. Positive language tends to be perceived as feminine (Chen et al., 2025), so positive reviews may attract feminine forms. The star rating records the author's evaluation independently of any classifier. Translators and LLMs never saw the rating, so any association with the rating runs through the text.

A rating, however, says how the author judged the product, not how the review is written or what it conveys. The text itself may express emotions, and emotions carry gender associations of their own: love and sadness are commonly believed to be expressed more by women, anger more by men (Plant et al., 2000). An exploratory study on the Russian part of the corpus suggested love as such a cue (Manakhimova and Lapshinova-Koltunski, 2026). We ask whether these emotions carry information beyond the rating, and we check that such associations are not simply a response to positive language.

Human translators and LLMs may respond to such cues differently. In a comparison of human annotations and MT output, Hackenbuchner et al. (2025) found that context shifted human gender as-

Proceedings of the 7th Workshop on Gender Bias in Natural Language Processing (GeBNLP), pages 145–156
November 9, 2026 ©2026 Association for Computational Linguistics

sociations more readily than machine translations, which more often retained a masculine default. We therefore compare professional translators, translation students, and three LLMs on the same reviews.

We ask three research questions:

RQ1: Is the assigned gender associated with the author's evaluation of the product?

RQ2: Do translators and LLMs select the reviewer's grammatical gender in line with the emotions expressed in the source review, as gender stereotypes would predict?

RQ3: Does this differ across translator groups: professional translators, translation students, and three LLMs?

All three questions are tested in Russian and in Croatian. The aim is to describe patterns in translation data, not to identify their causes or to mitigate them, and the analysis is confined to the binary gender the target grammars require.

2 Background

Most evidence on gendered emotion stereotypes comes from the United States. In Plant et al. (2000), U.S. respondents rated 14 of 19 emotions as differing by gender: love, sadness, fear, and sympathy were among the emotions believed to be experienced and expressed more by women, and anger and pride were the only two attributed to men. Other U.S. studies show that the warmth-related traits ascribed to women have persisted or grown stronger over decades (Haines et al., 2016; Eagly et al., 2020). Evidence also comes from outside the United States. Dutch respondents more strongly endorsed beliefs that women, rather than men, display powerless emotions and are made vulnerable by their emotional expression (Timmers et al., 2003). Recent cross-national work confirms this pattern: in public opinion polls across 22 nations in 1995 and 40 nations in 2023, including several in Central and Eastern Europe, women were perceived as more communal than men, with communion including traits such as caring, affectionate, and emotional (Nater et al., 2026). The direction of this stereotype is thus largely shared across cultures, although its strength varies between nations.

Emotion analysis in NLP has paid little attention to the fact that emotion perception is gendered (Plaza-del-Arco et al., 2024a). Where the two have been brought together, it has been in work on bias rather than in emotion analysis itself: the anger–men, sadness–women split found in LLMs (Plaza-del-Arco et al., 2024b) appears also when models are prompted in Bangla (Sadhu et al., 2024), and Weber et al. (2026) show that the stereotype is active in human annotators as well, who identify an author's emotion more accurately when the true label matches the gender stereotype. All of these designs run in one direction: gender is given, and emotion is judged. The converse direction has received less attention. Readers who guess an author's gender from tweets over-attribute joy to women and anger to men (Flekova et al., 2016), and texts rich in positive emotional language are judged more feminine in style (Chen et al., 2025).

Our design draws directly on earlier work on the corpus we use (Section 3) and its extension. Popović and Lapshinova-Koltunski (2024) established the annotation procedure we reuse and, beyond the masculine default and the product-category effect, left open which other source properties drive gendered choices. The GENDER1PERSON test suite extends the category finding to 1,000 reviews and the WMT 2025 systems (Popović and Lapshinova-Koltunski, 2025). Our earlier, exploratory study (Manakhimova and Lapshinova-Koltunski, 2026) found on the Russian data that reviews expressing love more often received feminine forms from the human translators and Llama. We extend this from one emotion to the positivity and emotions of the whole text, in two target languages. To our knowledge, no previous work has asked whether such properties are associated with the gender that translators and LLMs assign to an author the text does not identify.

3 Data

We use DiHuTra (Lapshinova-Koltunski et al., 2022), a public corpus of 196 English Amazon product reviews, 14 from each of 14 product categories, half of them positive (4–5 stars) and half negative (1–2 stars).[1] The reviewer's gender is unknown and, except in two reviews that indicate a female reviewer (Popović and Lapshinova-Koltunski, 2024), is not recoverable from the source.

Each review was translated into Russian and Croatian by professional translators and by translation students, who received no guidelines on grammatical gender (Popović and Lapshinova-Koltunski, 2024), and by GPT-4o mini, Llama 3.1

[1] https://github.com/katjakaterina/dihutra

	Group	f	m	mixed	x
RU	Prof	36	88	1	71
	Stud	27	91	0	78
	GPT-4o mini	21	113	1	61
	Llama	24	101	8	63
	Mistral	23	113	1	59
HR	Prof	60	79	0	57
	Stud	47	86	4	59
	GPT-4o mini	17	105	15	59
	Llama	15	117	7	57
	Mistral	24	105	7	60

Table 1: Distribution of gender labels by translator group and target language ($n = 196$ per row). **f/m**: all first-person gendered words feminine/masculine; **mixed**: both genders present, including forms that realise both at once; **x**: no first-person gendered word.

Nemotron 70B Instruct, and Mistral Large Instruct 2407 (hereafter Llama and Mistral), prompted zero-shot to translate (see Popović et al., 2025).[2] We call these five sources *translator groups*. According to the corpus metadata, the human translations were produced by 10 professionals and 10 students for Russian and by 4 professionals and 20 students for Croatian, most of them women (19 of 20 and 18 of 24).

First-person gendered forms, chiefly past-tense verbs and predicative adjectives, were annotated manually (Popović and Lapshinova-Koltunski, 2024). A translation is labelled **f** or **m** when all such forms are feminine or masculine, **mixed** when both occur, and **x** when there are none (Table 1).

4 Measuring Source-Text Emotions

4.1 Emotion Scores and Clusters

We measure emotion in the English source reviews using a RoBERTa-base model (Liu et al., 2019) fine-tuned on GoEmotions (Demszky et al., 2020).[3] The model was trained on Reddit comments annotated with 27 emotion categories and neutral. Each review receives a score between 0 and 1 for each category, and the scores do not sum to one. Scores are computed on the source only, so all five translations of a review share them.

Emotion annotation can refer to what an author felt, what a text expresses,...or what a reader attributes to the author (Mohammad, 2022), and annotations differ depending on the perspective annotators are asked to take (Buechel and Hahn, 2017). GoEmotions annotators judged the emotions expressed by the writer based on the text alone (Demszky et al., 2020). We therefore treat the scores as an approximate measure of reader-attributed emotion, not as a measure of what individual translators perceived.

We use continuous emotion scores rather than predicted labels. At a threshold of 0.5, 56 of the 196 reviews would receive no label, more often among negative than positive reviews (37% of 1-star and 39% of 2-star reviews against 21% of 4-star and 19% of 5-star reviews). Selecting only the highest-scoring label would also discard information about other emotions in the same review.

Love, sadness, and anger are selected test cases rather than an exhaustive set of gender-stereotyped emotions. Plant et al. (2000) rated broad emotions such as anger and sadness, whereas GoEmotions distinguishes 27 finer labels. We therefore group the labels into broad emotions, following the mapping to Ekman's basic emotions provided by the GoEmotions authors (Demszky et al., 2020). That mapping places love under joy; we keep love separate, as Plant et al. (2000) and Shaver et al. (1987) do:

- **love** = love + caring + desire
- **sadness** = sadness + disappointment + embarrassment + grief + remorse
- **anger** = anger + annoyance + disapproval

Each cluster score is the sum of its member scores, a composite index rather than a probability. The clusters were fixed before model fitting and z-standardised across the 196 reviews.

In product reviews, anger and sadness appear mostly in milder forms. The *anger* label never exceeds 0.104, and *sadness* exceeds 0.3 in only two reviews. Disappointment accounts for 80 % of the summed sadness score, and disapproval and annoyance for 96 % of the summed anger score. The two indices therefore mainly measure disappointment and disapproval and annoyance, which GoEmotions defines as sadness at unmet expectations and as mild anger (Table 6). We keep the head-label names for brevity; whether the stereotypes extend to these milder forms is discussed under Limitations.

4.2 Human Validation

Because our statistical analysis relies on automatically generated emotion scores, we assessed how

[2] Accessed via the ChatAI service (https://chat-ai.academiccloud.de/chat) with default settings, GPT-4o mini in December 2024, Llama and Mistral in January 2025.

[3] https://huggingface.co/SamLowe/roberta-base-go_emotions

well these scores correspond to human judgments of emotion in the source reviews. Two annotators independently labelled 80 of the 196 source reviews, covering all 14 product categories. They used nine frequent GoEmotions labels, which together account for 82 % of the classifier's scores in the corpus: *admiration*, *approval*, *love*, *joy*, *optimism*, *realization*, *disappointment*, *disapproval*, and *annoyance*, plus *neutral*. They were asked to mark the emotion a reader would perceive the author as expressing, not what the author felt, what they felt themselves, or how they would translate the review. They were not told the gender hypothesis (Appendix A.2).

Agreement on positive versus negative labels is high (Cohen's $\kappa = .74$; Cohen, 1960), but agreement on the nine fine-grained labels is weak (Krippendorff's $\alpha = .17$ with MASI distance; Artstein and Poesio, 2008; Passonneau, 2006), as in GoEmotions itself (Demszky et al., 2020).

We then compared the classifier's scores with the annotators' labels. For each label, we paired every review the annotators had marked with it against every review they had not, and counted how often the classifier scored the marked review higher. This is the area under the ROC curve (Hanley and McNeil, 1982), where .5 means chance. For four labels included in our clusters, the AUC values were .85 for love, .86 for disappointment, and .92 for both disapproval and annoyance. Thus, the classifier scores show correspondence with human judgments for these individual labels. However, this does not establish the validity of the complete emotion clusters, since several of their constituent labels were not evaluated and the composite scores themselves were not compared with human judgments. Low agreement on fine-grained emotion labels further limits what this validation can establish.

5 Statistical Analysis

5.1 Analysis sample

Each observation in the statistical analysis is a translation labelled as unambiguously masculine ($m = 0$) or feminine ($f = 1$). Translations without gendered forms (x) or with mixed forms are excluded. Most x labels are explained by the source text: in about three-quarters of cases, none of the five translations of that review has a gendered form. Mixed forms are rare (11 in Russian, 33 in Croatian) and occur mostly in LLM outputs. The *musical-instruments* category contains no feminine realisations, so it is excluded from the models. This yields a final sample of 607 Russian translations (from 139 source reviews) and 613 Croatian translations (from 143 source reviews). The two reviews with indicators of a female reviewer are kept.

5.2 Variables and hypotheses

The dependent variable is the grammatical gender of each translation, feminine against masculine. The models include star rating, product category, and translator group as baseline predictors. The three emotion scores are added in M_1 and M_2.

Product rating is coded as positive (4–5 stars) against negative (1–2 stars). It is a proxy for the evaluation expressed in the text, not a measure of it. Text sentiment is a second measure of evaluation, taken from the text the translators read. We score each English source review with VADER (Hutto and Gilbert, 2014), a lexicon- and rule-based sentiment analyser, and standardise its compound score (-1 to 1) across the 196 reviews. It correlates with the rating at $r = .57$. Product category is a fixed effect, since feminine assignment varies with it (Popović and Lapshinova-Koltunski, 2024, 2025). The translator group (*Group* in the model formulas) is a fixed effect with five levels: professional translators, translation students, GPT-4o mini, Llama, and Mistral.

The emotion predictors are the three cluster scores from Section 4.1. We report their effects as odds ratios (OR). An OR above 1 means that reviews with a higher score are translated with feminine forms more often. An OR below 1 means less often. Each OR is the change for one standard deviation of the score. Before fitting, we fixed the expected directions from the stereotype under test: love and sadness toward feminine forms, anger toward masculine ones ($OR_{\text{love}} > 1$, $OR_{\text{sadness}} > 1$, $OR_{\text{anger}} < 1$). We report each with a 95 % confidence interval and apply no correction for multiple tests.

5.3 Models

We fit logistic mixed-effects models (Jaeger, 2008) separately for Russian and Croatian, with a random intercept for source review. Emotion is measured once per source review and is the same for all five translations of that review. The effective sample for any emotion estimate is therefore the number of source reviews, 139 for Russian and 143 for Croatian, not the number of translations.

Three nested models answer the three questions:

$$M_0 : \text{Gender} \sim \text{Rating} + \text{Category} + \text{Group} + (1 \mid \text{Review}),$$
$$M_1 : M_0 + \text{Love} + \text{Sadness} + \text{Anger},$$
$$M_2 : M_1 + (\text{Love} + \text{Sadness} + \text{Anger}) \times \text{Group}.$$

M_0 tests whether star rating is associated with grammatical gender assignment (RQ1).

M_1 adds love, sadness, and anger to examine their associations with gender assignment after accounting for rating, product category, and translator group (RQ2). We examine the direction and 95% confidence interval of each emotion estimate and compare M_1 with M_0 to test whether the three emotions jointly improve model fit (likelihood-ratio test, 3 df).

M_2 adds interactions between emotions and translator group to test whether these associations differ across the five groups (RQ3). We compare M_2 with M_1 using a likelihood-ratio test (12 df). Group-specific estimates describe the pattern of variation but are not separate confirmatory tests.

Models are fitted in Python with 30-point non-adaptive Gauss–Hermite quadrature and independently refitted in R with `lme4::glmer` using 25-point adaptive Gauss–Hermite quadrature (`nAGQ = 25`) as an implementation check (Bates et al., 2015). Appendix A.1 repeats M_1 with the single *love* label in place of the love cluster.[4]

The sentiment score enters in two ways. For RQ1, it replaces the rating in M_0, giving a second, text-based estimate of the evaluation effect. For RQ2, it is added to M_0 and M_1, and the M_1–M_2 comparison is repeated, to check that the emotion associations are not simply a response to positive language. The primary models use the rating, as fixed before fitting.

6 Results

6.1 RQ1: Product rating and gender assignment

Feminine realizations are more frequent for reviews with high ratings than for reviews with low ratings in both languages (Table 2). The same pattern appears in every translator group, although its magnitude varies between the groups (Figure 1).

The M_0 models confirm this association after controlling for product category and translator group and accounting for repeated translations of the same source review: OR = 2.92 [1.10, 7.73] in Russian and OR = 2.66 [1.35, 5.24] in Croatian (Table 3).

Language	Negative % feminine	Positive % feminine	Difference (pp)
Russian	14.6	26.9	+12.3
Croatian	18.4	31.6	+13.2

Table 2: Observed proportion of feminine gender realizations for negative and positive reviews before adjustment for product category, translator group, and repeated translations. Differences are expressed in percentage points (pp).

With the emotion scores included (M_1), the star-rating estimate decreases, and its confidence interval includes 1 (Table 3). Low ratings tend to co-occur with high anger and sadness scores: of the 65 reviews highest in anger, 61 are negative ($r = -.59$; for sadness, $r = -.43$). With about 140 reviews, the model cannot tell them apart, so the rating interval widens. RQ1 is therefore answered by M_0. Only love, which is weakly related to the rating ($r = .23$), can be separated from it.

Evaluation measured from the text gives the same picture. With the sentiment score in place of the rating, M_0 gives OR = 1.72 [1.02, 2.90] in Russian and 1.69 [1.16, 2.45] in Croatian per standard deviation. The two measures overlap: sentiment adds nothing once the rating is in the model ($p = .36$ and .19).

6.2 RQ2: Emotion effects and their direction

In both Russian and Croatian, love (OR = 1.64 and 1.37) and sadness (OR = 1.27 and 1.33) are associated with higher odds of feminine assignment, whereas anger (OR = 0.81 and 0.72) is associated with lower odds. Thus, all six point estimates follow the stereotype-consistent directions. Most confidence intervals, however, include the null value of OR = 1; only the Russian love confidence interval lies entirely on the predicted side of the null (Table 3). Figure 2 visualises the corresponding effect estimates and their uncertainty. With the single *love* label in place of the cluster, the interval excludes 1 in both languages (Appendix A.1).

Adding the three emotion scores improves the model fit in Croatian ($\chi^2(3) = 8.01$, $p = .046$), with weaker evidence in Russian ($\chi^2(3) = 6.40$, $p = .094$). The model comparison tests the three emotion predictors jointly, whereas the confidence intervals describe individual emotion effects. The

[4]Data and code: https://github.com/happy522/Emotion-as-an-implicit-gender-cue-in-translation

Effect / comparison	Russian	Croatian
Analysis sample	$N = 607$ (139 reviews)	$N = 613$ (143 reviews)
Star rating, M_0 (positive vs. negative)	OR = 2.92 [1.10, 7.73]	OR = 2.66 [1.35, 5.24]
Star rating, M_1 (positive vs. negative)	OR = 2.27 [0.62, 8.38]	OR = 2.10 [0.85, 5.19]
Love	OR = 1.64 [1.07, 2.50]	OR = 1.37 [0.99, 1.89]
Sadness	OR = 1.27 [0.79, 2.04]	OR = 1.33 [0.95, 1.86]
Anger	OR = 0.81 [0.40, 1.63]	OR = 0.72 [0.43, 1.19]
Three emotion scores: M_0 vs. M_1	$\chi^2(3) = 6.40, p = .094$	$\chi^2(3) = 8.01, p = .046$
Emotion × group: M_1 vs. M_2	$\chi^2(12) = 24.82, p = .016$	$\chi^2(12) = 15.12, p = .235$

Table 3: Summary of the mixed-effects regression results. Star-rating ORs compare positive with negative reviews, without (M_0) and with (M_1) the emotion scores in the model. Emotion ORs are from M_1 and correspond to a one-standard-deviation increase in the source-review emotion score; values above 1 indicate higher odds of feminine assignment, values below 1 lower odds. Likelihood-ratio tests compare nested models. Bracketed values are 95% confidence intervals.

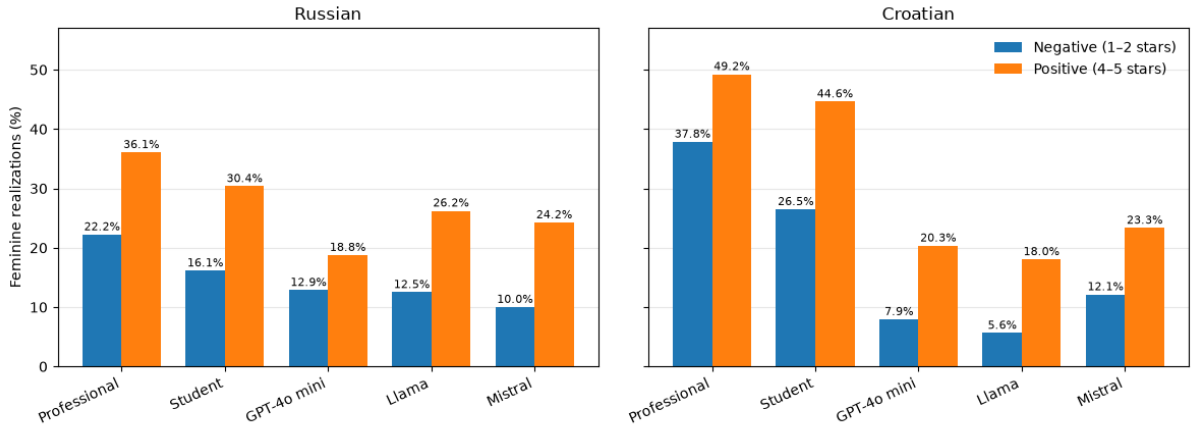


Figure 1: Proportion of feminine gender assigned for reviews with low and high ratings across translator groups in Russian and Croatian. Percentages are calculated within each star-rating and group combination. The figure is descriptive and does not adjust for product category or repeated translations of the same source; adjusted estimates are reported in Table 3.

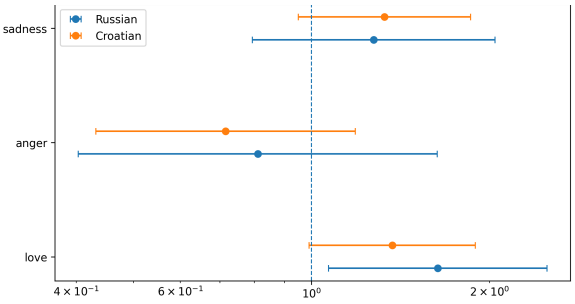


Figure 2: Emotion effects from M_1 for Russian and Croatian: odds ratios of feminine forms per standard deviation of each emotion score, with 95% confidence intervals. The dashed line marks no association (OR = 1).

two analyses therefore address different questions. Because the joint test is sensitive to how the rating is coded (Appendix A.1), we base our conclusions on the individual estimates. The difference in statistical significance does not establish a difference between languages. Adding the sentiment score to M_1 (Section 5.3) does not change these results. It does not improve the model ($\chi^2(1) = 0.42$, $p = .52$ in Russian; $\chi^2(1) = 1.31$, $p = .25$ in Croatian), and the three emotion estimates change by at most 0.09: love 1.63 [1.06, 2.49], sadness 1.33 [0.81, 2.17], and anger 0.90 [0.42, 1.92] in Russian; 1.37 [0.99, 1.89], 1.40 [0.99, 1.98], and 0.80 [0.47, 1.38] in Croatian. The M_1–M_2 comparison is also unchanged ($p = .016$ in Russian and .22 in Croatian, against .016 and .235 without the sentiment score). The emotion associations are therefore not explained by the overall sentiment of the text. The contrast between the two negative emotions is the pattern most relevant to the stereotype, because a response to positivity alone would not produce it. If translators and models responded only to how positive a review is, both sadness and anger would point toward masculine forms. Instead, in both languages, sadness points toward feminine forms (OR = 1.27 and 1.33) and anger toward masculine ones (0.81 and 0.72), as

the stereotype predicts, and the split remains with the sentiment score in the model (sadness 1.33 and 1.40; anger 0.90 and 0.80). All intervals, however, include 1.

6.3 RQ3: Variation across translator groups

The overall M_1–M_2 comparison provides evidence that emotion effects vary across translator groups in the Russian data, and no evidence of such variation in the Croatian data (Table 3). The point estimates differ most for love. In Russian, the estimated love effect is strongest for professional translators, students, and Llama, with ORs of 2.32, 2.55, and 2.57, respectively. At the same time, the GPT-4o mini and Mistral estimates are close to the null (Table 5, Appendix). In Croatian, Llama is the only group whose love interval excludes 1 (OR = 2.03 [1.09, 3.78]); the professional and student estimates are 1.38 and 0.92. For sadness, the group-specific estimates lie between 0.75 and 2.26, and every interval includes 1. For anger, several group-specific estimates are not reliably estimable: with few feminine translations of high-anger reviews in some groups, the intervals span several orders of magnitude (e.g. Llama in Croatian, OR = 0.04 [0.00, 2.19]). All group-specific anger intervals include 1 (Table 5, Appendix).

7 Discussion

Our results link the assigned gender to how positive a review is (RQ1), in line with the perception of positive language as feminine (Chen et al., 2025), and, less clearly, to the emotions it expresses (RQ2). What we cannot tell is what exactly drives the choice: the emotion, its intensity, or the words that express it. The associations are consistent with gendered emotion stereotypes but do not establish that stereotypes drive the choices.

Another possibility is that these associations partly reflect gender-related patterns in language use rather than stereotypes alone. Author gender leaves a trace in writing style: women use more social and emotion words, though the effects are small (Newman et al., 2008). Readers who infer author gender from texts rely mostly on such cues but over-emphasise some, assuming, for instance, that women express more joy (Flekova et al., 2016). We cannot test this account, since author gender is unknown for all but two reviews.

Regarding RQ3, translator groups differ in two ways. First, human translators assign feminine forms more often than any LLM: 23–29% of translations labelled f or m against 16–19% in Russian, and 35–43% against 11–19% in Croatian; in line with the finding that MT keeps the masculine default more often than human readers do (Hackenbuchner et al., 2025). Second, the groups differ in how closely they follow the emotion cues, and this difference does not split along the human–machine line: in Russian, Llama follows the love score about as closely as the two human groups, whereas GPT-4o mini and Mistral do not. The group-specific estimates are uncertain, but the pattern suggests that sensitivity to implicit cues can make a system more human-like and more stereotyped at the same time.

Although most human translators were women (Section 3), the female translators used feminine forms in only 26 % (Russian) and 41 % (Croatian) of their gendered translations, about as often as the male translators (33 % and 36 %). Their choices followed the cues this study examines instead: feminine forms were more frequent for positive than for negative reviews (34 % vs. 18 % in Russian, 48 % vs. 33 % in Croatian) and rose with the love score, from 9 % to 42 % and from 30 % to 49 % between the lowest and the highest third of reviews. Rather than projecting their own gender onto the author, the translators appear to infer it from the text: from its positivity and emotions, as well as from the product it describes (Popović and Lapshinova-Koltunski, 2024).

8 Conclusions

In our data, the gender that translators and LLMs assign to an unmarked author in Russian and Croatian is associated with how positive the review is. Specific emotions show weaker and less certain associations: their estimates point in the directions gender stereotypes predict, love and sadness toward feminine forms and anger toward masculine ones, but only the association with love in Russian is statistically significant. For research on gender bias in translation, our findings suggest that implicit cues in the source text deserve attention alongside the contextual cues commonly examined in previous work. LLMs differ from human translators in how often they default to masculine forms, and, at least in Russian, from one another in how closely they follow such cues, which bias evaluations of translation systems should take into account. Future work should examine larger datasets, more

language pairs and genres, and look beyond binary gender to non-binary forms and strategies that avoid gender marking altogether. It should also validate the emotion measures more thoroughly, with more annotators, all labels that enter the clusters, and a direct comparison of the cluster scores with human judgments.

Limitations

This study has a number of limitations. It covers only one genre, Amazon product reviews, and the sample is small: emotion is scored once per source review and shared by its five translations, so the emotion estimates rest on 139 reviews in Russian and 143 in Croatian. The group-specific estimates for RQ3 are even less precise, since feminine forms are rare in some groups, for example, 15 to 24 per LLM group in Croatian.

Our emotion scores come from a classifier built for emotion analysis, a field with known problems (Plaza-del-Arco et al., 2024a). The scores are a model-based reading of the emotion expressed in the English source. They do not measure what the reviewer felt, what a particular translator perceived, or what the translation conveys, and that a classifier and a translator read a review in the same way is an assumption we cannot test. Our human validation was small: two annotators labelled 80 reviews. They agreed well on positive versus negative labels but poorly on fine-grained ones, and the classifier ranked reviews much as they did. The human annotation provides partial support for four individual emotion labels but does not validate the complete cluster scores. Text sentiment is measured with a general-purpose lexicon, which captures positivity only approximately. Both properties we measure are also contested. Emotion has no single agreed definition: psychology offers competing taxonomies of basic emotions (Ekman, 1992), hierarchies (Shaver et al., 1987), and continuous gradients (Cowen and Keltner, 2017), and ...emotion annotation in NLP often yields low agreement (Weber et al., 2026), and the field lacks consensus on its constructs (Plaza-del-Arco et al., 2024a). The gender coding of specific emotions rests mainly on one U.S. study (Plant et al., 2000). Our clusters extend that coding from the head labels to milder related forms, chiefly disappointment, disapproval, and annoyance, and assume that the stereotypes carry over to them; seven of the eleven labels we aggregate were not rated in that study. Moreover, emotion lexicons carve the space differently across languages (Jackson et al., 2019; Wierzbicka, 1999). The stereotype associations we test have not been measured for speakers of Russian or Croatian. Finally, the target grammars restrict the analysis to a binary masculine–feminine choice, which excludes non-binary referents by design.

Ethical Considerations

Annotation was carried out by two student assistants employed by the research group and compensated as part of their regular paid employment.

Bias Statement

We acknowledge several potential biases that may have influenced our findings. The biases we deal with include harmfulness and stereotyping behavior of systems towards individuals based on their gender. Our focus on Russian and Croatian translations may involve biases related to cultural differences in language use, gender stereotypes, or translation practices. The study does not control for the review author demographics, which may involve biases related to age, gender, or cultural background.

References

Ron Artstein and Massimo Poesio. 2008. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Luisa Bentivogli, Beatrice Savoldi, Matteo Negri, Mattia A. Di Gangi, Roldano Cattoni, and Marco Turchi. 2020. Gender in danger? Evaluating speech translation technology on the MuST-SHE corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6923–6933, Online. Association for Computational Linguistics.

Sven Buechel and Udo Hahn. 2017. Readers vs. writers vs. texts: Coping with different perspectives of text understanding in emotion annotation. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 1–12, Valencia, Spain. Association for Computational Linguistics.

Hongyu Chen, Neele Falk, Michael Roth, and Agnieszka Falenska. 2025. "Feels feminine to me": Understanding perceived gendered style through human annotations. In *Proceedings of the 2025 Conference*

on Empirical Methods in Natural Language Processing, pages 31459–31480, Suzhou, China. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Alan S. Cowen and Dacher Keltner. 2017. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the National Academy of Sciences*, 114(38):E7900–E7909.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.

Alice H. Eagly, Christa Nater, David I. Miller, Michèle Kaufmann, and Sabine Sczesny. 2020. Gender stereotypes have changed: A cross-temporal meta-analysis of U.S. public opinion polls from 1946 to 2018. *American Psychologist*, 75(3):301–315.

Paul Ekman. 1992. An argument for basic emotions. *Cognition and Emotion*, 6(3–4):169–200.

Lucie Flekova, Jordan Carpenter, Salvatore Giorgi, Lyle Ungar, and Daniel Preoţiuc-Pietro. 2016. Analyzing biases in human perception of user age and gender from text. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 843–854, Berlin, Germany. Association for Computational Linguistics.

Lance Calvin Lim Gamboa, Yue Feng, and Mark G. Lee. 2025. Social bias in multilingual language models: A survey. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27857–27880, Suzhou, China. Association for Computational Linguistics.

Pascal Gygax, Ute Gabriel, Oriane Sarrasin, Jane Oakhill, and Alan Garnham. 2008. Generically intended, but specifically interpreted: When beauticians, musicians, and mechanics are all men. *Language and Cognitive Processes*, 23(3):464–485.

Nizar Habash, Houda Bouamor, and Christine Chung. 2019. Automatic gender identification and reinflection in Arabic. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 155–165, Florence, Italy. Association for Computational Linguistics.

Janiça Hackenbuchner, Arda Tezcan, and Joke Daems. 2025. Gender bias and the role of context in human perception and machine translation. *Computational Linguistics in the Netherlands Journal*, 14:215–239.

Elizabeth L. Haines, Kay Deaux, and Nicole Lofaro. 2016. The times they are a-changing . . . or are they not? A comparison of gender stereotypes, 1983–2014. *Psychology of Women Quarterly*, 40(3):353–363.

James A. Hanley and Barbara J. McNeil. 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36.

Clayton J. Hutto and Eric Gilbert. 2014. VADER: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1):216–225.

Joshua Conrad Jackson, Joseph Watts, Teague R. Henry, Johann-Mattis List, Robert Forkel, Peter J. Mucha, Simon J. Greenhill, Russell D. Gray, and Kristen A. Lindquist. 2019. Emotion semantics show both cultural variation and universal structure. *Science*, 366(6472):1517–1522.

T. Florian Jaeger. 2008. Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, 59(4):434–446.

Ekaterina Lapshinova-Koltunski, Maja Popović, and Maarit Koponen. 2022. DiHuTra: A parallel corpus to analyse differences between human translations. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1751–1760, Marseille, France. European Language Resources Association.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *Preprint*, arXiv:1907.11692.

Shushen Manakhimova and Ekaterina Lapshinova-Koltunski. 2026. Can emotions signal gender? Investigating implicit cues in human and LLM translations of Amazon reviews. In *Proceedings of the 4th Workshop on Gender-Inclusive Translation Technologies (GITT 2026)*, pages 73–80, Tilburg, The Netherlands. European Association for Machine Translation.

Saif M. Mohammad. 2022. Ethics sheet for automatic emotion recognition and sentiment analysis. *Computational Linguistics*, 48(2):239–278.

Christa Nater, David I. Miller, Alice H. Eagly, and Sabine Sczesny. 2026. Gender stereotypes across nations relate to the social position of women and men: Evidence from cross-cultural public opinion polls. *Proceedings of the National Academy of Sciences*, 123(2):e2510180122.

Matthew L. Newman, Carla J. Groom, Lori D. Handelman, and James W. Pennebaker. 2008. Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes*, 45(3):211–236.

Rebecca J. Passonneau. 2006. Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, Genoa, Italy. European Language Resources Association (ELRA).

E. Ashby Plant, Janet Shibley Hyde, Dacher Keltner, and Patricia G. Devine. 2000. The gender stereotyping of emotions. *Psychology of Women Quarterly*, 24(1):81–92.

Flor Miriam Plaza-del-Arco, Alba A. Cercas Curry, Amanda Cercas Curry, and Dirk Hovy. 2024a. Emotion analysis in NLP: Trends, gaps and roadmap for future directions. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5696–5710, Torino, Italia. ELRA and ICCL.

Flor Miriam Plaza-del-Arco, Amanda Cercas Curry, Alba Curry, Gavin Abercrombie, and Dirk Hovy. 2024b. Angry men, sad women: Large language models reflect gendered stereotypes in emotion attribution. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7682–7696, Bangkok, Thailand. Association for Computational Linguistics.

Maja Popović and Ekaterina Lapshinova-Koltunski. 2024. Gender and bias in Amazon review translations: By humans, MT systems and ChatGPT. In *Proceedings of the 2nd International Workshop on Gender-Inclusive Translation Technologies*, pages 22–30, Sheffield, United Kingdom. European Association for Machine Translation (EAMT).

Maja Popović and Ekaterina Lapshinova-Koltunski. 2025. GENDER1PERSON: Test suite for estimating gender bias of first-person singular forms. In *Proceedings of the Tenth Conference on Machine Translation*, pages 800–822, Suzhou, China. Association for Computational Linguistics.

Maja Popović, Ekaterina Lapshinova-Koltunski, and Anastasiia Göldner. 2025. Did I (she) or I (he) buy this? Or rather I (she/he)? Towards first-person gender neutral translation by LLMs. In *Proceedings of the 3rd Workshop on Gender-Inclusive Translation Technologies (GITT 2025)*, pages 64–73, Geneva, Switzerland. European Association for Machine Translation.

Marcelo O. R. Prates, Pedro H. Avelar, and Luís C. Lamb. 2020. Assessing gender bias in machine translation: A case study with Google Translate. *Neural Computing and Applications*, 32(10):6363–6381.

Spencer Rarrick, Ranjita Naik, Varun Mathur, Sundar Poudel, and Vishal Chowdhary. 2023. GATE: A challenge set for gender-ambiguous translation examples. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, pages 845–854, Montréal, QC, Canada. Association for Computing Machinery.

Jayanta Sadhu, Maneesha Saha, and Rifat Shahriyar. 2024. An empirical study of gendered stereotypes in emotional attributes for Bangla in multilingual large language models. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 384–398, Bangkok, Thailand. Association for Computational Linguistics.

Beatrice Savoldi, Jasmijn Bastings, Luisa Bentivogli, and Eva Vanmassenhove. 2025. A decade of gender bias in machine translation. *Patterns*, 6(6):101257.

Phillip Shaver, Judith Schwartz, Donald Kirson, and Cary O'Connor. 1987. Emotion knowledge: Further exploration of a prototype approach. *Journal of Personality and Social Psychology*, 52(6):1061–1086.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

Monique Timmers, Agneta H. Fischer, and Antony S. R. Manstead. 2003. Ability versus vulnerability: Beliefs about men's and women's emotional behaviour. *Cognition and Emotion*, 17(1):41–63.

Eva Vanmassenhove, Christian Hardmeier, and Andy Way. 2018. Getting gender right in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3003–3008, Brussels, Belgium. Association for Computational Linguistics.

Sabine Weber, Lynn Greschner, and Roman Klinger. 2026. Less is more? The role of demographic author information in emotion classification of ambiguous text. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference*, pages 8147–8161, Palma de Mallorca, Spain. ELRA Language Resources Association.

Anna Wierzbicka. 1999. *Emotions across Languages and Cultures: Diversity and Universals*. Cambridge University Press, Cambridge.

A Appendix

A.1 Sensitivity analyses

Effect / comparison	Russian	Croatian
Love (single label), OR	1.73 [1.15, 2.61]	1.41 [1.02, 1.95]
Three emotion scores, M_0 vs. M_1	$p = .041$	$p = .033$

Table 4: Sensitivity analysis replacing the composite love cluster (*love* + *caring* + *desire*) with the single GoEmotions *love* score. Odds ratios correspond to a one-standard-deviation increase in the standardised single-love score. Bracketed values are 95% confidence intervals. The sadness and anger clusters and all covariates are unchanged from the primary M_1 specification.

We ran two robustness checks.

Single love label. The love cluster combines three labels (love, caring, and desire). Replacing it with the single love label strengthens the love association: its 95% confidence interval now excludes 1 in both languages (Russian OR = 1.73 [1.15, 2.61]; Croatian OR = 1.41 [1.02, 1.95]; Table 4). The love result therefore does not depend on how the cluster is built.

Numeric rating. Coding the rating as 1, 2, 4, or 5 stars instead of positive versus negative weakens the joint test of the three emotion scores ($p = .114$ instead of .094 in Russian; $p = .062$ instead of .046 in Croatian). The joint test is thus sensitive to how the rating is coded, which is one reason we interpret the individual emotion estimates rather than this test. The test of differences between translator groups is unaffected: there is evidence of variation in Russian ($p = .018$) and, as before, no evidence of it in Croatian ($p = .253$).

A.2 Annotation Guideline

This appendix documents the annotation task described in Section 4.2.

Setup. The task was run in a spreadsheet. Each row held one English source review, and each column one label. Annotators marked a label by typing `x` in its cell. A free-text *notes* column was available for comments.

The file contained 88 rows: 80 distinct reviews, eight of which appeared a second time. The rows were split into four sessions of 22. Each repeated review fell in a different session from its first appearance. Row order was randomised separately for each annotator. The two annotators worked independently. There was no adjudication and no discussion of individual items.

Instructions. The annotators received the following text before starting.

> *This is part of a research project on how emotion expressed in a text affects the translation process. Your job is to say how each review comes across to you as a reader.*
>
> *There is no answer key. We are not checking your answers against anything, and there are no right or wrong responses. Where annotators disagree, that disagreement is itself informative.*
>
> *88 reviews, four sessions of 22. Please leave a break between sessions and do not go back to revise earlier ones.*
>
> *Your answers are stored under a letter, not your name, and are reported only in aggregate. Participation is voluntary — you may stop, skip anything, or withdraw at any point.*
>
> *If anything is unclear, please ask before you start rather than guessing.*
>
> *How to fill it in. Put an x in a column if that label applies. Leave the cell empty if it does not. You can mark several labels for one review.*
>
> *Mark the emotion a reader would perceive the author as expressing — not what you think the author truly felt, not what you feel while reading, and not how you would handle it as a translator.*

Labels. Table 6 lists the labels as they were shown. The nine emotion labels are GoEmotions categories, with the definitions used in that dataset (Demszky et al., 2020). The classifier and the annotators therefore worked from the same labels and the same definitions. Three further options were available: *neutral*, *other_emotion*, and *too_difficult*. The task was multi-label. Any number of labels could be marked, and no ranking or dominant emotion was requested.

Use of the extra options. The two non-emotion options were barely used. *other_emotion* was marked once, and *too_difficult* never. *Neutral* was marked 14 times by one annotator and three times by the other. The analysis in Section 4.2 is therefore based on the nine emotion labels and *neutral*.

Emotion	Translator group	Russian OR [95% CI]	Croatian OR [95% CI]
Love	Professional	2.32 [1.13, 4.76]	1.38 [0.81, 2.37]
	Student	2.55 [1.22, 5.33]	0.92 [0.52, 1.64]
	GPT-4o mini	1.14 [0.59, 2.23]	1.27 [0.70, 2.31]
	Llama	2.57 [1.19, 5.53]	2.03 [1.09, 3.78]
	Mistral	1.02 [0.54, 1.96]	1.43 [0.81, 2.52]
Anger	Professional	0.72 [0.26, 1.99]	0.76 [0.39, 1.47]
	Student	1.88 [0.73, 4.86]	0.73 [0.35, 1.51]
	GPT-4o mini	1.18 [0.41, 3.42]	0.12 [0.01, 1.76]
	Llama	0.32 [0.06, 1.83]	0.04 [0.00, 2.19]
	Mistral	0.10 [0.01, 1.61]	1.02 [0.43, 2.39]
Sadness	Professional	1.75 [0.89, 3.45]	1.29 [0.79, 2.09]
	Student	0.75 [0.34, 1.63]	1.54 [0.92, 2.58]
	GPT-4o mini	1.35 [0.65, 2.82]	1.43 [0.66, 3.09]
	Llama	1.32 [0.60, 2.91]	2.26 [0.92, 5.59]
	Mistral	1.60 [0.68, 3.76]	0.97 [0.51, 1.84]

Table 5: Group-specific emotion effects estimated from M_2. Odds ratios (ORs) represent the change in the odds of feminine grammatical-gender assignment associated with a one-standard-deviation increase in the corresponding source-review emotion score within each translator group. Bracketed values are 95% confidence intervals. These estimates are provided to characterise the pattern underlying the RQ3 overall M_1–M_2 interaction test and are not interpreted as separate confirmatory pairwise tests between groups.

Label	Definition shown
admiration	Finding something impressive or worthy of respect.
approval	Having or expressing a favorable opinion.
love	A strong positive emotion of regard and affection.
joy	A feeling of pleasure and happiness.
optimism	Hopefulness and confidence about the future or the success of something.
realization	Becoming aware of something.
disappointment	Sadness or displeasure caused by the nonfulfillment of one's hopes or expectations.
disapproval	Having or expressing an unfavorable opinion.
annoyance	Mild anger, irritation.
other_emotion	Some emotion is expressed, but none of the nine above capture it. Write which one, in your own words.
neutral	You are not certain that any emotion is expressed. Not "no emotion here" — "I am not confident any emotion is expressed". A normal answer, not a cop-out.
too_difficult	The review is particularly difficult to label. Tick this and select no emotions.

Table 6: Labels and definitions shown to the annotators. The nine emotion labels and their definitions are taken from GoEmotions (Demszky et al., 2020).