Crowd-Based Evaluation of Emotion Intensity Preservation in Spanish–Basque Tweet Machine Translation

Nora Aranberri
HiTZ Basque Center for Language Technologies - Ixa NLP Group
University of the Basque Country (UPV/EHU)
nora.aranberri@ehu.eus

Abstract

Machine translation (MT) systems perform well on standard benchmarks, yet their ability to preserve emotional meaning in informal user-generated content—particularly for low-resource languages—remains underexplored. We investigate the preservation of emotion intensity in Spanish–Basque tweet translation, focusing on Basque, an under-represented language in MT research. We compile a small, controlled corpus of Spanish reaction tweets and evaluate Basque translations from three publicly available systems through a crowd-based study. While all systems achieve comparable and above mid-range accuracy and fluency, emotion intensity is systematically attenuated in the translations, with greater loss for more emotionally intense inputs. A follow-up on highly emotional tweets shows that LLM prompting reduces emotion loss, yet substantial attenuation remains, highlighting emotion preservation as a persistent challenge in Spanish–Basque MT.

1 Introduction

Machine translation (MT) quality has improved substantially in recent years, with neural approaches—particularly transformer-based architectures and large multilingual models—now dominating research and practice (Ataman et al., 2025; Chatterji et al., 2025). According to the WMT 2025 findings, current systems perform well on standard benchmarks, especially for descriptive, emotionally neutral, and well-structured texts. These gains, however, are uneven across languages and domains, motivating the development of more challenging evaluation settings with linguistically complex data (Kocmi et al., 2025). In this context, creative and noisy user-generated content (UGC), such as tweets, remains particularly challenging for MT (Barbieri et al., 2020; Popovic et al., 2024), and exposes limitations in handling linguistic variability, metaphorical language, and contextual cues. These challenges are amplified in low-resource settings, where limited parallel data and cultural specificity constrain both training and evaluation (Joshi et al., 2020). Basque, despite recent corpus development efforts (Etchegoyhen et al., 2018; Barnes et al., 2018; Romero et al., 2022; Heredia et al., 2025), remains under-represented in MT research on UGC, particularly in informal, affect-rich domains.

Prior work shows that the preservation of emotional intensity is particularly important when translating between languages with unequal resources or sociolinguistic status, yet MT systems often fail to retain fine-grained affective cues (Briakou and Carpuat, 2021). Still, many online platforms (e.g., X) automatically translate UGC without human supervision. While this can increase accessibility for minority languages like Basque, it also risks distorting emotionally salient content that shapes cross-linguistic interpretation. Given the pervasive and emotionally charged nature of social media, it is essential to examine whether—and to what extent—emotional content is preserved in MT outputs, particularly for these languages. In this paper, we investigate Spanish–Basque tweet translation, analyzing how different MT systems handle accuracy, fluency, and emotion intensity preservation. We construct a small tweet corpus and conduct a crowd-based evaluation with Basque-speaking intensity increases, emotion intensity loss in translation becomes progressively more pronounced.

2 Related Work

Tweets differ markedly from formal written genres in their linguistic, structural, and pragmatic properties. Twitter has been described as a highly conversational medium shaped by platform-specific conventions such as mentions, retweets, and hashtags (Honeycutt and Herring, 2009). Tweets are typically brief and informal, exhibiting non-standard orthography, creative spelling, abbreviations, ex-

The Proceedings for the 15th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis (WASSA 2026), pages 123–133

pressive punctuation, and multimodal markers such as emojis and hashtags (Eisenstein, 2013; Zappavigna, 2012). Prior work highlights the frequent use of stance-taking, irony, and humor, particularly in emotionally charged contexts (Barbieri and Saggion, 2014). Twitter discourse also relies heavily on shared context, ongoing events, and cultural references (Feldman et al., 2018), while lexical variation reflects social identity factors such as gender and community affiliation (Bamman et al., 2014). These characteristics make tweets a challenging genre for MT and emotion-related tasks.

Tweets and microblogs have been widely studied in recent work on emotion and MT, with analyses focusing on how emotion intensity is altered during translation. Fukuda and Jin (2022) adopt a quantitative approach to emotion preservation by measuring how emotion scores assigned to tweets change before and after translation. Using a Japanese–English Twitter dataset, they apply an automatic emotion analysis system, enabling direct comparison of emotion intensity across multiple categories (joy, relief, fear, anger...). They show that MT systematically alters emotional profiles, with translated tweets exhibiting reduced variance and attenuated intensity, particularly for high-arousal emotions. Crucially, these effects are observed even when translations are otherwise fluent and semantically adequate, indicating that emotion distortion is not captured by standard MT evaluation metrics.

Extending this analysis to a multilingual setting, Saadany et al. (2023) use emotion-annotated Twitter data in English, Arabic, and Spanish, and analyze translations by commercial MT services, including Google Translate, Microsoft Bing Translator, and Amazon Translate. Their study combines automatic detection of emotion mismatches with targeted manual error analysis, and allows them to identify Twitter-specific linguistic phenomena—such as slang, emojis, code-switching, and non-standard spelling—that frequently trigger emotion distortion even when translations remain fluent and semantically plausible.

Qian et al. (2023) continue in this line for Chinese–English translation of emotion-loaded microblog texts. Using translations generated by Google Translate, they conduct fine-grained manual evaluation based on an MQM-inspired scheme that explicitly targets emotion preservation and distortion. Again, their analysis shows that a substantial proportion of MT outputs fail to preserve the original emotional meaning, and their error analysis highlights recurring triggers such as polysemy, negation, and abbreviations.

To the best of our knowledge, no prior work has examined the preservation of emotional intensity of translated social media content into Basque, although recent studies have begun to characterize informal language in social media (ILSC) for this language. The challenges typically associated with ILSC—non-standard orthography, emojis, abbreviations, cultural references, and emotionally charged expressions—are amplified by Basque's status as a minority language undergoing revitalization, in which informal digital registers are still emerging. From a sociolinguistic perspective, Elordui (2025) identifies three defining features of Basque ILSC—transcription of spoken language, dialectal variation, and code-switching—based on the GazteSare corpus (Elordui et al., 2020). However, as GazteSare draws on Facebook and Instagram data, its findings may not fully capture ILSC on platforms such as X, where communication is more public, reactive, and performative, often intensifying expressive and affective language.

These findings motivate a focused investigation of emotion preservation in the machine translation of tweets into Basque.

3 Experimental Setup

3.1 Spanish tweets

Our first step was to obtain a corpus of informal Spanish tweets. Although a parallel Spanish–Basque dataset would be ideal, no such resource exists to our knowledge. While several tweet collections have been developed for Spanish and Basque (Alegria et al., 2013; Pérez et al., 2019; Amores et al., 2021; Camargo Fernández, 2021), these are typically domain-specific, thematically limited, or dated. Also, they do not all involve informal language and tend to rely on translated data. Given the rapidly evolving nature of online discourse, we constructed a new dataset tailored to our study.

Because direct collection of real tweets is constrained by platform policies and legal considerations, we designed a controlled elicitation task to simulate authentic tweet production. To capture informal, creative, and emotionally expressive language, we recruited 19 native Spanish-speaking volunteers aged 19–25, a demographic chosen for their frequent use of platforms such as X.

Volunteers were asked to produce reaction

tweets within a form-based interface. To approximate a realistic user experience, we embedded the stimuli in a X-like layout: each stimulus displayed a headline and subheading tweet adapted from either a mainstream Spanish newspaper (El País) or a satirical outlet (El Mundo Today), accompanied by an illustrative image, metadata such as likes and retweets, and surrounding interface elements (e.g., recommended accounts, trending topics). The topics were chosen to encourage spontaneous, playful, or cheeky reactions; our interest lay primarily in the linguistic form, tone, and stylistic features of the responses. For each stimulus, volunteers were provided with an empty text box in which to write their reaction tweets.

Before beginning the task, we conducted a brief group discussion in which participants reflected on what they perceived as characteristic features of tweets (e.g., brevity, irony, informality). This served to situate them in an everyday social media mindset and mitigate the risk of producing unnaturally formal or academically influenced language. In total, we collected 472 reaction tweets, with each volunteer contributing an average of approximately 24–25 instances.

3.2 Baseline Basque translations

To assess how different systems handle the automated translation of informal tweet language, we compared the outputs of three publicly accessible MT and LLM-based tools. The first system was Itzuli, a neural MT service developed and maintained by the Basque Government[1]. The second was ChatGPT 4-o-mini, a proprietary large language model accessible via the ChatGPT platform[2]. Finally, we included Latxa, a Llama-based open model specifically adapted for Basque and recently released by the HiTZ Center (Sainz et al., 2025)[3]. Together, these systems represent a generic NMT engine, a generic commercial multilingual LLM, and an open-source Basque-specialized LLM, allowing us to examine a diverse range of approaches to Spanish–Basque translation.

All translations were obtained using each system's default configuration. For the LLMs, we used a simple translation prompt instructing the model to "translate the following tweet," without providing additional information or constraints. This design choice allows us to compare the baseline performance of each model in a minimal-supervision setting and focus on the inherent capabilities of each system in preserving the linguistic and emotional intensity of informal social-media discourse.

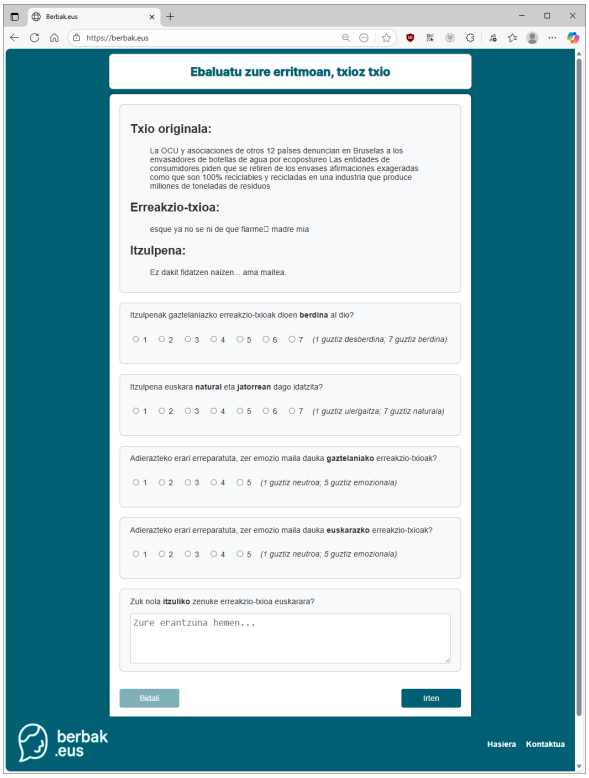


Figure 1: Sample screenshot of the evaluation interface.

3.3 Evaluation attributes and platform

Although our primary focus is the preservation of emotional intensity, we begin with the two dimensions that are typically central in translation quality evaluation: whether the target text preserves the meaning of the source (accuracy) and whether it reads naturally (fluency). However, tweets place particular emphasis on emotional content, given their informality, expressiveness, nonstandard spelling, and dense affective cues. Accordingly, we include emotion intensity as a separate, explicit evaluation attribute, assessing whether the translation preserves the emotional intensity of the original text.

This issue is especially relevant in the case of Basque. First, most training data–whether for NMT systems or LLMs–primarily covers standard or formal registers, which means that models have limited exposure to non-standard, colloquial uses of the language. Second, although Basque has undergone substantial normalization, its informal varieties continue to evolve unevenly across regions

[1]https://www.euskadi.eus/itzuli/
[2]https://chatgpt.com/
[3]https://latxa.hitz.eus/

and speaker communities. With 5 dialects, about 15 subdialects and over 50 recognized varieties, as well as widespread code-switching with Spanish/French and English, what counts as "natural" Basque in informal settings can differ significantly across speakers. As a consequence, the perception and expression of emotion in Basque is highly variable, making it an especially challenging language for evaluating the adequacy of informal MT. To address this gap, we also gather reference tweets by participants for further research.

We designed a dedicated web page to collect fully anonymous evaluations from the Basque-speaking community (Figure 1). Participants were asked to indicate their linguistic background (native/C1 level, specialization, and social media usage), but no verification was required. Previous crowd-based evaluation campaigns (Aranberri et al., 2017; Aranberri, 2024; Sainz et al., 2025) have shown that contributions from non-Basque speakers are extremely rare, probably because the initiative is not relevant to this profile and because the dissemination channels typically reach Basque-speaking networks. Lowering participation barriers leads to higher engagement, and given the low risk of contamination, we opted not to perform additional checks. For anomalous cases, reference translations could be manually reviewed and responses discarded.

Once participants completed the introductory section and profile questions, they proceeded to the evaluation interface. For each item, they were shown the initial Spanish tweet, its Spanish reaction tweet, and the automatically translated Basque version of the latter. They rated accuracy ("does the translation express the same meaning as the Spanish reaction tweet?") and fluency ("does the translation read naturally in Basque?") using a 7-point Likert scale. They also rated the emotional intensity of both the Spanish reaction tweet and its Basque translation on a 5-point Likert scale. Although absolute intensity values were collected, note that our main interest lies in the comparison between the source and target languages. Finally, participants were offered an open text box to provide their own Basque translation under the prompt: "How would you translate this reaction tweet into Basque?"

Given the size of the dataset, we set to collect a single evaluation per translation instance. While this limits our generalizability, it provides a valuable first step toward understanding how informal, emotion-rich tweet language is translated into Basque and lays the groundwork for more extensive evaluation in future work.

4 Crowd assessment results for baseline translations

We collected 1,098 individual evaluation responses. Because the same reaction tweet was not presented to evaluators twice and the task was open for a limited amount of time, the resulting dataset exhibited certain imbalances: some translations did not receive any evaluation, while others received several, independent assessments. To improve the comparability of the results, we cleaned the data.

First, we removed all reaction tweets for which any of the three MT systems lacked evaluation. Next, we resolved duplicate evaluations, defined as cases in which the same reaction tweet–system pair received more than one assessment. To do so, we distinguished between specialist and non-specialist evaluators. We classified evaluators as specialists if they were native or C1-level Basque speakers and met at least one of the following criteria: training in translation studies or philology, professional experience as translators, or work in communication-related fields. All remaining evaluators—including those who did not report their language proficiency or lacked a language-related background—were classified as non-specialists.

For duplicate cases evaluated exclusively by non-specialists, we merged the assessments by averaging their scores. Similarly, for duplicates evaluated only by specialists, we averaged the specialist scores. Finally, for cases with mixed evaluator profiles, we retained only the specialist assessments; when multiple specialist evaluations were available, these were again merged by averaging the scores.

After completing this cleaning process, our dataset consisted of 1,002 evaluations for 334 reaction tweets, with exactly one evaluation per MT system for each reaction tweet.

The evaluations were predominantly provided by self-reported proficient participants: 75% of the responses came from native Basque speakers and 91% from evaluators who reported a C1 level or higher. Only 28 responses (2.79%) were submitted by non-native evaluators who did not report a C1 level. After manually reviewing the reference translations produced by the latter group, we chose to retain all responses, as their contributions were of satisfactory quality.

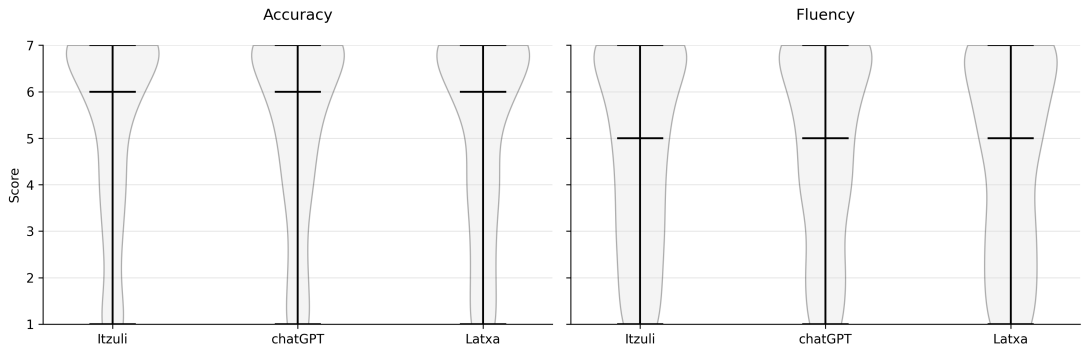


Figure 2: Distribution of accuracy and fluency scores per system. Horizontal lines indicate median values.

4.1 Accuracy and Fluency

Focusing on general translation quality, we analyze two attributes: accuracy and fluency (Figure 2). In a 1-7 Likert scale, accuracy scores fall in the range of approximately 5.1-5.3, while fluency scores are slightly lower, around 4.6-4.8. On average, this reflects a difference of roughly 0.5 scale points between accuracy and fluency, with accuracy consistently rated higher. Both dimensions are about 0.5-1.7 points above the scale midpoint of 4, indicating that the systems clearly surpass mid-range performance. Yet, at the same time, the gap of nearly two points between current performance and the upper end of the scale suggests substantial remaining room for improvement.

Accuracy scores are similar across systems, with Itzuli (M = 5.28, SD = 2.03) and chatGPT (M = 5.21, SD = 1.97) performing nearly identically and Latxa slightly lower (M = 5.10, SD = 2.05). A comparable pattern is observed for fluency, where chatGPT (M = 4.78, SD = 2.02) and Itzuli (M = 4.70, SD = 2.05) again score similarly, followed by Latxa (M = 4.58, SD = 2.08). All three systems cover the full 1-7 scale, with distributions showing similar variability and reflecting variation in translation performance.

To test whether these differences are statistically meaningful, we applied complementary analyses. A one-way ANOVA found no significant effect of system for either accuracy (F = 0.67, p = .512) or fluency (F = 0.83, p = .436). The Kruskal–Wallis test, which does not assume interval-scale Likert data confirmed this resu (accuracy: H = 1.69, p = .429; fluency: H = 1.58, p = .453).

4.2 Preservation of emotion intensity

We next turn to emotion preservation. Spanish intensity scores span the entire scale (1–5), with a mean of 3.61 (SD = 1.14). The distribution is skewed toward mid-to-high intensity: the median is 4, and the interquartile range spans from 3 to 4. This indicates that most reaction tweets express noticeable, though not extreme, emotion intensity (Figure 3). Although assigning emotion intensity to a tweet is inherently subjective, evaluators rated the Spanish reaction tweets set consistently across systems (Itzuli: 3.63, ChatGPT: 3.61, Latxa: 3.60), making general comparisons possible (Itzuli: 3.63, ChatGPT: 3.61, Latxa: 3.60).

When examining how systems preserved this emotional intensity in their Basque outputs, we observe a clear downward shift (Figure 3). While Basque scores also cover the full available range, the mean drops to 2.80 (SD = 1.29), nearly one full point lower than the Spanish originals. This indicates that translations tend to dampen emotion intensity, producing outputs that are more neutral or less expressive than their Spanish counterparts. The median Basque intensity score is 3, reflecting a general attenuation rather than an abrupt loss of affect.

This trend is consistent across systems, with similar mean scores for Itzuli (2.83), ChatGPT (2.76), and Latxa (2.82). The consistency of this reduction across models suggests that emotion dampening is not system-specific, but rather reflects a broader challenge in MT.

Aware of the subjectivity involved in determining emotional intensity, our primary interest lies not in absolute scores but in how intensity shifts during translation. We therefore analyse the *intensity shift*, defined as the difference between the Basque and Spanish emotion intensity ratings for each translation. Negative values indicate emotional attenuation, while positive values reflect amplification.

Across all systems, the mean intensity shift is consistently negative. To better understand this ef-

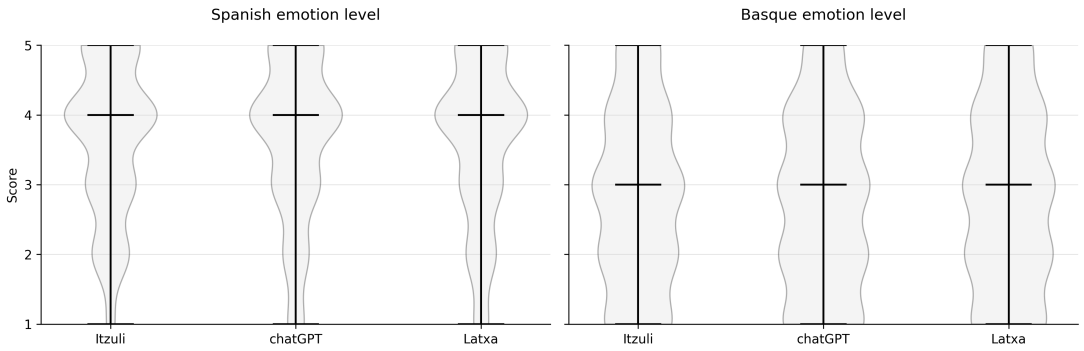


Figure 3: Distribution of emotion intensity scores for Spanish source tweets and Basque translations. Horizontal lines indicate median values.

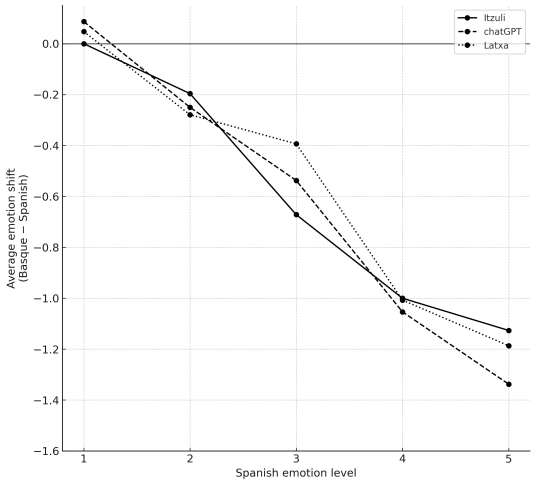
fect, we examine how shifts vary across the Spanish scale and whether systems differ in their handling of low-, medium-, and high-intensity inputs.

Across the three systems, emotion intensity preservation behaviour is remarkably consistent at each Spanish emotion level, with inter-system differences generally below 0.5 points (Figure 4). For tweets with low intensity (levels 1 and 2), all systems preserve emotion closely, with shifts near zero. As source intensity increases, however, attenuation becomes more pronounced: tweets rated at level 3 lose approximately half a point on average, while high-intensity tweets (levels 4 and 5) lose between one and 1.3 points. This trend indicates that input with higher emotional intensity leads to greater loss in translation, regardless of the system used.

To assess whether differences between systems are significant, we compared the intensity-shift scores using both parametric and non-parametric tests. A one-way ANOVA found no significant effect of system (F = 0.42, p = .659). A Kruskal–Wallis test similarly showed no significant differences (H = 1.67, p = .434). Together, these results indicate that the system used to obtain the translation does not influence the degree of intensity shift.

4.3 Emotion intensity shift and translation quality

Finally, we checked whether the degree of emotion intensity shift in translation correlates with the global translation quality attributes studied, namely, accuracy and fluency. Because both metrics are ordinal, we first computed Spearman rank correlations. Intensity shift exhibits moderately strong positive correlations with both accuracy ($\rho = 0.514$, $p < 0.001$) and fluency ($\rho = 0.481$, $p < 0.001$), which suggests that translations that preserve more emotional intensity also tend to receive higher quality ratings. However, the association is moderate rather than deterministic: attenuation is observed even in translations judged as reasonably accurate and fluent. This pattern, robust across systems, differs from previous findings (Fukuda and Jin, 2022; Saadany et al., 2023), which emphasize a stronger independence between emotion preservation and general translation quality. One possible explanation is that differences in overall system quality may partly account for this relationship, motivating further investigation into the interplay between emotion intensity preservation and global quality metrics.



Spanish Emotion Level	Itzuli	chatGPT	Latxa
1	0.00 (n=11)	0.09 (n=23)	0.05 (n=21)
2	-0.20 (n=51)	-0.25 (n=36)	-0.28 (n=43)
3	-0.67 (n=67)	-0.54 (n=67)	-0.39 (n=61)
4	-1.00 (n=126)	-1.05 (n=131)	-1.01 (n=134)
5	-1.13 (n=79)	-1.34 (n=77)	-1.19 (n=75)

Figure 4: Degree of intensity shift (Basque–Spanish) across Spanish levels.

5 High emotion intensity tweet translation

Given that translations of more emotionally intense tweets appear to be particularly challenging, and noting that our initial experiments did not fully exploit the capabilities of LLMs, we conducted a targeted follow-up evaluation.

We selected reaction tweets with high emotion intensity (scores of 4 or 5) whose Basque translations had previously received low intensity ratings ($=< 3$). This procedure resulted in a subset of 86 reaction tweets. These tweets were retranslated into Basque using Latxa under three different prompting conditions:

Prompt 1: A prompt written in Basque that assigns the LLM the role of a professional translator, explicitly describes the communicative context and stylistic characteristics of tweets, and requests high-quality translations with respect to accuracy, fluency, and style.

Prompt 2: An extension of the previous prompt with five real, in-context examples of Basque tweets, corresponding to a few-shot setting.

Prompt 3: A variation of the second prompt with the temperature increased from 0.7 to 1.2 (Top-P at 0.9), allowing greater generative flexibility.

We then conducted a crowd-based evaluation. Using the same platform as before, volunteers were again asked to assess the accuracy and fluency of the Basque translations, as well as the emotion intensity of both the Spanish source tweets and their Basque counterparts.

We collected a total of 499 evaluation responses covering the 86 reaction tweets (258 translations), with at least one assessment per prompt for each reaction tweet. Regarding participant proficiency, 84.37% of responses were provided by native Basque speakers, and 94.56% by evaluators reporting a C1 level or higher. Only 11 responses (2.20%) came from non-native evaluators without a reported C1 level. As before, duplicate evaluations were resolved by applying the procedure for distinguishing specialist and non-specialist evaluators and merging multiple assessments accordingly.

We begin by examining global quality metrics. As expected, all three prompting strategies outperform the baseline (from now on Prompt 0) on this high-intensity subset, where scores fail to reach the scale midpoint (accuracy: M = 3.94, SD = 1.95; fluency: M = 3.47, SD = 1.98). Mean ratings indicate that all new prompting strategies achieve above mid-range quality for both attributes; however, the remaining gap of more than two points to the top of the scale suggests considerable room for improvement (Figure 5). The few-shot prompt (Prompt 2) yields the highest mean scores for both accuracy (M = 4.61, SD = 1.98) and fluency (M = 4.84, SD = 1.93), followed closely by the description-rich prompt (Prompt 1; accuracy: M = 4.56, SD = 1.98; fluency: M = 4.68, SD = 1.96). Increasing the decoding temperature (Prompt 3) results in slightly lower scores (accuracy: M = 4.35, SD = 2.02; fluency: M = 4.55, SD = 2.00).

No significant differences across systems are observed for accuracy (ANOVA: $F = 2.08$, $p = .102$; Kruskal–Wallis: $H = 6.40$, $p = .094$). In contrast, fluency differs significantly across systems (ANOVA: $F = 8.52$, $p < .001$; Kruskal–Wallis: $H = 23.25$, $p < .001$). Tukey's HSD confirms that the baseline prompt (prompt0) scores significantly lower than all enhanced prompting conditions (prompt1–prompt3; all $p < .01$), while no significant differences emerge among the enhanced prompts (all $p > .75$).

We next focus on emotion intensity preservation. Mean Spanish intensity scores range from approximately 3.8 to 4.1 in a 1-5 scale, clustering around the high-section of the scale. This shows slightly more variation in assigning an emotional intensity to a rich-emotion tweet. Basque scores are substantially lower across prompts.

Across all prompting conditions, emotion shift values are consistently negative, indicating a systematic attenuation of emotional intensity in Basque relative to Spanish. The baseline prompt (Prompt 0) shows the largest loss (M = -1.85), while all three enhanced prompting strategies substantially reduce the magnitude of this shift. The few-shot prompt (Prompt 2) achieves the smallest average loss (M = -1.11), corresponding to an improvement of approximately 0.6 points in emotional preservation. Nevertheless, even under the best-performing prompt, emotion intensity is reduced by more than one scale point on average, suggesting that effective emotion preservation may require more complex prompting or fine-tuning approaches that add knowledge or better access particular language characteristics.

To test whether intensity-shift differences are statistically meaningful, we applied complementary significance tests. A one-way ANOVA showed a

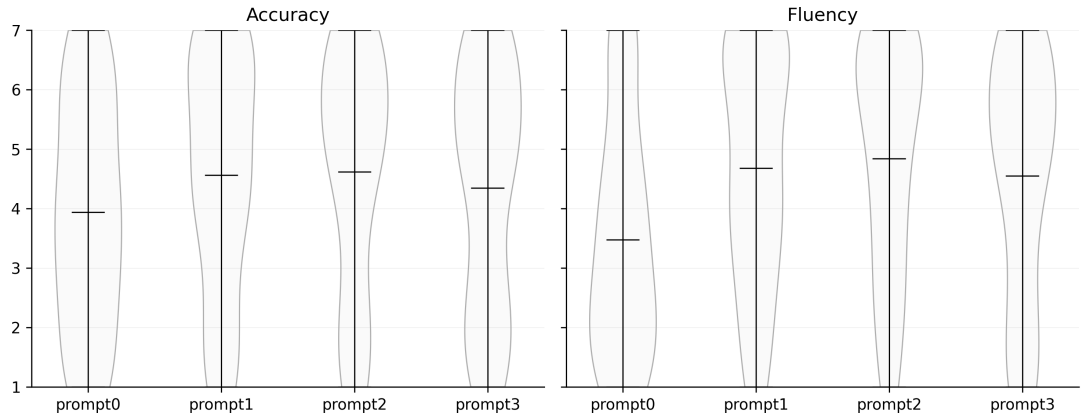


Figure 5: Accuracy and Fluency score distributions per prompt. Horizontal lines indicate median values.

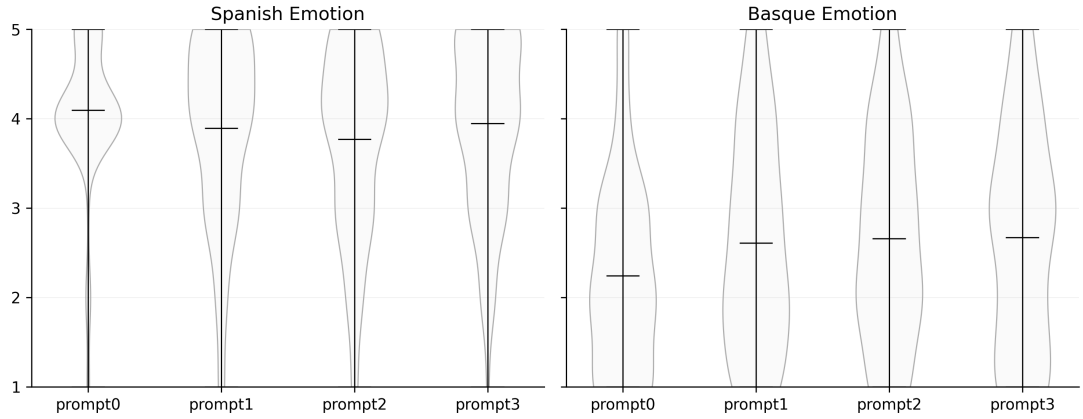


Figure 6: Spanish and Basque emotion intensity score distributions per prompt. Horizontal lines indicate median values.

significant effect of prompting strategy ($F = 4.69$, $p = .001$), which was confirmed by a Kruskal–Wallis test ($H = 18.45$, $p = .001$). Post-hoc Tukey HSD comparisons revealed that the baseline prompt (Prompt 0) differs significantly from all enhanced prompts (Prompt 1–Prompt 3; all $p < .05$), while no significant differences were found among the enhanced prompts themselves (all $p > .80$).

6 Conclusions

We studied the MT of emotion-rich tweets from Spanish into Basque, focusing on translation quality and preservation of emotion intensity. Overall, accuracy and fluency scores lie above the scale midpoint, indicating reasonably good performance, though with substantial room for improvement; for high intensity tweets, however, overall quality falls below mid-range levels, highlighting the difficulty of this subset. We observe no significant quality differences between a traditional NMT system and LLM-based models in baseline settings without task-specific instructions. Across systems, emotion intensity is systematically attenuated in translation, with an average loss of about one point on a 1–5 scale that increases with source emotion intensity. Contrary to previous work, quality attributes seem to correlate with intensity shift, possibly reflecting differences in overall system quality. Finally, basic LLM prompting strategies—contextual instructions, few-shot examples, and increased temperature—yield only modest reductions in intensity loss for highly emotional tweets, leaving intensity preservation an open challenge in Spanish–Basque MT.

Several limitations should be acknowledged. The dataset is small, limiting generalizability, and the evaluation relies on crowd-based judgments from participants with mixed language backgrounds. Although prior work shows that evaluators with linguistic or translation training tend to be stricter and expertise was considered during data cleaning, its effect was not systematically

tested. Moreover, most translations received only a single evaluation, and judgments of quality and emotion intensity are inherently subjective, adding variability. Future work would benefit from more controlled designs that balance emotion intensity and cover different emotion types, and from examining how the development and use of informal Basque—both in MT training data and in speaker expectations—relate to translation performance and evaluation. Despite these limitations, our results point to emotion intensity preservation as a meaningful challenge in Spanish–Basque tweet MT.

Acknowledgments

This work was partially supported by the MOLVI project (PID2024-157855OB-C32), funded by MICIU/AEI/10.13039/501100011033 and FEDER, EU, and by the project Desarrollo de Modelos ALIA, Resol. SEDIA 19.08.2024, within the framework of the National Language Technologies Plan (ENIA 2024), funded by MTDFP, PRTR, and the European Union–NextGenerationEU. The author thanks the volunteer participants for their contributions to the data creation and crowd-based evaluation.

References

Inaki Alegria, Nora Aranberri, Víctor Fresno, Pablo Gamallo, Lluis Padró, Inaki San Vicente, Jordi Turmo, and Arkaitz Zubiaga. 2013. Introducción a la tarea compartida tweet-norm 2013: Normalización léxica de tuits en español. In *Tweet-Norm@ SEPLN*, pages 1–9.

Javier J Amores, David Blanco-Herrero, Patricia Sánchez-Holgado, and Maximiliano Frías-Vázquez. 2021. Detectando el odio ideológico en twitter. desarrollo y evaluación de un detector de discurso de odio por ideología política en tuits en español. *Cuadernos. info*, (49):98–124.

Nora Aranberri. 2024. Analysis of the annotations from a crowd MT evaluation initiative: Case study for the Spanish-Basque pair. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 548–559, Sheffield, UK. European Association for Machine Translation (EAMT).

Nora Aranberri, Gorka Labaka, Arantza Díaz de Ilarraza, and Kepa Sarasola. 2017. Ebaluatoia: crowd evaluation for english–basque machine translation. *Language Resources and Evaluation*, 51(4):1053–1084.

Duygu Ataman, Alexandra Birch, Nizar Habash, Marcello Federico, Philipp Koehn, and Kyunghyun Cho. 2025. Machine translation in the era of large language models:a survey of historical and emerging problems. *Information*, 16(9).

David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. 2014. Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2):135–160.

Francesco Barbieri, Jose Camacho-Collados, Leonardo Neves, and Luis Espinosa-Anke. 2020. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1644–1650.

Francesco Barbieri and Horacio Saggion. 2014. Modelling irony in twitter. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 56–64. Association for Computational Linguistics.

Jeremy Barnes, Toni Badia, and Patrik Lambert. 2018. MultiBooked: A corpus of Basque and Catalan hotel reviews annotated for aspect-level sentiment classification. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Eleftheria Briakou and Marine Carpuat. 2021. A case for evaluating emotion in machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL/IJCNLP)*, pages 2855–2869.

Laura Camargo Fernández. 2021. El nuevo orden discursivo de la extrema derecha española: de la deshumanización a los bulos en un corpus de tuits de vox sobre la inmigración. *Culture, Language and Representation, 2021, vol. 26, p. 63-82*.

Aaron Chatterji, Thomas Cunningham, David J Deming, Zoe Hitzig, Christopher Ong, Carl Yan Shan, and Kevin Wadman. 2025. How people use chatgpt. Working Paper 34255, National Bureau of Economic Research.

Jacob Eisenstein. 2013. What to do about bad language on the internet. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 359–369. Association for Computational Linguistics.

Agurtzane Elordui. 2025. Heteroglossic management in instagram: Emerging ideological dynamics among basque youth. *Journal of Linguistic Anthropology*, 35(1):1–23.

Agurtzane Elordui, Jokin Aiestaran, Garbiñe Bereziartua, Irantzu Epelde, Orreaga Ibarra, Oroitz Jauregi, Libe Mimenza, Beñat Muguruza, and Ane Odria. 2020. Gaztesare corpus and data-base.

Thierry Etchegoyhen, Lindsay Bywood, Mark Fishel, Panayota Georgakopoulou, and 1 others. 2018. Mt for under-resourced languages: The case of basque. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 3761–3766.

Oded Feldman, Dan Ariely, and Erez Shmueli. 2018. Characterizing the linguistic style of conversational tweets. In *Proceedings of the Twelfth International AAAI Conference on Web and Social Media*, pages 100–109. AAAI.

Karin Fukuda and Qun Jin. 2022. Analyzing change on emotion scores of tweets before and after machine translation. In *Social Computing and Social Media: Design, User Experience and Impact*, pages 294–308, Cham. Springer International Publishing.

Maite Heredia, Jeremy Barnes, and Aitor Soroa. 2025. EuskañolDS: A naturally sourced corpus for Basque-Spanish code-switching. In *Proceedings of the 7th Workshop on Computational Approaches to Linguistic Code-Switching*, pages 1–5, Albuquerque, New Mexico, USA. Association for Computational Linguistics.

Courtenay Honeycutt and Susan C. Herring. 2009. Beyond microblogging: Conversation and collaboration via twitter. In *Proceedings of the 42nd Hawaii International Conference on System Sciences*, pages 1–10. IEEE.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 6282–6293.

Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica Lundin, Christof Monz, Kenton Murray, and 10 others. 2025. Findings of the WMT25 general machine translation shared task: Time to stop evaluating on easy test sets. In *Proceedings of the Tenth Conference on Machine Translation*, pages 355–413, Suzhou, China. Association for Computational Linguistics.

C Pérez, A Rebollar, and M Pérez. 2019. Construcción de un corpus lingüístico a partir de tweets tomando como base un tema específico. *Jornada de Ciencia y Tecnología Aplicada Tecnológico Nacional de México/CENIDET*, 2:15–19.

Maja Popovic, Ekaterina Lapshinova-Koltunski, and Maarit Koponen. 2024. Effects of different types of noise in user-generated reviews on human and machine translations including ChatGPT. In *Proceedings of the Ninth Workshop on Noisy and User-generated Text (W-NUT 2024)*, pages 17–30, San Ġiljan, Malta. Association for Computational Linguistics.

Shenbin Qian, Constantin Orasan, Felix Do Carmo, Qiuliang Li, and Diptesh Kanojia. 2023. Evaluation of Chinese-English machine translation of emotion-loaded microblog texts: A human annotated dataset for the quality assessment of emotion translation. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 125–135, Tampere, Finland. European Association for Machine Translation.

Eneko Romero, Haritz González, and Arantza Díaz de Ilarraza. 2022. The basque parliament corpus: Construction and applications for low-resource mt. In *Proceedings of the Thirteenth International Conference on Language Resources and Evaluation (LREC 2022)*, pages 1125–1133.

Hadeel Saadany, Constantin Orasan, Rocio Caro Quintana, Felix Do Carmo, and Leonardo Zilio. 2023. Analysing mistranslation of emotions in multilingual tweets by online MT tools. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 275–284, Tampere, Finland. European Association for Machine Translation.

Oscar Sainz, Naiara Perez, Julen Etxaniz, Joseba Fernandez de Landa, Itziar Aldabe, Iker García-Ferrero, Aimar Zabala, Ekhi Azurmendi, German Rigau, Eneko Agirre, Mikel Artetxe, and Aitor Soroa. 2025. Instructing large language models for low-resource languages: A systematic study for Basque. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29124–29148, Suzhou, China. Association for Computational Linguistics.

Michele Zappavigna. 2012. *Discourse of Twitter and Social Media: How We Use Language to Create Affiliation on the Web*. Continuum, London.

A Prompt wording

- **Prompt 0:**

  Itzuli hurrengo tuitak euskarara.

- **Prompt 1:**

  Itzultzaile profesionala zara. Eskatu dizute sare sozial bateko erreakzio txioak itzultzeko gaztelaniatik euskarara. Itzulpenek gaztelaniako erreakzio txioen esanahia islatu behar dute eta hizkuntza jatorra erabili. Txioek hizkera informala erabili ohi dute, ironia eta sarkasmoa ere bai, intentsitate emozional handikoak ere izan ohi dira, eta erreferentzia kulturalak izaten dituzte. Horregatik, garrantzitsua da itzulpenek gaztelaniako txioen estiloa mantentzea.

Hemen duzu itzuli beharreko erreakzio txioen zerrenda. Erreakzio-txio bakoitzaren aurretik hasierako txioa ageri da testuinguru gisa erabiltzeko. Itzuli erreakzio txioa bakarrik. Itzulpenak bakarrik erakutsi.

- **Prompt 2:** (Promt 1 + examples)

  Hona hemen erreakzio txioen adibideak:

  - Buah zelako temazoa txo
  - pelikula bateko gidoia izan liteke baina esukal herriko jaixetan danetarik gerta liteke, aupa gaztediiiiii!!
  - Pitxartxarreraz dakidan apurraren arabera, "Ospa hemendik, zomorroide, txitatzen ari gaituk eta!”
  - Tibet? etxe ondoan bertan! ai eneeeeeee! ez joan hain urrun ...
  - Ez da Simón edo Agirrezabala... baia hau be ez da atezain txarra gero!