Sentiment Analysis as a Quality Assurance Tool in Translator Training: A Pedagogical Case Study

Gabriel Cabrera-Méndez
gabriel.cabrera@uah.es
Universidad de Alcalá, Spain
Raquel Lázaro-Gutiérrez
raquel.lazaro@uah.es
Universidad de Alcalá, Spain

Cabrera-Méndez, Gabriel & Lázaro-Gutiérrez, Raquel (2025). Sentiment Analysis as a Quality Assurance Tool in Translator Training: A Pedagogical Case Study. *Language Value*, 18(1), 1-17. Universitat Jaume I ePress: Castelló, Spain. Http://www.languagevalue.uji.es.
July 2025
DOI: https://www.doi.org/10.6035/languagev.8811
ISSN 1989-7103

ABSTRACT

This paper presents a pedagogical case study on the use of sentiment analysis as a quality assurance tool in translator training. Conducted at the University of Alcalá (Spain), the study involved 37 undergraduate students who analysed the sentiment of English source texts and their Spanish translations using an AI model. Working with institutional, journalistic, and political texts, students applied a three-step methodology: initial sentiment analysis of source texts, translation using CAT tools (excluding MT), and final sentiment analysis of their translations. Neutrality coefficients ranging from -1 to +1 were used to quantify sentiment shifts. Results suggest that sentiment analysis can complement traditional quality assessment methods, particularly for politically sensitive texts. Students found the approach pedagogically valuable. Despite limitations related to sample size and reliance on a single AI tool, the study supports incorporating sentiment analysis into translator education and potentially into professional workflows for ensuring emotional and pragmatic consistency.

Keywords: *Sentiment analysis; Translation quality assessment; Translator training; Neutrality coefficient; Artificial intelligence in translation.*

I. INTRODUCTION

The integration of artificial intelligence into translation studies has opened up new avenues for both research and pedagogy. Among the most promising developments is the use of sentiment analysis, a technique originating in the field of Natural Language Processing (NLP), which enables the automatic evaluation of a text's emotional valence, typically classifying it as positive, negative, or neutral. While sentiment analysis is commonly applied in areas such as marketing or political discourse analysis, its potential to inform translation quality assessment remains underexplored, particularly in educational contexts.

This paper presents the results of a pedagogical experience conducted in a university course on Computer-Assisted Translation Tools at the University of Alcalá (Spain), in which undergraduate students explored the potential of AI-based sentiment analysis as a quality assurance resource in translator training. The study did not seek to determine whether translations consistently preserve the sentiment of the source texts, but rather to investigate how sentiment analysis might be used pedagogically to raise students' awareness of tonal and emotional fidelity. By focusing on the neutrality coefficient —a numerical scale ranging from -1 (strongly negative) to +1 (strongly positive), with 0 indicating perfect neutrality— students engaged with translation quality through a quantifiable and replicable lens.

The challenge of evaluating translation quality remains a central concern in both professional and academic contexts. Traditional methods often prioritise lexical accuracy or syntactic fluency, yet they tend to overlook pragmatic and affective dimensions such as tone, neutrality, or emotional consistency, elements that are particularly relevant in politically sensitive or institutional texts. Although various metrics exist to assess surface-level correspondence, few tools offer insights into how meaning and sentiment are transferred across languages.

Recent advances in NLP have made sentiment analysis more accessible, and although some studies have applied this technique to machine translation evaluation, its use in human translator training remains limited. The present study seeks to address this gap by proposing a didactic framework in which students use sentiment analysis tools to assess the emotional consistency of their translations. In doing so, it contributes to the growing body of research advocating for the integration of computational tools into translator education, while also fostering a deeper awareness of how affective meaning may shift during the translation process.

II. LITERATURE REVIEW

II.1. Early Developments in Sentiment Analysis and Translation

The emergence of sentiment analysis as a subfield of Natural Language Processing (NLP) began in the early 2000s, with a particular focus on the computational detection of opinions and emotions in texts (Pang & Lee, 2008; Liu, 2012). Early sentiment classifiers relied heavily on lexicon-based approaches, which later evolved to incorporate machine learning and, more recently, neural network models.

While sentiment analysis was initially developed for monolingual texts, the intersection with translation studies surfaced when researchers began to investigate how affective meaning is preserved —or altered— through translation (Nida, 1964; House, 1997). Initial contributions in this space suggested that translation often entails shifts in sentiment, whether due to cultural constraints, lexical asymmetry, or the translator's own stance (Baker, 2006; Munday, 2012). These affective shifts were difficult to quantify until sentiment analysis tools became more accessible.

II.2. Quality Evaluation in Machine Translation and Human Translation

As machine translation (MT) systems evolved, so did efforts to evaluate their output. Metrics such as BLEU (Papineni et al., 2002), METEOR (Banerjee & Lavie, 2005), and TER (Snover et al., 2006) became standard tools. However, these metrics primarily assess surface-level features such as lexical overlap or edit distance, overlooking aspects such as semantic equivalence and pragmatic fidelity, including sentiment consistency.

Neural Machine Translation (NMT) marked a significant shift in the field, yielding more fluent and context-aware translations (Bahdanau et al., 2015). Nevertheless, studies showed that even NMT systems were susceptible to sentiment shifts, particularly when translating emotionally charged content (Castilho et al., 2017). These findings highlighted the need for evaluation tools that go beyond syntax and lexis to capture deeper layers of meaning.

II.3. Sentiment Preservation in Translation Studies

The last decade has seen increasing scholarly interest in sentiment preservation as an indicator of translation quality. In their systematic review, Han, Smeaton, and Jones (2021) underscore how traditional quality assessment overlooks emotional fidelity and call for more comprehensive approaches integrating sentiment-aware metrics. Similarly, Rivera-Trigueros (2022) identifies sentiment preservation as a key gap in MT evaluation procedures and proposes integrating sentiment analysis tools into the assessment pipeline.

Notably, Saadany et al. (2021) introduced the Sentiment-Aware Measure (SAM), a metric designed to compare sentiment polarity between source and target texts. Their findings confirmed that sentiment transfer is not always guaranteed, even when translations are grammatically and semantically accurate. This aligns with earlier theoretical claims by Munday (2012), who argued that affect is one of the most vulnerable dimensions in translation due to its reliance on cultural and emotional resonance.

II.4. Sentiment Analysis as a Pedagogical Tool in Translator Training

A more recent trend has been the incorporation of sentiment analysis tools into translation pedagogy. This approach offers future translators a quantitative method to assess emotional consistency between the source and target texts, complementing traditional qualitative feedback (Biel, 2011; Angelelli & Baer, 2016). By leveraging sentiment analysis, students can become more aware of implicit tone, connotative meaning, and neutrality —elements that are crucial in institutional and political texts where objectivity is paramount.

Moreover, the work by Saadany et al. (2021) demonstrates that sentiment-aware evaluation metrics can provide valuable insight not only in assessing machine translation output but also in developing pedagogical applications. Their Sentiment-Aware Measure (SAM) quantifies discrepancies in emotional polarity between source and translated texts, showing how computational tools can help visualise and interpret affective variation. This capacity to make emotional shifts explicit empowers students to make more informed decisions regarding lexical choices and translation strategies.

II.5. Current Gaps and the Rationale for the Present Study

Despite these advancements, there remains a paucity of empirical research on how sentiment analysis can be used in classroom settings to assess translation neutrality. Most existing work either focuses on machine translation (Saadany et al., 2021) or offers theoretical reflections on affect and tone in translation without hands-on application (Munday, 2012). The present study aims to bridge this gap by reporting on a classroom-based experience where students used AI-powered sentiment analysis tools to compare neutrality scores in original texts and their own translations rendered by them using CAT tools, but not MT.

By focusing on the coefficient of neutrality —understood as the degree to which a text refrains from emotional polarity— this research contributes to current debates on translation quality and introduces a replicable pedagogical model.

II.6. Approaches and Tools for Sentiment Analysis

Over the years, various methodologies have been developed to perform sentiment analysis, ranging from early lexicon-based models to advanced machine learning and deep learning systems. The lexicon-based approach, relying on pre-defined lists of positive and negative words, constituted the earliest method for detecting sentiment in texts (Taboada et al., 2011). Although relatively simple, this approach struggled with contextual nuances such as irony, sarcasm, or negation.

The advent of machine learning introduced supervised classifiers such as Naïve Bayes, Support Vector Machines (SVMs), and Decision Trees, which significantly improved sentiment detection accuracy by learning from annotated corpora (Pang, Lee, &

Vaithyanathan, 2002). Later, the rise of deep learning techniques, particularly Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), allowed for more sophisticated sentiment analysis models capable of capturing complex linguistic patterns (Kim, 2014).

More recently, Transformer-based models such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) have become the standard in sentiment analysis tasks. These models, pre-trained on massive corpora and fine-tuned for sentiment classification, offer state-of-the-art performance by leveraging contextual embeddings and attention mechanisms.

Accessible platforms such as Hugging Face have democratized the use of advanced sentiment analysis models, offering APIs and pre-trained transformers ready for application in multiple languages and domains (Hugging Face, n.d.). The availability of these resources enables not only researchers but also practitioners in education and industry to perform high-quality sentiment analysis with minimal technical barriers.

In particular, Hugging Face’s "Sentiment Analysis" pipeline provides an off-the-shelf solution that classifies texts into positive, negative, or neutral categories based on fine-tuned transformer models. This ease of access has significantly contributed to the integration of sentiment analysis in diverse fields, including translation studies, journalism, and political communication (Hugging Face, n.d.).

These technological developments underpin the methodological choices in the present study, which leverages AI-powered sentiment analysis tools to explore shifts in neutrality across translated and original texts. Specifically, the students used OpenAI’s ChatGPT, a transformer-based large language model built on the GPT architecture, to perform sentiment analysis in both English and Spanish. The model provided sentiment classifications (positive, neutral, negative) along with numerical neutrality scores on a scale from -1 to +1. Although ChatGPT is not a sentiment-dedicated model like those fine-tuned via Hugging Face’s pipelines (e.g., Taboada et al., 2011; Liu, 2012), its performance was considered to be suitable for educational purposes due to its multilingual capabilities, ease of access, and ability to justify outputs qualitatively. The choice of ChatGPT was motivated by pedagogical considerations, prioritising usability and student engagement over technical optimisation.

III. RESULTS AND DISCUSSION

III.1 Methodology

The present study was conducted within the framework of a practical classroom exercise in the subject "Computer-Assisted Translation Tools" at the University of Alcalá (Spain). The aim was to explore the potential of sentiment analysis as a complementary method for evaluating translation quality, specifically in relation to tonal fidelity and neutrality preservation.

Students worked with a curated selection of authentic English-language texts, encompassing three main categories: (1) institutional documents (e.g., government statements, official reports), (2) press releases from major international organisations, and (3) news articles from media outlets with distinct political leanings. These texts were chosen to represent a range of expected neutrality coefficients, from highly neutral to potentially polarised content.

The experimental procedure followed three successive stages:

Stage 1: Sentiment analysis of the original texts

Students used the artificial intelligence model ChatGPT (OpenAI), a transformer-based large language model, to perform sentiment analysis on the English source texts. For each text, they entered a prompt requesting both a qualitative sentiment classification (positive, neutral, or negative) and a numerical neutrality coefficient on a scale from -1 (extremely negative) to +1 (extremely positive), with 0 indicating perfect neutrality. This prompted the model to provide a combined output including both a polarity label and a brief rationale explaining the score. To ensure consistency, students were given a standardised template prompt to use across all texts. The outputs were then recorded in a comparative table and used as a baseline for subsequent analysis of their Spanish translations. ChatGPT was selected for its accessibility, multilingual capacity, and suitability for classroom environments. Its ability to justify sentiment scores qualitatively encouraged students to engage in critical reflection on the pragmatic tone of both source and target texts. This approach is in line with recent calls to explore the integration of automatic quality assessment tools in translator training settings (Han, Smeaton, & Jones, 2021).

Stage 2: Translation into Spanish using CAT tools

Students then translated the original English texts into Spanish, utilising theComputer-Assisted Translation (CAT) tool Wordfast Anywhere. It is important to note that students were explicitly instructed not to employ Machine Translation systems, ensuring that the translation process was human-driven while still assisted by translation memory and terminology management features.

Stage 3: Sentiment analysis of the translated texts

Finally, the students applied the same AI-based sentiment analysis model to their Spanish translations. The neutrality coefficients and sentiment categories were recorded and compared to those obtained for the original texts.

The primary research question guiding the exercise was whether the sentiment, particularly in terms of neutrality, would be preserved during the translation process. This approach aligns with recent scholarship advocating for the integration of computational tools in translation studies to assess quality parameters beyond mere lexical fidelity (Läubli et al., 2020; Ribeiro, 2017).

To facilitate understanding, the workflow of the experiment is illustrated in Diagram 1 below:

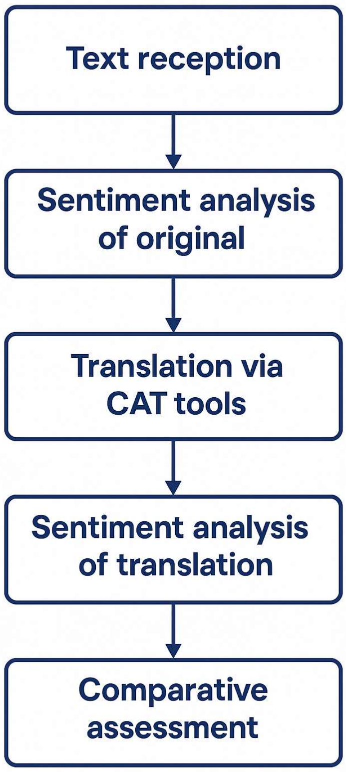

Figure 1. *Diagram showing the workflow during this task in the classroom.*

This methodology allowed for an objective, replicable, and relatively low-cost approach to assessing tonal shifts during translation, offering a promising avenue for further research and pedagogical innovation in translator training programmes (Han, Smeaton & Jones (2021).

III.2 Participant Profile

The classroom activity involved a cohort of 37 undergraduate students enrolled in the course on Computer-Assisted Translation Tools at the University of Alcalá (Spain). The participants were aged between 20 and 23 years, with a notable gender imbalance: 81% were women and 19% were men. Out of the 37 students, 28 completed and submitted the full set of tasks.

Although the examples presented here are illustrative of common tendencies, the discussion of sentiment alignment should be interpreted within the pedagogical framework of the study. Rather than aiming for a definitive assessment of translation quality, the activity sought to raise students' awareness of emotional consistency and tone management in translated texts. Based on in-class observation and informal feedback, students reported finding the exercise novel, engaging and intellectually stimulating. Many noted that the neutrality scores prompted them to reflect on translation choices they might otherwise have considered unproblematic, particularly when dealing with politically sensitive language. While no formal post-task questionnaire was administered, the high level of classroom participation and spontaneous commentary suggest that the method helped to develop a critical approach to pragmatic equivalence. A more structured evaluation of student perception —e.g. via reflective reports or surveys— would be beneficial in future iterations of the activity.

III.3. Results

The sentiment analysis exercise produced interesting insights into the extent to which emotional neutrality was preserved in translation. The students worked on a series of 5 texts that they all had to translate and analyse, including institutional press releases, political news articles, and transcripts of press conferences. Each original English text was first analysed to obtain a neutrality score, and then the students translated the texts into Spanish using CAT tools (without the aid of machine translation), after which the sentiment analysis was reapplied to the translations.

The following section presents selected examples that reflect the tendencies observed during the classroom exercise. These examples are interpreted in light of the study's pedagogical objective: to examine how sentiment analysis can be used to raise students' awareness of tonal and emotional consistency in translation. Rather than testing a linguistic hypothesis in the strict sense, the aim is to assess whether AI-generated sentiment scores can serve as useful indicators for student self-evaluation and classroom reflection.

A selection of representative examples illustrates the patterns observed:

Institutional Press Releases

Source text in English: "The government remains committed to ensuring the safety and prosperity of all its citizens through comprehensive policy measures."

The neutrality score for the original text was 0.05, indicating near-complete neutrality.

The Spanish translation: "El gobierno sigue comprometido a garantizar la seguridad y la prosperidad de todos sus ciudadanos mediante medidas de política integral."

In this case, the neutrality score of the translated text was 0.03, indicating a minimal and acceptable shift. For the purposes of this classroom activity, variations within a ±0.10 range were considered acceptable, as they typically reflect only minor tonal differences unlikely to alter the perceived intent or sentiment of the source text. This operational threshold was not intended as a rigid standard, but rather as a pedagogical guideline to help students distinguish between significant and negligible shifts in emotional tone.

Political News Articles

Source text in English: "The administration's failure to address the crisis has left countless families without essential support."

This text had a negative sentiment score of -0.65.

Spanish translation: "La falta de acción del gobierno ante la crisis ha dejado a innumerables familias sin apoyo esencial."

The sentiment score adjusted to -0.45, signalling a slight mitigation of negativity.

Press Conference Transcripts

Source English text: "We are taking every possible step to stabilise the situation and provide immediate assistance."

The English source text received an original neutrality score of 0.10, indicating a slightly positive but largely neutral tone.

Spanish translation: "Estamos tomando todas las medidas posibles para estabilizar la situación y ofrecer asistencia inmediata."

The sentiment analysis achieved a neutrality score of 0.08, preserving the intended tone with only a negligible shift.

The overall results are summarised in the following tables:

Table 1. *Summary of Sentiment Shift by Text Type*

Text Type	Average Neutrality Shift	General Observations
Institutional Press Releases	±0.02	High fidelity to original neutrality.
Political News Articles	±0.15	Moderate shifts, often reducing sentiment extremity.
Press Conference Transcripts	±0.05	Minor shifts, with tone generally well preserved.

Table 2. *Sentiment Shift Outcomes Across Submissions*

Outcome	Number of Cases	Percentage
Faithful Preservation of Neutrality	17	61%
Minor but Acceptable Shift	8	29%
Significant Sentiment Alteration	3	10%

These results suggest that while a majority of students maintained the intended neutrality or sentiment of the original texts, the complexity and emotional charge of some source materials posed challenges. Translation of institutional texts demonstrated the highest levels of fidelity, while political news articles were more susceptible to sentiment shifts during translation.

Example of a Significant Sentiment Alteration

While the majority of student translations managed to preserve the overall sentiment of the original texts, a few instances displayed notable deviations, with shifts large enough to alter the perceived emotional tone of the message. One such example came from a politically charged opinion piece.

Source text in English: "While the reforms present opportunities, critics warn that the changes may disproportionately affect low-income families."

This sentence received a neutrality score of -0.10, indicating a very slightly negative tone due to the cautionary element but still largely balanced.

Spanish translation: "Aunque las reformas traen oportunidades, los críticos aseguran que estas decisiones perjudicarán seriamente a las familias pobres."

This version yielded a sentiment score of -0.48, reflecting a strongly negative tone. The phrase "perjudicarán seriamente a las familias pobres" introduces a much more dramatic and emotionally loaded construction than the original "may disproportionately affect", which is hedged and more tentative.

This shift significantly altered the overall tone of the message. It reflects a common translation pitfall: intensifying the emotional impact, either intentionally or inadvertently, which may compromise the fidelity of the translation in contexts where neutrality is critical (e.g., journalism, diplomacy, or institutional communication).

Table 3. *Example of Significant Sentiment Drift*

Original Text Sentiment	Translated Text Sentiment	Shift Value	Commentary
-0.10	0.48	-0.38	Over-intensification of the critical tone

This and other similar examples reinforce the potential of sentiment analysis as a training tool for identifying areas where students may unconsciously distort meaning, particularly in sensitive or politically nuanced texts.

IV. CONCLUSIONS

The pedagogical experience outlined in this study illustrates the innovative potential of integrating sentiment analysis into translator training. Students responded enthusiastically to the exercise, appreciating its practical relevance and the immediate feedback it provided on their translation choices.

The results suggest that sentiment analysis could serve as a valuable complement to traditional quality assurance practices within Language Service Providers (LSPs), offering an objective means of assessing emotional consistency alongside linguistic accuracy. Consequently, this technique will be permanently incorporated into the Computer-Assisted Translation Tools curriculum at the University of Alcalá.

Notably, the exercise revealed that political texts are particularly susceptible to shifts in neutrality during translation, more so than institutional or press conference materials. This finding underlines the critical need for heightened sensitivity when translating ideologically charged content.

However, several limitations must be acknowledged. The study was conducted with a single cohort of undergraduate students, and the selection of texts, while varied, was not exhaustive. Additionally, sentiment analysis results depend on the capabilities and training of the AI model used, which may introduce variability. Future research should address these limitations by expanding the corpus, involving more diverse participant groups, and testing different sentiment analysis models.

Despite these constraints, the novelty of applying neutrality coefficients as a training tool marks a significant contribution to translator education. With further practice to refine both methodological rigour and student familiarity, this approach could be successfully transferred to professional translation workflows, enriching quality assurance protocols and fostering greater pragmatic and affective fidelity in cross-linguistic communication.

In response to the pedagogical research question guiding this study, the results indicate that sentiment analysis can serve as a practical and reflective tool for raising students' awareness of emotional fidelity in translation and for supporting quality assurance practices in translator education.

REFERENCES

Angelelli, Claudia V., & Baer, Brian James (Eds.). (2016). Researching Translation and Interpreting. Routledge.

Bahdanau, Dzmitry, Cho, Kyunghyun, & Bengio, Yoshua. (2015). Neural machine translation by jointly learning to align and translate. International Conference on Learning Representations (ICLR). https://arxiv.org/abs/1409.0473

Baker, Mona. (2006). Translation and Conflict: A Narrative Account. Routledge. https://doi.org/10.4324/9780429438240

Banerjee, Satanjeev, & Lavie, Alon. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization (pp. 65–72). https://aclanthology.org/W05-0909/

Biel, Łucja. (2011). Training translators or translation service providers? EN 15038:2006 and its training implications. The Journal of Specialised Translation, 16, 61–76. https://jostrans.org/issue16/art_biel.pdf

Castilho, Sheila, Gaspari, Federico, Moorkens, Joss, & Way, Andy. (2017). Is neural machine translation the new state of the art? The Prague Bulletin of Mathematical Linguistics, 108(1), 109–120. https://doi.org/10.1515/pralin-2017-0013

Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton, & Toutanova, Kristina. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT (pp. 4171–4186). https://aclanthology.org/N19-1423/

Han, Liting, Smeaton, Alan F., & Jones, Gareth J. F. (2021). Translation quality assessment: A brief survey on manual and automatic methods. In Proceedings for the First Workshop on Modelling Translation: Translatology in the Digital Age (pp. 15–33). ACL. https://doi.org/10.48550/arXiv.2105.03311

House, Juliane. (1997). Translation Quality Assessment: A Model Revisited. Gunter Narr Verlag.

Hugging Face. (n.d.). Sentiment analysis. Hugging Face. Retrieved April 26, 2025, from https://huggingface.co/tabularisai/multilingual-sentiment-analysis

Kim, Yoon. (2014). Convolutional neural networks for sentence classification. In Proceedings of EMNLP 2014 (pp. 1746–1751). https://aclanthology.org/D14-1181/

Liu, Bing. (2012). Sentiment Analysis and Opinion Mining. Morgan & Claypool. https://doi.org/10.2200/S00416ED1V01Y201204HLT016

Liu, Yinhan, Ott, Myle, Goyal, Naman, Du, Jingfei, Joshi, Mandar, Chen, Danqi & Stoyanov, Veselin. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692. https://arxiv.org/abs/1907.11692

Munday, Jeremy. (2012). Evaluation in Translation: Critical Points of Translator Decision-Making. Routledge. https://doi.org/10.4324/9780203117743

Nida, Eugene A. (1964). Toward a Science of Translating: With Special Reference to Principles and Procedures Involved in Bible Translating. Brill.

Pang, Bo, Lee, Lillian, & Vaithyanathan, Shivakumar. (2002). Thumbs up? Sentiment classification using machine learning techniques. In Proceedings of EMNLP 2002 (pp. 79–86). https://aclanthology.org/W02-1011/

Pang, Bo, & Lee, Lillian. (2008). Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1–2), 1–135. https://doi.org/10.1561/1500000011

Papineni, Kishore, Roukos, Salim, Ward, Todd, & Zhu, Wei-Jing. (2002). BLEU: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (pp. 311–318). https://aclanthology.org/P02-1040/

Ribeiro, Fernando. (2017). On the role of machine translation output in translator training: A preliminary study. Journal of Specialised Translation, 28, 129–152. https://jostrans.org/issue28/art_ribeiro.php

Rivera-Trigueros, Irene. (2022). Machine translation systems and quality assessment: A systematic review. Language Resources and Evaluation, 56, 593–619. https://doi.org/10.1007/s10579-021-09537-5

Saadany, Hossam, Orasan, Constantin, Mohamed, Ehab, & Tantawy, Ahmed. (2021). Sentiment-aware measure (SAM) for evaluating sentiment transfer by machine translation systems. arXiv preprint arXiv:2109.14895. https://arxiv.org/abs/2109.14895

Snover, Matthew, Dorr, Bonnie, Schwartz, Richard, Micciulla, Linnea, & Makhoul, John. (2006). A study of translation edit rate with targeted human annotation. In Proceedings of AMTA 2006 (pp. 223–231). https://aclanthology.org/2006.amta-papers.25/

Taboada, Maite, Brooke, Julian, Tofiloski, Milena, Voll, Kimberly, & Stede, Manfred. (2011). Lexicon-based methods for sentiment analysis. Computational Linguistics, 37(2), 267–307. https://doi.org/10.1162/COLI_a_00049

Läubli, Samuel, Castilho, Sheila, Neubig, Graham, Sennrich, Rico, Shen, Qinlan., & Toral, Antonio. (2020). A set of recommendations for assessing human–machine parity in language translation. Journal of artificial intelligence research, 67, 653-672. https://doi.org/10.1613/jair.1.11371

Received: 15 May 2025

Accepted: 9 July 2025