Sentiment-Aware Machine Translation for Indic Languages

Amulya Ratna Dash, Harpreet Singh Anand and Yashvardhan Sharma
Birla Institute of Technology and Science, Pilani, Rajasthan, India
{*p20200105, f20212416, yash*}*@pilani.bits-pilani.ac.in*

Keywords: Natural Language Processing, Sentiment Analysis, Machine Translation, Low Resource Languages, Large Language Model.

Abstract: Machine Translation (MT) is a critical application in the field of Natural Language Processing (NLP) that aims to translate text from one language to another language. Indic languages, characterized by their linguistic diversity, often encapsulate emotional and sentimental expressions that are difficult to map accurately when translated from English. In-order to bridge the gap in language barrier, text(reviews) in English should be translated to multiple languages while preserving the sentiment. In this paper, we focus on the machine translation of English into three low resource Indic languages by employing sentiment-aware in-context learning techniques with large language models. Our approach helps improve the average translation score by +4.74 absolute points.

1 INTRODUCTION

The translation of emotions and sentiments into Indic languages presents a significant challenge in the field of Machine Translation (MT). Indian languages, characterized by their rich cultural and linguistic diversity, often encapsulate nuanced emotional expressions that are difficult to convey accurately in translations. Traditional MT systems frequently overlook these subtleties, leading to translations that may be technically correct but lack the emotional resonance of the original text. This inadequacy is particularly pronounced in low-resource languages, where the scarcity of sentiment lexicons further complicates the task of sentiment analysis(Malinga et al., 2024).

Moreover, the effectiveness of MT systems is heavily reliant on the availability of parallel corpora, which are often limited for Indic languages (Lample et al., 2018). As a result, the challenge of preserving sentiment during translation becomes even more critical, as it directly impacts the quality and usability of translated content in various applications, including social media and educational contexts (Saadany and Orašan, 2021). Recent advancements in Neural Machine Translation(NMT) have shown promise in addressing these issues by incorporating sentiment-aware mechanisms (Kumari et al., 2021; Si et al., 2019). However, there remains a pressing need for research focused specifically on Indic languages to enhance the accuracy and emotional fidelity of translations, thereby ensuring that the sentiments embedded in the source texts are preserved in the target languages.

In recent times, Large Language Models(LLM) have shown impressive performance in various NLP tasks like question answering, summarization, machine translation, etc. In this work, we propose a novel approach to preserve the sentiment of original source language text during the process of machine translation. We use In-Context Learning(ICL) techniques for sentiment-aware machine translation using LLMs.

The rest of the paper is organized as follows. Section 2 and 3 presents the background and review of related works. The methodology is briefly described in Section 4. Sections 5 and 6 detail the experiments, results and analysis, followed by the conclusion and future scope in Section 7.

2 BACKGROUND

In recent years, there has been a rise in the application of sentiment analysis to gain a deeper understanding of public opinion and sentiment towards various entities, products and events. Although much of the research in Sentiment Analysis has been focused mainly on the English language, there has been a growing interest in applying sentiment analysis techniques to other languages(Mabokela et al., 2022).

Dash, A. R., Anand, H. S. and Sharma, Y.
Sentiment-Aware Machine Translation for Indic Languages.
DOI: 10.5220/0013404000003890

In *Proceedings of the 17th International Conference on Agents and Artificial Intelligence (ICAART 2025) - Volume 1*, pages 948-953
ISBN: 978-989-758-737-5; ISSN: 2184-433X

The wide diversity of languages and dialects spoken across the country of India pose a key challenge in sentiment analysis for Indian languages (Basile et al., 2018). It has been shown that most of the research in sentiment analysis in Indian languages has focused on Hindi, followed by other languages such as Bengali, Tamil, Malayalam, Urdu, and Kannada. This is mostly due to the fact that there is still a lack of comprehensive datasets and resources for many of these languages, which has hindered the development of robust sentiment analysis models (Patra et al., 2015).

Researchers are exploring various approaches to sentiment analysis in Indian languages. For example, utilizing supervised learning techniques, such as machine learning algorithms, to classify text into positive, negative, or neutral sentiment (Kulkarni et al., 2021; Si et al., 2019) and leveraging lexical resources, such as SentiWordNet, to determine the sentiment of words and phrases (Rikters and Kāle, 2023; Shelke et al., 2022). Additionally, there has been a growing interest in the use of transformer-based models, such as BERT and IndicBERT, which have shown promising results in capturing the semantic nuances and linguistic patterns inherent in Indian languages (Kannan et al., 2021; Kumar and Albuquerque, 2021).

The prevalence of code-mixing i.e. switching between multiple languages within a single text also poses a challenge to the sentiment analysis for Indian languages. Several approaches have been explored to address this problem, such as developing specialized datasets and models for code-mixed text (Chakravarthi et al., 2022).

There have been significant advances in Neural Machine Translation (NMT) models achieving state-of-the-art performance (Luong et al., 2015). However, the translation of rare and unknown words still remains an open-vocabulary problem. Researchers have therefore explored the use of subword units to enable open-vocabulary translations(Sennrich et al., 2016). Most of the times, these rare words are important in conveying and transferring the sentiment and emotion of the original text to translated text. LLMs are trained on huge amount of data, which enables them to understand and generate these rare words in a better manner compared to specialized NMT models.

3 RELATED WORK

Sentiment preservation has emerged as a promising approach to improving accuracy of machine translation by incorporating emotional and attitudinal information. In Sentiment Analysis, many researchers use machine translation to analyze the sentiment of text in non-English language. The goal is to preserve sentiment information during the translation process, as sentiment can be crucial for understanding the meaning and nuance of text (Saadany and Orăsan, 2020).

Neural Machine Translation (NMT) systems often encounter challenges in accurately preserving sentiment when translating sentiment-ambiguous words due to their reliance on training data and contextual cues. (Si et al., 2019) introduced a sentiment-aware NMT framework to address this issue. By employing methods like sentiment labeling and valence-sensitive embeddings, the system successfully generated translations that reflected user-defined sentiment labels. Their results demonstrated superior performance in terms of BLEU scores and sentiment preservation compared to baseline models, especially for ambiguous lexical items where multiple sentiment-driven translations are possible.

Troiano et al., 2020 explored the challenges of emotion preservation in neural machine translation (NMT) through a back-translation setup. Their findings revealed that emotion nuances are often diluted or altered during translation. To mitigate this, they introduced a re-ranking approach using an emotion classifier, which improved the retention of affective content. Additionally, the framework enabled emotion style transfer, allowing translation outputs to reflect different emotional tones. Kumari et al., 2021 finetune a pre-trained NMT model to preserve sentiments using deep reinforcement learning and curriculum learning based techniques and evaluate using English-Hindi and French-English datasets.

Brazier and Rouas, 2024 proposed integrating emotion information into Large Language Models (LLMs) to improve translation quality. They fine-tuned LLMs on an English-to-French translation task, incorporating emotional dimensions such as arousal, dominance, and valence into the input prompts. The study demonstrated that including arousal information led to significant improvements in translation scores, showcasing how emotion-aware prompts can enhance semantic fidelity and expressiveness in translations.

Our work explores performance of LLMs for machine translation of Indic languages along with sentiment preservation using in-context learning techniques.

4 METHODOLOGY

4.1 Models

To investigate the sentiment-preserved translation capability of LLMs, we shortlisted LLaMA-3(Dubey et al., 2024) and Gemma-2(Team et al., 2024) as the preferred open-source LLM due to better multilingual ability. NLLB(Costa-jussà et al., 2022), a state-of-the-art encoder-decoder neural machine translation model provides translation capabilities for over 200 languages. We use pretrained LLaMA[1], Gemma[2] and NLLB[3] model for experiments. The prompt that yields a better response from both LLMs is shown in Figure 1.

4.2 Languages and Evaluation Data

To evaluate our approach, we consider 3 languages: Hindi, Odia, Punjabi. All three Indic languages belong to the Indo-Aryan branch of Indo-European language family. Hindi is a low - medium resource language, whereas Odia and Punjabi are very low resource languages. Hindi is spoken by more than 300 million speakers, and Odia/Punjabi is spoken by 50 million speakers in India.

To evaluate we use IndicSentiment(Doddapaneni et al., 2023) dataset which contains reviews from various categories and sub-categories in 13 Indic languages. Each review is labeled with sentiment polarity (neutral, negative, and positive).

4.3 Metrics

We use Character n-gram F-score(ChrF) (Popović, 2015) and COMET-22 (Rei et al., 2022) scores as evaluation metrics. ChrF is a F-score which balances precision and recall. ChrF calculates character-level overlap between the translation and reference texts. COMET is a learned metric to evaluate translations and is better aligned with human judgement. COMET considers the context and semantics of the source and translated text while predicting translation accuracy.

5 EXPERIMENTS AND RESULTS

We consider three translation tasks: *English* → *Hindi*, *English* → *Odia* and *English* → *Punjabi*. The test split (1000 records) of the IndicSentiment dataset is used for evaluation.

Seven experiments were designed per language: 1 using NLLB model, 3 using LLaMA-3.1 8B model and 3 using Gemma-2 2B model. In-Context Learning technique was explored in the LLaMA and Gemma experiments to preserve sentiments during translation tasks.

The evaluation scores for the three translation tasks with respect to different approaches are available in Table 1 (NLLB), Table 2 (LLaMA) and Table 3 (Gemma). Based on promising results from the small 2B Gemma-2 model, we experimented with the Gemma-2 9B model on *English* → *Hindi* translation task. The comparative translation performance for Gemma 2B and 9B are available in Table 4.

Table 1: Translation performance for NLLB-200 based experiment.

Language Pair	ChrF	COMET
English - Hindi	51.04	73.34
English - Odia	49.84	80.15
English - Punjabi	56.09	81.52

The response from Gemma model for the baseline experiment had additional information like 'Explanations', 'Important Notes' and 'Consideration' along with the translated text. As part of post-processing, an attempt was made to remove the additional information before calculating the evaluation metrics. Improvement in scores for the post-processed response was observed only for Odia translations as seen in Table 3 where the raw score is in parentheses.

6 ANALYSIS

- Our few-shot in-context learning approach yields **+ 4.74** Average COMET score compared to vanilla prompting approach for LLaMA-3.
- The COMET score for *English* → *Odia* translation improved the maximum (**+ 10.82**).
- Compared to state-of-the-art encoder-decoder(NLLB-200), the COMET score for *English* → *Hindi* improved by **+ 1.91** using LLM based approach.
- The encoder-decoder NMT model performed better for very low resource languages (Odia and Punjabi). This may be due to extremely low representation in LLaMA-3's training data as compared to English or Hindi.
- Sample inference examples are shown in Figure 3 and 2.

[1]https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

[2]https://huggingface.co/google/gemma-2-2b-it

[3]https://huggingface.co/facebook/nllb-200-distilled-600M

Role	Instruction and Response Template
Human	Translate the following review from English to <*Target_Language (Hindi/Odia/Punjabi)*>. The sentiment of the review is <*Sentiment (Positive/Negative/Neutral)*>. Provide only the translated text and ensure that the translation accurately conveys the sentiment without any justification or extra output. For example, if the English review is 'The service was terrible and I am very disappointed' (Negative), the translation should be 'सेवा बहुत खराब थी और मैं बहुत निराश हूं'. <*Text in Source language*>
Assistant (LLM)	<*Text in Target language*>

Figure 1: Prompt template used for sentiment-aware machine translation.

Table 2: Translation performance of LLaMa 3.1 based experiments.

	W/o Sentiment		**LLaMa 3.1 (0-Shot)**		**LLaMa 3.1 (1-Shot)**	
Language Pair	ChrF	COMET	ChrF	COMET	ChrF	COMET
English - Hindi	**48.27**	74.73	46.36	74.37	47.86	**75.25**
English - Odia	**23.90**	48.15	13.21	57.66	23.18	**58.97**
English - Punjabi	36.69	71.77	38.10	74.09	**39.18**	**74.64**
Average	36.28	64.88	32.55	68.70	**36.74**	**69.62**

Table 3: Translation performance of Gemma 2 (2B) based experiments.

	Baseline		**Sentiment Aware**	
Language Pair	ChrF	COMET	ChrF	COMET
English - Hindi	22.86	40.96	41.83	**71.59**
English - Odia	(28.57) 31.28	(33.79) 37.53	36.61	**40.76**
English - Punjabi	1.55	26.62	1.23	**31.65**
Average	18.56	35.12	26.55	48

Table 4: Comparison of Gemma 2 and LLaMA 3.1 on sentiment-aware *English* → *Hindi* translation.

Model	ChrF	COMET
Gemma 2 (2B)	41.83	71.59
Gemma 2 (9B)	**51.77**	**78.07**
LLaMa 3.1 (8B)	47.86	75.25

- In the case of Punjabi, the performance of the Gemma (2B) model exhibited notable shortcomings. A significant issue was the inappropriate use of scripts, where a substantial portion of translations was rendered in the Urdu script, rather than the Gurmukhi script, which is predominantly used in India. Similar observation was seen in the case of Odia, as some of the generated response was in Bengali script instead of Odia script. This deviation from the expected script posed a major challenge in achieving linguistically appropriate translations as reflected in the evaluation scores.
- Furthermore, Odia and Punjabi responses from Gemma (2B)contained redundant and repetitive tokens, resulting in translations that were spurious and lacked coherence. These anomalies were particularly evident more in the baseline translations. However, it is worth noting that these issues declined substantially in sentiment-aware translations. Despite this, the overall translation quality remained suboptimal, indicating room for improvement in both linguistic adequacy, fluency and contextual understanding for low resource Indic languages in LLMs.
- In the case Hindi, Gemma (9B) performed better than LLaMa (8B) and Gemma (2B) with difference of **+ 2.82** and **+ 6.48** in COMET scores respectively.

English **Sentiment**	Just a heck of a movie. Story looks like loosly knitted plot in a hurry to make a movie. **Negative**
Reference (Hindi):	बस एक बकवास फिल्म। कहानी एक फिल्म बनाने के लिए जल्दी में बनाया गया प्लॉट लग रहा है।
NLLB-200(Hindi):	कहानी एक फिल्म बनाने की जल्दी में ढीली बुना सा सा सा सा सा लगता है।
LLaMa-3.1 (Hindi):	एक बेहद बुरा फिल्म। कहानी दिखती है जैसे कि एक जल्दबाजी में बनाई गई फिल्म का ढीला बुना हुआ प्लॉट।

Figure 2: Sample sentiment preserved translation of English → Hindi.

English	After I watched Raya and the Last Dragon, this movie is just WOW!!! The character development of dragons is outstanding and yes, MORE WOW!
Sentiment	Positive
Reference(Punjabi)	"ਰਾਇਆ ਐਂਡ ਦ ਲਾਸਟ ਡ੍ਰੈਗਨ ਦੇਖਣ ਤੋਂ ਬਾਅਦ, ਇਹ ਫ਼ਿਲਮ ਬਹੁਤ ਵਧੀਆ ਲੱਗੀ!!! ਡਰੈਗਨ ਦੇ ਕਿਰਦਾਰ ਦਾ ਵਿਕਾਸ ਸ਼ਾਨਦਾਰ ਹੈ ਅਤੇ ਹਾਂ, ਵਾਕਈ ਬਹੁਤ ਵਧੀਆ!"
NLLB-200(Punjabi)	ਰਾਇਆ ਅਤੇ ਆਖਰੀ ਡ੍ਰੈਗਨ ਦੇਖਣ ਤੋਂ ਬਾਅਦ, ਇਹ ਫਿਲਮ ਸਿਰਫ ਵੌਅ ਹੈ!
LLaMa-3.1(Punjabi)	ਮੈਂ ਰਾਇਆ ਅਤੇ ਅੰਤਮ ਡਰੈਗਨ ਨੂੰ ਵੇਖਣ ਤੋਂ ਬਾਅਦ, ਇਹ ਫਿਲਮ ਬਹੁਤ ਹੀ ਆਉਂਦੀ ਹੈ!!! ਡਰੈਗਨਜ਼ ਦੇ ਕਿਰਦਾਰ ਦਾ ਵਿਕਾਸ ਬਹੁਤ ਚੰਗਾ ਹੈ ਅਤੇ ਹਾਂ, ਬਹੁਤ ਹੀ ਆਉਂਦਾ ਹੈ!!!

Figure 3: Sample sentiment preserved translation of English → Punjabi.

7 CONCLUSION AND FUTURE WORK

Every new release of large language models enhances and advances their multilingual capabilities. In-context Learning via prompts and Adapter based finetuning of LLMs have demonstrated exceptional results in addressing various NLP challenges, including information retrieval, sentiment analysis, and machine translation.

Our study concludes that using sentiment-aware in-context learning techniques with large language models helps to preserve the sentiment of original text and improve the overall accuracy of translations. Our approach can be generalized to other languages with very minimal changes, without any need of training or finetuning effort. In future, we would experiment with multiple open-source large language models and include additional languages. Also, we will finetune a LLM to improve the translation ability and further finetune the model to learn sentiment conditioning while translating reviews.

REFERENCES

Basile, P., Basile, V., Croce, D., and Polignano, M. (2018). Overview of the evalita 2018 aspect-based sentiment analysis task (absita). In *EVALITA@CLiC-it*.

Brazier, C. and Rouas, J.-L. (2024). Conditioning llms with emotion in neural machine translation. In *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pages 33–38.

Chakravarthi, B. R., Priyadharshini, R., Muralidaran, V., Jose, N., Suryawanshi, S., Sherly, E., and McCrae, J. P. (2022). Dravidiancodemix: Sentiment analysis and offensive language identification dataset for dravidian languages in code-mixed text. *Language Resources and Evaluation*, 56(3):765–806.

Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., et al. (2022). No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Doddapaneni, S., Aralikatte, R., Ramesh, G., Goyal, S., Khapra, M. M., Kunchukuttan, A., and Kumar, P. (2023). Towards leaving no indic language behind: Building monolingual corpora, benchmark and models for indic languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12402–12426.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle,

A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al. (2024). The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Kannan, R. R., Rajalakshmi, R., and Kumar, L. (2021). Indicbert based approach for sentiment analysis on code-mixed tamil tweets. In *FIRE (Working Notes)*, pages 729–736.

Kulkarni, A., Mandhane, M., Likhitkar, M., Kshirsagar, G., and Joshi, R. (2021). L3cubemahasent: A marathi tweet-based sentiment analysis dataset. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 213–220.

Kumar, A. and Albuquerque, V. H. C. (2021). Sentiment analysis using xlm-r transformer and zero-shot transfer learning on resource-poor indian language. *Transactions on Asian and Low-Resource Language Information Processing*, 20(5):1–13.

Kumari, D., Chennabasavaraj, S., Garera, N., and Ekbal, A. (2021). Sentiment preservation in review translation using curriculum-based re-inforcement framework. In *Proceedings of machine translation summit XVIII: Research track*, pages 150–162.

Lample, G., Ott, M., Conneau, A., Denoyer, L., and Ranzato, M. (2018). Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049.

Luong, T., Pham, H., and Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. In Màrquez, L., Callison-Burch, C., and Su, J., editors, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.

Mabokela, K. R., Celik, T., and Raborife, M. (2022). Multilingual sentiment analysis for under-resourced languages: a systematic review of the landscape. *IEEE Access*, 11:15996–16020.

Malinga, M., Lupanda, I., Nkongolo, M. W., and van Deventer, P. (2024). A multilingual sentiment lexicon for low-resource language translation using large languages models and explainable ai. *arXiv preprint arXiv:2411.04316*.

Patra, B. G., Das, D., Das, A., and Prasath, R. (2015). Shared task on sentiment analysis in indian languages (sail) tweets-an overview. In *Mining Intelligence and Knowledge Exploration: Third International Conference, MIKE 2015, Hyderabad, India, December 9-11, 2015, Proceedings 3*, pages 650–655. Springer.

Popović, M. (2015). chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.

Rei, R., De Souza, J. G., Alves, D., Zerva, C., Farinha, A. C., Glushkova, T., Lavie, A., Coheur, L., and Martins, A. F. (2022). Comet-22: Unbabel-ist 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585.

Rikters, M. and Kāle, M. (2023). The future of meat: Sentiment analysis of food tweets. In *Proceedings of the 11th International Workshop on Natural Language Processing for Social Media*, pages 38–46.

Saadany, H. and Orašan, C. (2021). Bleu, meteor, bertscore: Evaluation of metrics performance in assessing critical translation errors in sentiment-oriented text. In *Proceedings of the Translation and Interpreting Technology Online Conference*, pages 48–56.

Saadany, H. and Orăsan, C. (2020). Is it great or terrible? preserving sentiment in neural machine translation of arabic reviews. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 24–37.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In Erk, K. and Smith, N. A., editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Shelke, M. B., Alsubari, S. N., Panchal, D., and Deshmukh, S. N. (2022). Lexical resource creation and evaluation: sentiment analysis in marathi. In *Smart Trends in Computing and Communications: Proceedings of SmartCom 2022*, pages 187–195. Springer.

Si, C., Wu, K., Aw, A., and Kan, M.-Y. (2019). Sentiment aware neural machine translation. In *Proceedings of the 6th Workshop on Asian Translation*, pages 200–206.

Team, G., Riviere, M., Pathak, S., Sessa, P. G., Hardin, C., Bhupatiraju, S., Hussenot, L., Mesnard, T., Shahriari, B., Ramé, A., et al. (2024). Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Troiano, E., Klinger, R., and Padó, S. (2020). Lost in back-translation: Emotion preservation in neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4340–4354.