

ORIGINAL RESEARCH

Open access

Enhancing Low-Resource Language Sentiment Classification through Lexicon-Guided Fine-Tuning of Multilingual Language Models

Santiago Morales^{1*}, Laura Rojas², Camila Vega¹, Diego Molina³

Abstract

Pre-trained multilingual language models (PMLMs), including mBERT and XLM-R, have demonstrated strong capabilities in cross-lingual transfer. Nonetheless, these models are not explicitly optimized to capture cross-lingual indicators related to sentiment-bearing terms. This limitation disproportionately affects low-resource languages (LRLs), which suffer from under-representation within these architectures. To optimize the fine-tuning of these models for sentiment analysis in LRLs, this paper presents a novel intermediate task fine-tuning (ITFT) approach centered on a sentiment lexicon from a high-resource language (HRL). The authors evaluate this methodology using a 3-class sentiment classification framework across three LRLs: Sinhala, Tamil, and Bengali. The findings indicate that the proposed strategy surpasses standard vanilla fine-tuning of the PMLM, while also performing on par with or exceeding conventional ITFT techniques that utilize HRL sentiment classification datasets.

Keywords Sentiment analysis, Pre-trained Multilingual Language Models (PMLMs), Cross-lingual transfer, Low-Resource Languages (LRLs)

*Correspondence:

Santiago Morales

santiago.morales@fi.uba.ar

¹ Department of Natural Language Processing and Low-Resource Languages, Faculty of Engineering, University of Buenos Aires, Buenos Aires, Argentina

² Department of Lexicon-Guided Fine-Tuning, Faculty of Engineering, Pontifical Catholic University of Chile, Santiago, Chile

³ Department of Sentiment Analysis and Multilingual Models, Faculty of Engineering, National University of La Plata, La Plata, Argentina

Introduction

Pre-trained multilingual language models (PMLMs) have yielded highly encouraging outcomes in text classification, particularly within resource-constrained linguistic contexts [1]. Nonetheless, language representation within these PMLMs remains unbalanced; LRLs are heavily under-represented, as their presence is dictated by the volume of monolingual data available for each language during the pre-training phase [2]. Consequently, when models are fine-tuned using training sets of identical proportions, the performance achieved for well-represented languages significantly outpaces that of languages with lower representation [3]. Furthermore, the quantity of task-specific training data allocated for downstream PMLM fine-tuning serves as a critical determinant of performance [3, 4]. Because assembling labeled datasets for LRLs presents a substantial bottleneck, exploring alternative optimization paths for LRL text classification via PMLMs is imperative.

Intermediate Task Fine-tuning (ITFT) represents a powerful methodology for boosting PMLM efficacy on downstream assignments, including sentiment analysis, when resources are scarce. In a standard ITFT framework, the PMLM

undergoes an initial round of fine-tuning on a dataset derived from a separate language or an alternative task. Subsequently, the model undergoes secondary fine-tuning tailored to the primary target task of the chosen language [5–7]. We define this approach as basic ITFT.

Departing from this basic ITFT paradigm—which depends on a sentiment-labeled dataset from an HRL—we introduce two distinct intermediate tasks (TransIT and AuxIT) engineered from a high-resource language (HRL) lexicon:

- **TransIT**—We convert the expressions within the HRL lexicon into the target LRL using an openly accessible Machine Translation platform. These potentially imperfect translations are paired with the source lexicon expressions to form positive and negative training examples, determined by their respective valence ratings. Utilizing these pairings, the model undergoes intermediate fine-tuning framed as a binary classification objective.
- **AuxIT**—We construct a collection of artificial phrases (termed Auxiliary Phrases [APs]) from the source HRL lexicon, which are then prefixed to every training instance within the target LRL dataset. This data augmentation strategy facilitates a binary classification objective wherein a positive instance occurs when the AP and the text sample express matching sentiment, and a negative instance occurs otherwise. This binary objective functions as an intermediate alignment mechanism, mapping target LRL sentiment terms to their corresponding equivalents in the HRL.

In both methodologies, once the intermediate fine-tuning phase concludes, the model is subsequently fine-tuned using the sentiment classification dataset of the target LRL. Both proposed intermediate procedures seek to supply the PMLM with an extra cross-lingual signal, illuminating the connections between sentiment-bearing words across distinct languages.

The proposed approaches align with the framework established by Ke *et al.* [8] regarding the integration of external knowledge repositories; however, our strategy leverages lexicon-driven fine-tuning rather than modifying the initial pre-training phase. For our methodologies to operate effectively, we assume the availability of an HRL sentiment lexicon containing valence scores for each entry. This assumption is highly realistic, as multiple lexicons of this nature exist for the English language [9, 10].

We conduct evaluations across three LRLs—Sinhala, Tamil, and Bengali—focusing on a sentiment classification task. The empirical outcomes demonstrate that our AuxIT intermediate objective outbalances the vanilla fine-tuning baseline. Notably, this pattern remains consistent during evaluations involving an English dataset. AuxIT also matches or exceeds the capabilities of basic ITFT, which depends on an HRL sentiment classification dataset. Moving further, we explore sequential ITFT by pairing AuxIT with an intermediate task derived from an HRL sentiment classification dataset (the conventional task type utilized in basic ITFT, as detailed above). Simply put, the PMLM is fine-tuned sequentially using our newly designed task alongside the HRL sentiment classification objective. Ultimately, this model is fine-tuned on the sentiment classification data of the target LRL. This sequential ITFT configuration delivers additional performance gains over both vanilla fine-tuning and basic ITFT for the Sinhala language.

We demonstrate that the efficacy of the model is governed more by the caliber of the HRL lexicon than by the linguistic proximity between the HRL and the target LRL. Interestingly, when generating APs via our second methodology, utilizing an HRL lexicon yields superior outcomes compared to utilizing an imperfect, noisy lexicon from the target language itself.

Related Work

Intermediate task fine-tuning

ITFT has emerged as a strategy capable of boosting target task performance when applied to a pre-trained language model (PLM). Phang *et al.* [11] are widely acknowledged as the pioneers who introduced the ITFT concept for PLMs. The authors noted that ITFT aims to mitigate catastrophic forgetting within the PLM while enhancing its overall robustness.

However, they also emphasized that identifying the optimal alignment between intermediate tasks and downstream target tasks remains a complex hurdle. As highlighted by Vu *et al.* [12], elements such as dataset magnitude, domain alignment, and the degree of similarity between the source and target tasks heavily influence the success of ITFT. Pruksachatkun *et al.* [6] executed an empirical assessment of various ITFT configurations to decode the underlying dynamics of cross-task transfer. Their analysis indicated that target objectives requiring reasoning elements, such as question answering, derive the greatest advantage from ITFT.

The application of ITFT spans both token-level and sequence-level contexts, including objectives analogous to sentiment classification. Within sequence-level domains, Savini and Caragea [5] leveraged pre-trained BERT-centric architectures to execute a sarcasm detection task. By incorporating multiple intermediate objectives—such as emotion classification from public tweets and sentiment analysis of film reviews—they discovered that distinct ITFT objectives support the primary task in complementary ways. For token-level operations, de la Rosa [13] applied ITFT toward a target task of loanword detection. Furthermore, ITFT has proven highly viable in sequence-to-sequence operations, including Neural Machine Translation [14].

Use of external knowledge bases

The integration of external knowledge repositories, such as lexicons or knowledge graphs, has been put forward as an alternate strategy to boost PLM performance in downstream NLP applications. Lauscher *et al.* [15] introduced an approach to augment BERT, enabling superior performance across GLUE benchmark [16] evaluations through the infusion of supplementary lexical information. Similarly, Peters *et al.* [17] developed the KnowBERT architecture, which integrates external knowledge via entity-linking mechanisms alongside diverse external references like Wikipedia and WordNet, yielding heightened accuracy in tasks such as Word Sense Disambiguation. Along the same lines, Liu *et al.* [18] infused PLMs with Wikidata knowledge triplets, demonstrating enhanced capabilities in knowledge-dependent downstream tasks.

External linguistic repositories have likewise been utilized to optimize sentiment classification outcomes. Teng *et al.* [19] introduced a straightforward weighted-sum approach that harnesses lexicons to capture context-sensitive features for sentiment modeling. Qian *et al.* [20] achieved enhanced sentiment classification accuracy using LSTM networks supported by a lexicon derived from MPQA [21] and the SST dataset. Suresh and Ong [22] introduced a strategy centered on synthesized vector embeddings to supply models with external knowledge. In particular, sentiment lexicons have frequently served as supplementary knowledge engines to boost the sentiment classification proficiency of deep learning architectures, including CNNs [23] and RNNs [24–26]. Lexicons have also been employed to refine sentiment-aware representations within basic Transformer models [27]. Ke *et al.* [8] designed an approach to instill word-level linguistic information into language models like BERT, utilizing a label-aware pre-training objective combined with SentiWordNet [28] to isolate sentiment-carrying words. Nevertheless, these prior endeavors do not explicitly embed ITFT as a mechanism for target task optimization on PLMs.

Materials and Methods

Our framework leverages a combination of ITFT and sentiment lexicons. We begin by introducing two distinct intermediate objectives, TransIT and AuxIT, built using a sentiment lexicon from an HRL. The PMLM undergoes independent fine-tuning on each of these tasks prior to its final fine-tuning phase for the LRL sentiment classification task. The rationale for implementing these intermediate objectives is to introduce an external cross-lingual alignment signal. This signal leverages the model's comprehension of HRL terms to enhance its capacity to interpret low-resource expressions. Subsequently, we construct a sequential ITFT pipeline by merging the most effective of our new intermediate tasks with an intermediate objective derived from an HRL sentiment classification dataset.

Baselines

We evaluate our approach against three standard frameworks:

- **Zero-Shot Transfer:** The PMLM undergoes training exclusively on an HRL sentiment dataset and is evaluated directly on the LRL evaluation data.
- **Direct Fine-Tuning:** The model is trained solely using the target LRL sentiment dataset, bypassing any intermediate phases.
- **Standard ITFT:** A two-stage training sequence where the PMLM is initially optimized on an HRL sentiment classification dataset before undergoing final fine-tuning on the target LRL data.
-

Bilingual sentiment-word phrases as intermediate task data (TransIT)

Our introductory strategy follows a direct implementation model. We utilize an HRL sentiment lexicon containing specific valence +metrics for each lexical entry. These valence scores quantify word-level sentiment polarity, where values approaching 1 denote escalating positive sentiment and values nearing 0 signify deepening negative sentiment [10, 29].

We compile bilingual phrase sets linking HRL terms with their corresponding LRL translations. These expressions are filtered by their valence classifications, prioritizing strongly positive terms with elevated scores. Each valid phrase pair is assigned a target label of 1 to reflect their shared sentiment orientation.

The formulation of a positive sample (label = 1) is demonstrated below:

Example: For Tamil; “good Nalla[SEP]affection Pācam[EOS]”; (label = 1).

Conversely, negative instances (labeled as 0) are engineered by pairing the original HRL source terms with LRL translations that represent an opposing sentiment quadrant:

Example: “good (Nalla[SEP]toxic Naccu[EOS]”; (label = 0).

In this scenario, English source words are intentionally mismatched with Tamil terms that carry a divergent sentiment meaning. These synthetic phrase pairings form a binary classification objective used to prime the PMLM.

Augmented LRL data as intermediate task data (AuxIT)

In a manner parallel to the TransIT framework, this alternative technique also synthesizes auxiliary phrases (APs) by conditioning on the valence distributions of the HRL lexicon. However, a key distinction lies in our verification step: we filter these phrases to guarantee they accurately project the intended sentiment alignment to the model. This is achieved by running the generated combinations through a standalone sentiment classification model to ensure they are categorized into the predicted sentiment bins.

We hypothesize that these verified APs function as an external orientation anchor for the model. Because each AP is mathematically bound to a verified, pre-established sentiment classification, we expect it to deliver a stabilizing alignment signal for that target sentiment domain during the intermediate learning phase.

To partition the lexicon into distinct positive, neutral, and negative word sets, we first establish static valence score thresholds. This segmentation is performed by mapping word distributions across valence intervals and manually defining the boundary lines for each of the three sentiment categories. We validate these interval boundaries by processing a sample set of English APs through a model optimized on the English corpus, verifying that the system maps the phrases to the predicted sentiment classes.

Following word selection, we generate candidate phrases by computing all possible permutations of the isolated terms. To distill these down to the most effective AP variations, we leverage an independent PMLM previously optimized on a 3-class sentiment corpus from the same parent HRL.

Let the initial pool of candidate APs constructed from the lexicon term permutations be represented as $\mathbf{S}_k$, where k denotes the target sentiment classes ($k \in \{\text{neutral}, \text{negative}, \text{positive}\}$) and $|\mathbf{S}_k| = N$, $N \in (\mathbb{Z}_+)$.

The ultimate selection of the highest-performing AP(s) (defined as the subset s) for a designated sentiment category is calculated via:

$$s = \underset{1 \leq i \leq N}{\operatorname{argmax}}_{i \in \mathbf{S}_k} M(i) \quad (1)$$

where M represents the language model optimized on the English training set, and $|s| \geq 1$.

As outlined in the workflow of **Figure 1**, the best neutral APs are isolated by passing the candidate strings through the pre-trained PMLM and extracting the specific sequences that maximize the positive logit output for that targeted sentiment class. Each AP possesses a dedicated sentiment profile dictated by its constituent vocabulary. While this design bears structural similarities to the concept of “Universal Adversarial Triggers” [30], our method explicitly samples sentiment-bearing terms from a structured lexicon, whereas Wallace *et al.* [30] discover trigger patterns through optimization over a targeted subset of the model’s vocabulary without sentiment-specific constraints.

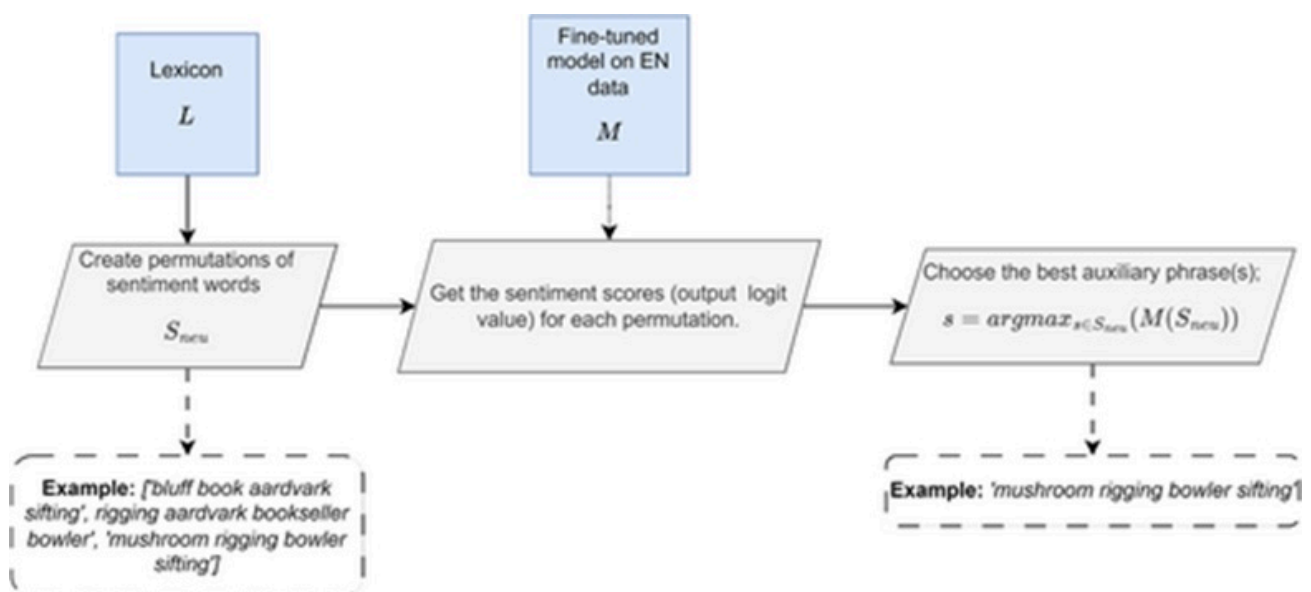


Figure 1. Creating APs from the lexicon. A neutral AP is considered as an example. Dotted boxes connected by blue arrows show example instances in respective steps. Dotted arrows represent inputs from lexicons/datasets.

These filtered APs are subsequently appended to the front of the original text samples in the target language corpus. If the appended AP shares a sentiment orientation matching the attached LRL data sample, the combined instance is labeled 1; otherwise, it receives a 0. A practical application using Sinhala is illustrated in **Figure 2**, where the component words of the AP register neutral scores within the (0.4, 0.6) valence interval, yielding a neutral phrase. The underlying Sinhala text segments translate to “There’s more here than we know” and “This work should be given maximum punishment.”

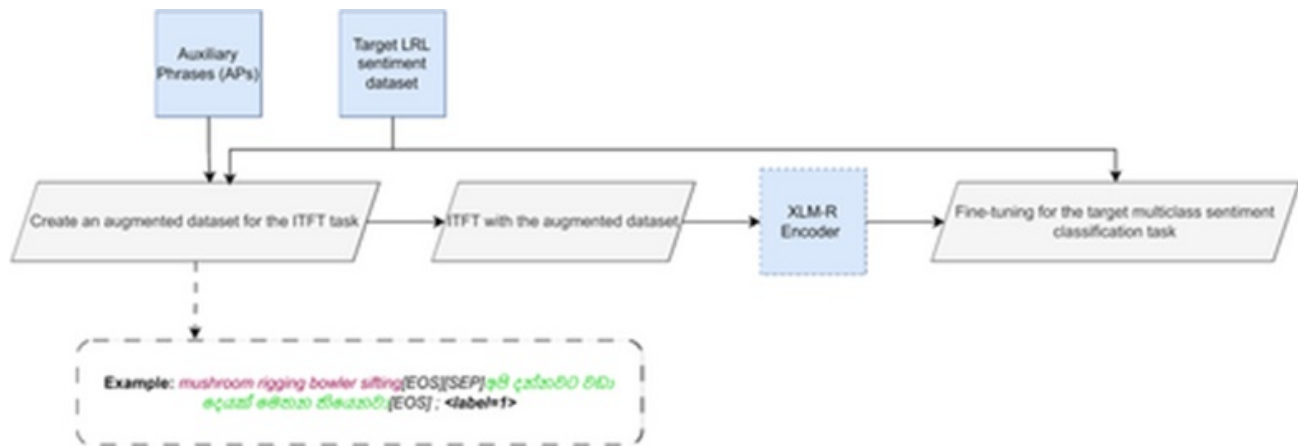


Figure 2. Proposed two-stage fine-tuning method using auxiliary phrases. The neutral AP from **Figure 1** is considered. An example of using AP with an actual data sample from the Sinhala dataset is shown at the bottom. The phrase in 'red' denotes the created AP which is appended by a sentence of the target LRL dataset (in 'green'). Here, the label = 1 is assigned as both phrases carry the same sentiment value (neutral).

The PMLM is fine-tuned on this newly consolidated, augmented corpus. Following this intermediate stage, the model undergoes a final fine-tuning step using the authentic LRL dataset as the definitive learning task (**Figure 2**).

Sequential ITFT

Our experimental evaluations revealed that only the AuxIT framework consistently surpassed the baseline metrics. To build upon this, we merged this specific intermediate objective sequentially with our third baseline setup (which trains the PMLM on an HRL sentiment corpus). Stated plainly, we fine-tune the PMLM on one intermediate task before immediately training it on the other, concluding the pipeline with a final fine-tuning stage on the target LRL training set.

Experiments

Datasets and lexicons

Our English experiments utilize the US Twitter Airline Sentiment corpus (representing a broad, general domain) alongside a specialized financial domain dataset [31]. English acts as the primary HRL across all setups. For the LRL evaluations, we adapt a 4-class Sinhala sentiment corpus [32] consisting of reader commentary sourced from online journalism platforms; for the purposes of this study, the 'conflict' class was removed. The Tamil and Bengali evaluations utilize the data blocks compiled by Hande *et al.* [33] and Islam *et al.* [34], respectively, with the Tamil corpus including code-mixed sentence examples.

The VAD sentiment lexicon [10] serves as our foundational resource for HRL sentiment terms. This database provides validated valence, dominance, and arousal parameters for approximately 20k English terms, alongside translations spanning 102 target languages. Additionally, we run parallel tests using the VADER lexicon [9], which logs 7,520 sentiment-carrying elements (including idiomatic emojis) paired with their respective valence metrics.

Training setup

We adopt XLM-R-base as our core multilingual model, given its comprehensive coverage of our target language suite and its proven efficacy in sentiment classification benchmarks [35, 36]. For the TransIT method, we assemble a 26,000-sample binary corpus drawn from the complete word pool of the VAD lexicon, enforcing a uniform length of 4 terms per instance. For the AuxIT method, we prefix exactly 50% of the training sentences with sentiment-matched APs and the

remaining half with mismatched APs to maintain a perfectly balanced binary class split. All recorded metrics reflect macro-averaged F1 scores computed across three independent training iterations initialized with unique random seeds.

Comparative results for different ITFT setups

The performance metrics are detailed in **Table 1**. Our first configuration, TransIT, demonstrated a performance degradation relative to the baseline models for both Sinhala (achieving a macro-F1 of 69.33%) and Tamil (stopping at 61.08%). On account of these poor early results, we do not include TransIT in the master data table or pass it into the subsequent testing phases. While the TransIT framework is straightforward to execute, its synthetic text combinations often fail to project a cohesive sentiment signal onto the model. Simply containing sentiment-aligned words does not ensure that the combined sequence retains that specific polarity in context. We validated this by routing the TransIT phrases into a fine-tuned XLM-R sentiment model, which routinely failed to assign the predicted target labels. This indicates that the phrases are likely too brief and structurally sparse to preserve essential sentiment context [37, 38].

In contrast, **Table 1** reveals a pronounced performance advantage when comparing the AuxIT method against the direct Fine-tuning baseline (Baseline 2). The most notable improvements occur within the English and Sinhala evaluation tracks. Furthermore, the AuxIT pipeline performs on par with or noticeably better than the standard ITFT pipeline (Baseline 3). It is worth noting that for our English runs—where English functions as the native HRL—the APs are derived internally from the same language, unlike the cross-lingual mapping used for the remaining three languages. Additionally, because English is the source provider for the HRL data, standard ITFT cannot be applied to it. Crucially, even within its native high-resource context, AuxIT yields clear optimization benefits, highlighting the utility of lexicon injection when fine-tuning PMLMs even for high-resource language targets.

Table 1. Macro-F1 scores of experiments with different methods. The best performance for each experiment is indicated with bold numbers.

Dataset	AuxIT→basic ITFT	Basic ITFT→AuxIT	AuxIT	Baseline 3	Baseline 2	Baseline 1
English (<i>Tweets</i>)	-	-	81.32	-	80.17	-
English (<i>Finance</i>)	-	-	88.77	-	87.92	-
Sinhala	71.45	69.10	71.19	70.30	69.61	62.23
Tamil	63.86	61.97	64.67	64.65	63.87	43.46
Bengali	39.14	37.59	43.26	43.31	42.73	40.20

When analyzing the sequential variations, running the standard ITFT phase prior to the AuxIT phase resulted in lower overall metrics. However, reversing the execution order (initiating with AuxIT followed by standard ITFT) unlocked the top performance threshold for the Sinhala language track.

Ablation Study

We execute a series of ablation investigations on the Sinhala corpus to assess the individual contributions of the parameters outlined below. For the first three evaluations, we utilize exactly 1 AP per sentiment class and evaluate the resulting performance metrics against the baseline established through direct fine-tuning of XLM-R (referenced in **Table 2**).

- Impact of the source language chosen for AP synthesis (utilizing the tweets corpus) [6]
- Influence of word-level valence distributions
- Choice of underlying sentiment repository

- Word count variation within a single AP
- Volumetric density of AP instances

Table 2. Results (Macro-F1) for experiments with varying attributes of the APs for Sinhala sentiment dataset. The best performance obtained for each experiment is indicated with bold numbers.

Experiment ID	Modified AP Characteristic	Configuration	F1 Score
–	Reference Model (Baseline 2 – Standard Fine-Tuning)	–	69.61
1	APs Generated in Multiple Languages	English	70.56
		Sinhala	69.66
		Tamil	70.28
		Bengali	69.16
		Hindi	69.73
2	APs Chosen at Random	Randomly Selected APs	67.66
3	APs Constructed Using Alternative Lexicons	VAD Sentiment Lexicon	70.56
		VADER	69.99

Effect of the language used to create APs

To evaluate how the source language (specifically regarding genetic language proximity) of the vocabulary repository impacts model calibration, we configure APs using alternative language origins. We synthesize sentiment lexicons for Hindi, Tamil, and Bengali by translating the primary English VAD database via Google Translate. The corresponding performance variations are documented under Experiment 1 in [Table 2](#).

Even though Hindi and Bengali share a common language family branch with Sinhala, and Tamil shares immediate geographic proximity, the resulting metrics fall short of the scores achieved using the English repository. We hypothesize that this discrepancy stems from the disproportionately high representation profile that English commands within the XLM-R pre-training corpus [1]. Furthermore, translating core APs into alternative target languages introduces a high probability of automated translation errors; the resulting linguistic noise acts as an additional destabilizing factor. Concrete instances of these translation discrepancies include the following items:

- The term “flop” (assigned a highly negative valence metric of 0.081) lacks an authentic semantic equivalent in Sinhala, defaulting instead to a phonetic transliteration of the English word.
- The expression “forge” (positioned within the neutral region at a valence score of 0.52) maps to a Sinhala translation that isolated only a single semantic variation—“deceptive imitation”—which properly belongs within a negative sentiment quadrant.
- The multi-word phrase “pissing me off” (noted as negative with a valence rating of 0.208) is incorrectly mapped to a Sinhala expression carrying an inverted, highly positive meaning: “pibidev” (the English translation of this target Sinhala word corresponds to “Arise”).
- The noun “abbot” completely lacks any translated equivalent within the target language output.

Such systematic errors frequently appear when lexicon compilers attempt to adapt foundational databases into LRL environments using automated translation mechanisms [39, 40].

Effect of the valence scores of lexicon words

To isolate the operational significance of using exact valence ratings when constructing APs, we generated arbitrary AP instances by sampling terms at random from across each valence boundary. This randomized grouping precipitated a measurable contraction in classification performance (detailed under Experiment 2 in Table 2). This degradation indicates that random sampling frequently captures terms with weak or ambiguous sentiment orientations (i.e., valence metrics that lack polarized intensity). These empirical findings validate the targeted AP optimization and filtering routine embedded within our AuxIT protocol.

Effect of the sentiment lexicon

Experiment 3 in Table 2 tracking lexicon selection displays comparative metrics between the VADER and VAD databases. The VAD framework demonstrates superior performance, a capability we attribute to its extensive vocabulary variety, particularly its dense collection of terms representing the neutral classification domain.

Effect of the length of APs and number of APs

We expanded our research to isolate both the ideal token count within an individual AP and the optimal quantity of AP structures deployed per sentiment category. As shown by the orange trajectory in Figure 3, increasing the word density within an AP results in a steady decline in accuracy, which likely indicates that the model undergoes localized over-fitting during the intermediate training phase. We capped our word-length trials at 8 tokens per AP due to the exponential computation times required to compute and run all token permutations. Concurrently, our testing indicates that deploying exactly 2 distinct AP configurations yields peak efficiency; scaling beyond this number degrades final performance metrics, as illustrated by the blue trajectory in Figure 3.

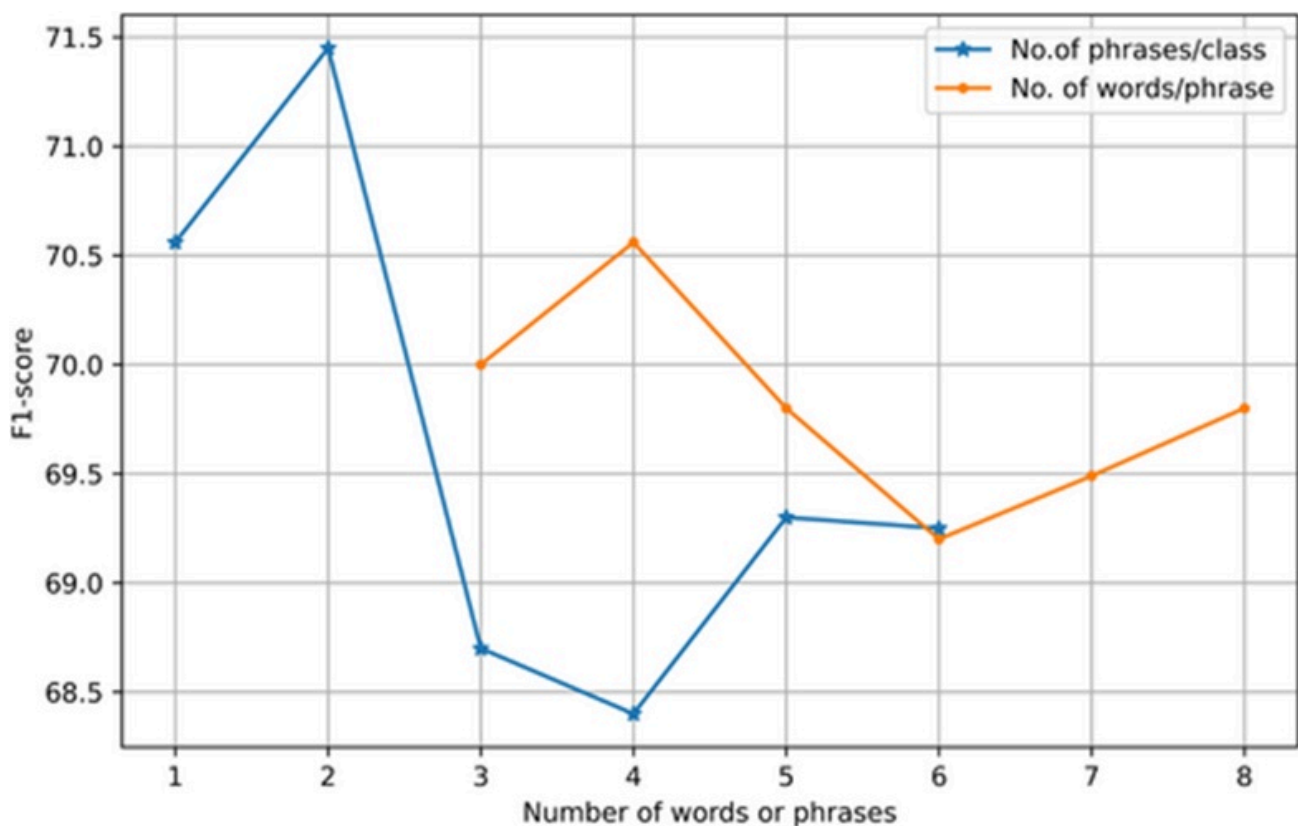


Figure 3. macro-F1 score with varying number of APs per class and no. words per AP.

Impact of ITFT on cross-lingual alignment of sentiment words

By integrating our two proposed intermediate optimization pipelines, we intend to project an enhanced cross-lingual orientation anchor into the PMLM framework using AP structures. To verify if this mapping operates as intended, we map out and examine the latent-space topography (word vector embeddings) for isolated sentiment tokens across English and Sinhala, contrasting a vanilla setup against our AuxIT intermediate task pipeline. We manually selected a reference set of positive and negative expressions from both languages. The spatial mapping in [Figure 4](#) confirms that equivalent cross-lingual sentiment expressions group into significantly closer proximity following the intermediate training iteration. This clustering effect is exceptionally pronounced within the negative sentiment cluster [\[7\]](#).

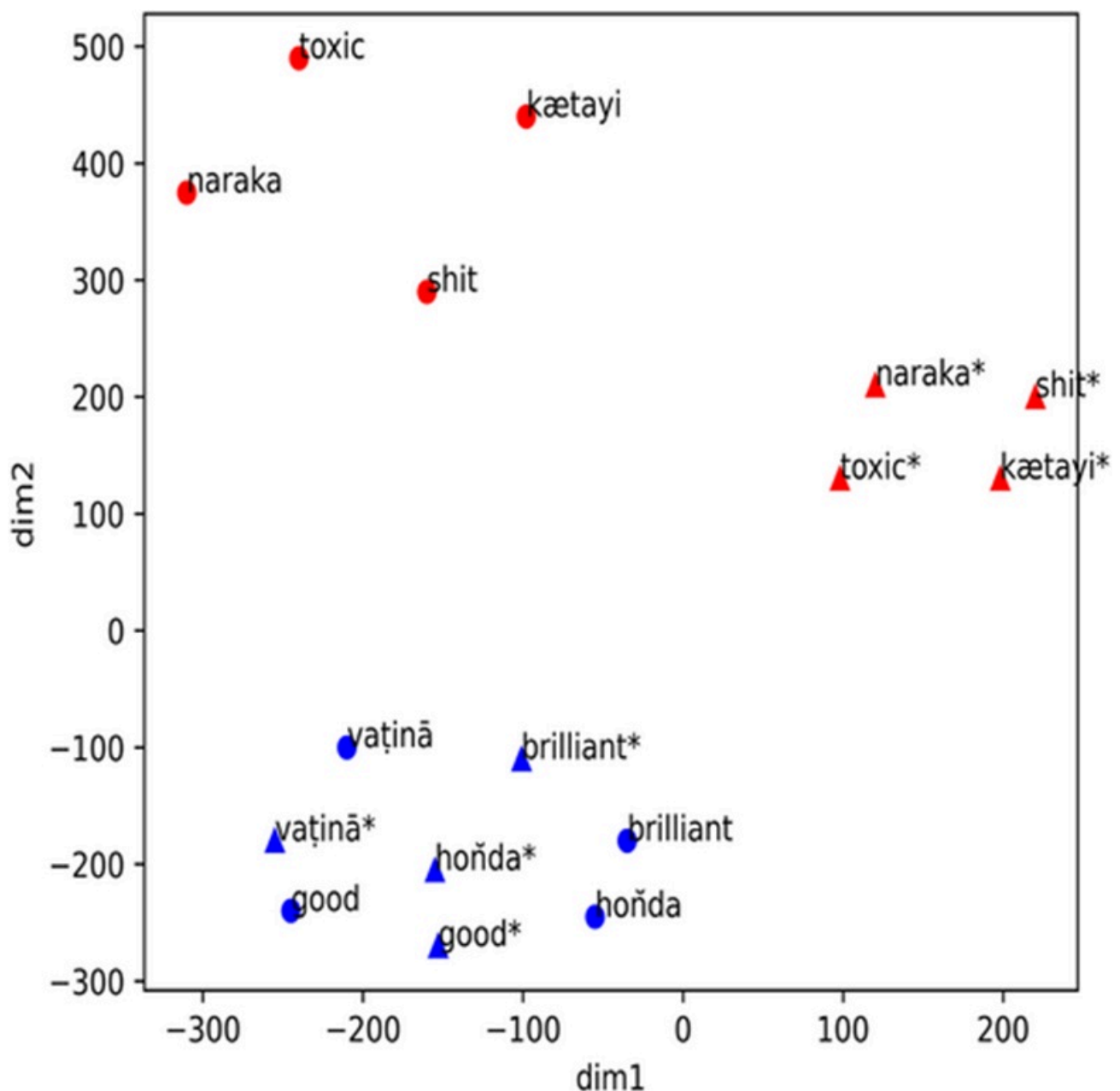


Figure 4. Word embeddings visualisation for positive (blue), negative (red) words in Sinhala and English. The circle markers show embeddings from a vanilla fine-tuned model (on Sinhala dataset) and the triangle markers show

embeddings from our approach.

Conclusion

This paper introduced two distinct cross-lingual intermediate task fine-tuning methodologies for optimizing LRL sentiment classification via PMLMs, driven by an HRL sentiment lexicon. Among these strategies, optimizing the model on an augmented training corpus built from an HRL database (AuxIT) generated substantial performance margins over standard vanilla training tracks. The AuxIT framework systematically equaled or outperformed conventional ITFT baselines. We demonstrated that these performance gains are explicitly driven by the newly introduced intermediate pipeline (AuxIT), which delivers a targeted cross-lingual alignment signal that guides the PMLM to recognize semantic equivalencies among sentiment-bearing tokens across disparate languages. Additionally, we formulated a sequential ITFT pipeline that optimizes the underlying PMLM by linking AuxIT and standard ITFT in a consecutive training stream.

The current system architecture was assessed solely using languages natively supported within the XLM-R model. Future research efforts will focus on expanding this methodology to alternative language models and low-resource languages missing from current pre-trained vocabularies. A parallel optimization goal involves refining the framework to handle fine-grained sentiment classification targets, shifting beyond the coarse-grained, three-class sentiment definitions analyzed in this study.

Acknowledgements

None

Conflict of interest

None

Financial support

Vinura Dhananjaya was funded by a Senate Research Committee grant (SRC/LT/2020/11) of the University of Moratuwa.

Ethics statement

None

Received: 03 Apr 2023

Revised: 21 Apr 2023

Accepted: 20 Jun 2023

Published online: 15 September 2023

Rights and permissions

Open Access The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at <https://creativecommons.org/licenses/by-nc-sa/4.0/>.

References

- Hu J, Ruder S, Siddhant A, Neubig G, Firat O, Johnson M. XTREME: a massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In: Proceedings of the 37th International Conference on Machine Learning (ICML). PMLR; 2020. p. 4411-21.
- Ranathunga S, de Silva N. Some languages are more equal than others: probing deeper into the linguistic disparity in the NLP world. In: Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (AACL-IJCNLP). Association for Computational Linguistics; 2022. p. 823-48.
- Wu S, Dredze M. Are all languages created equal in multilingual BERT? In: Proceedings of the 5th Workshop on Representation Learning for NLP (RepL4NLP). Association for Computational Linguistics; 2020. p. 120-30.
- Doddapaneni S, Ramesh G, Khapra MM, Kunchukuttan A, Kumar P. A primer on pretrained multilingual language models. arXiv [Preprint]. 2021.
<https://doi.org/10.48550/arXiv.2107.00676>.
- Savini E, Caragea C. Intermediate-task transfer learning with BERT for sarcasm detection. *Mathematics*. 2022;10(5):844.
<https://doi.org/10.3390/math10050844>.
- Pruksachatkun Y, Phang J, Liu H, Htut PM, Zhang X, Pang RY, et al. Intermediate-task transfer learning with pretrained language models: when and why does it work? In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL). Association for Computational Linguistics; 2020. p. 5231-47.
- Chang TY, Lu CJ. Rethinking why intermediate-task fine-tuning works. In: Findings of the Association for Computational Linguistics: EMNLP 2021. Association for Computational Linguistics; 2021. p. 706-13.
- Ke P, Ji H, Liu S, Zhu X, Huang M. SentiLARE: sentiment-aware language representation learning with linguistic knowledge. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics; 2020. p. 6975-88.
- Hutto C, Gilbert E. VADER: a parsimonious rule-based model for sentiment analysis of social media text. *Proc Int AAAI Conf Web Soc Media*. 2014;8(1):216-25.
- Mohammad S. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL). Association for Computational Linguistics; 2018. p. 174-84.
- Phang J, Févry T, Bowman SR. Sentence encoders on STILTs: supplementary training on intermediate labeled-data tasks. arXiv [Preprint]. 2018.
<https://doi.org/10.48550/arXiv.1811.01088>.
- Vu T, Wang T, Munkhdalai T, Sordoni A, Trischler A, Mattarella-Micke A, et al. Exploring and predicting transferability across NLP tasks. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics; 2020. p. 7882-92.
- DelaRosa J. ADoBo 2021: the futility of STILTs for the classification of lexical borrowings in Spanish. In: IberLEF@SEPLN. 2021. p. 947-55.
- Nayak S, Ranathunga S, Thillainathan S, Hung R, Rinaldi A, Wang Y, et al. Leveraging auxiliary domain parallel data in intermediate task fine-tuning for low-resource translation. arXiv [Preprint]. 2023.
<https://doi.org/10.48550/arXiv.2306.01382>.

Lauscher A, Vulić I, Ponti EM, Korhonen A, Glavaš G. Informing unsupervised pretraining with external linguistic knowledge. arXiv [Preprint]. 2019.
<https://doi.org/10.48550/arXiv.1909.02339>.

Wang A, Singh A, Michael J, Hill F, Levy O, Bowman SR. GLUE: a multi-task benchmark and analysis platform for natural language understanding. In: Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. Association for Computational Linguistics; 2018. p. 353-5.

Peters ME, Neumann M, Logan R, Schwartz R, Joshi V, Singh S, et al. Knowledge enhanced contextual word representations. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Association for Computational Linguistics; 2019. p. 43-54.

Liu L, Li X, He R, Bing L, Joty S, Si L. Enhancing multilingual language model with massive multilingual knowledge triples. arXiv [Preprint]. 2021.
<https://doi.org/10.48550/arXiv.2111.10962>.

Teng Z, Vo DT, Zhang Y. Context-sensitive lexicon features for neural sentiment analysis. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics; 2016. p. 1629-38.

Qian Q, Huang M, Lei J, Zhu X. Linguistically regularized LSTM for sentiment classification. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL). Association for Computational Linguistics; 2017. p. 1679-89.

Wilson T, Wiebe J, Hoffmann P. Recognizing contextual polarity in phrase-level sentiment analysis. In: Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics; 2005. p. 347-54.
<https://doi.org/10.3115/1220575.1220619>.

Suresh V, Ong DC. Using knowledge-embedded attention to augment pre-trained language models for fine-grained emotion recognition. In: 2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII). IEEE; 2021. p. 1-8.

Shin B, Lee T, Choi JD. Lexicon integrated CNN models with attention for sentiment analysis. In: Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. Association for Computational Linguistics; 2017. p. 149-58.

Kumar A, Kawahara D, Kurohashi S. Knowledge-enriched two-layered attention network for sentiment analysis. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics; 2018. p. 253-8.

Margatina K, Baziotis C, Potamianos A. Attention-based conditioning methods for external knowledge integration. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL). Association for Computational Linguistics; 2019. p. 3944-51.

Ma Y, Peng H, Cambria E. Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive LSTM. Proc AAAI Conf Artif Intell. 2018;32(1).
<https://doi.org/10.1609/aaai.v32i1.12048>.

Zhong P, Wang D, Miao C. Knowledge-enriched transformer for emotion detection in textual conversations. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Association for Computational Linguistics; 2019. p. 165-76.

Baccianella S, Esuli A, Sebastiani F. SentiWordNet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). European Language Resources Association; 2010.

Mehrabian A. Pleasure-arousal-dominance: a general framework for describing and measuring individual differences in temperament. *Curr Psychol*. 1996;14(4):261-92.

Wallace E, Feng S, Kandpal N, Gardner M, Singh S. Universal adversarial triggers for attacking and analyzing NLP. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics; 2019. p. 2153-62.

Malo P, Sinha A, Korhonen P, Wallenius J, Takala P. Good debt or bad debt: detecting semantic orientations in economic texts. *J Assoc Inf Sci Technol*. 2014;65(4):782-96.
<https://doi.org/10.1002/asi.23062>.

Senevirathne L, Demotte P, Karunanayake B, Munasinghe U, Ranathunga S. Sentiment analysis for Sinhala language using deep learning techniques. *arXiv [Preprint]*. 2020.
<https://doi.org/10.48550/arXiv.2011.07280>.

Hande A, Hegde SU, Priyadharshini R, Ponnusamy R, Kumaresan PK, Thavareesan S, et al. Benchmarking multi-task learning for sentiment analysis and offensive language identification in under-resourced Dravidian languages. *arXiv [Preprint]*. 2021.
<https://doi.org/10.48550/arXiv.2108.03867>.

Islam KI, Islam MS, Amin MR. Sentiment analysis in Bengali via transfer learning using multilingual BERT. In: *2020 23rd International Conference on Computer and Information Technology (ICCIT)*. IEEE; 2020. p. 1-5.

Rathnayake H, Sumanapala J, Rukshani R, Ranathunga S. Adapter-based fine-tuning of pre-trained multilingual language models for code-mixed and code-switched text classification. *Knowl Inf Syst*. 2022;64(7):1937-66.
<https://doi.org/10.1007/s10115-022-01698-1>.

Dhananjaya V, Demotte P, Ranathunga S, Jayasena S. BERTifying Sinhala: a comprehensive analysis of pre-trained language models for Sinhala text classification. In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC 2022)*. 2022. p. 7377-85.

Zhan J, Dahal B. Using deep learning for short text understanding. *J Big Data*. 2017;4:34.
<https://doi.org/10.1186/s40537-017-0095-2>.

Hussein DME. A survey on sentiment analysis challenges. *J King Saud Univ Eng Sci*. 2018;30(4):330-8.
<https://doi.org/10.1016/j.jksues.2016.04.002>.

Wan Y, Yang B, Wong DF, Chao LS, Yao L, Zhang H, et al. Challenges of neural machine translation for short texts. *Comput Linguist*. 2022;48(2):321-42.
https://doi.org/10.1162/coli_a_00435.

Bapna A, Caswell I, Kreutzer J, Firat O, van Esch D, Siddhant A, et al. Building machine translation systems for the next thousand languages. *arXiv [Preprint]*. 2022.
<https://doi.org/10.48550/arXiv.2205.03983>.