MaskedVerbalizer: Automatic Verbalizer Construction for Few-Shot Text Classification in Low-Resource Right-to-Left Languages

Faizad Ullah[1,+,*], **Furqan Sikandar**[2,+], **Areeba Waqar**[2], **Faizan Ali**[2],
Muhammad Sohaib Ayub[3], **Mubashar Mushtaq**[2], **Asim Karim**[1]
[1]Department of Computer Science, Lahore University of Management Sciences (LUMS), Pakistan
[2]Department of Computer Science, Forman Christian College University, Pakistan
[3]Data Science Institute, University of Galway, Ireland
20030057@lums.edu.pk,furqansikandar45@gmail.com,areebawaqar07@gmail.com,
r6s1337.xyz@gmail.com, muhammadsohaib.ayub@universityofgalway.ie,
mubasharmushtaq@fccollege.edu.pk,akarim@lums.edu.pk
[+] Equal contribution by the first two authors.
[*] Corresponding Author

Abstract

Text classification in low-resource right-to-left languages faces significant challenges due to the scarcity of annotated data and the morphological richness of languages such as Arabic, Urdu, Sindhi, and Pashto. Arabic and Urdu alone are spoken by over $380+$ million and $246+$ million people worldwide, respectively. Pashto is the national language of Afghanistan, highlighting the importance of effective language technologies. While multilingual Pre-trained Language Models (PLMs) have shown promising results, they typically require extensive labeled datasets and computationally expensive fine-tuning to achieve better performance. Such limitations make these PLMs impractical for the low-resource settings described above. Therefore, we employ a few-shot strategy (zero, 4, or 8 shots) to achieve results comparable to those of standard fine-tuning. In this work, we propose MaskedVerbalizer, a novel technique designed for few-shot text classification. Our method introduces an automatic verbalizer construction approach that generates class-specific label words in 4-shot settings, eliminating the need for extensive manual intervention. Despite maintaining a simple model architecture, MaskedVerbalizer achieves effective performance in classification benchmarks. Experimental results demonstrate that our method effectively addresses the core challenges of low-resource text classification, providing a practical, computationally efficient solution. We achieved accuracies of 90.43% and 92.72% with mBERT and XLM-RoBERTa, respectively, representing improvements of 25–30% over soft and automatic verbalizers. The code for MaskedVerbalizer is publicly available at https://github.com/Furqann-hue/MV.

Keywords: Prompt Engineering, Verbalizer Construction, Low-Resource Languages, MaskedVerbalizer, Few-Shot Learning

1. Introduction

Low-resource languages such as Arabic and Urdu pose significant classification challenges due to sparse labeled data and complex linguistic features. Although multilingual Pre-trained Language Models (PLMs) like mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020) have improved generalization across languages, standard fine-tuning techniques remain dependent on large quantities of labeled data and substantial computational resources. Prompt-based fine-tuning has recently gained attraction as a more data-efficient alternative, particularly in few-shot learning scenarios (Ullah et al., 2023). This approach reformulates classification tasks as Masked Language Modeling (MLM) problems using prefix and cloze-style templates (Ullah et al., 2025) and label words (Gao et al., 2021). By leveraging the language understanding already embedded in PLMs, prompt-based methods can generalize effectively from few annotated examples, e.g., 4 or 8-shot. However, existing prompt-based methods still face noticeable limitations. For instance, automatically generated or manually crafted templates and verbalizers often fail to capture nuanced linguistic patterns, especially in morphologically rich, script-diverse languages such as Arabic, Urdu, Pashto, and Sindhi (Faheem et al., 2026). These hand-crafted approaches struggle to adapt across tasks, resulting in inconsistent performance and limited scalability. Consequently, there remains a need for adaptive, language-aware prompt-optimization techniques that better exploit the representational power of PLMs in low-resource contexts.

To address these limitations, we propose a novel technique for automatic verbalizer generation, called MaskedVerbalizer. While many techniques—including automatic template generation and prompt-based optimization—have been explored, most existing methods for verbalizer generation remain inconsistent or impractical across

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 795–804
11–16 May 2026.

tasks and languages. Our approach uses few-shot (4-shot) examples with prefix and cloze-style templates and a base verbalizer initialized with two seed words from GPT. The model is then fine-tuned (prompt-based) to adapt effectively from minimal supervision. Following fine-tuning, we evaluate the model on a 50% dataset split and collect predictions at the `[MASK]` position. For each correctly classified instance, the predicted token at the mask index is extracted and aggregated separately for each class. The most frequently occurring tokens per class are then selected as the final label words, forming a task-specific verbalizer without manual intervention.

We evaluate our framework on multiple right-to-left languages, including Arabic, Urdu, Pashto, and Sindhi, on various classification tasks using 4-shot training. Results show that MaskedVerbalizer achieves up to 93% accuracy (XLM-R on Arabic Tweets) and consistently outperforms Soft (Ding et al., 2022) and Automatic Verbalizers (Schick et al., 2020), with gains of 10–25% across datasets. Even on noisy data like Urdu Corpus v1 (Khan and Nizami, 2020), it maintains competitive accuracy (e.g., 64.97% with XLM-roberta-Large), while delivering effective performance on other datasets such as iNLTK Urdu News (89.77%) (Arora, 2020) and Arabic Spam-Ham (92.72%) (Kaddoura and Henno, 2024). To evaluate the effectiveness of our framework on left-to-right languages, we also included baseline results on the English dataset (Socher et al., 2013). The main contributions of this work are as follows:

- We propose MaskedVerbalizer, a novel self-supervised approach that automatically constructs task-specific verbalizers by leveraging the model's own masked token predictions.
- We demonstrate that MaskedVerbalizer achieves substantial improvements (25–30%) over existing verbalizers (soft and automatic) in 4-shot settings across multiple low-resource right-to-left languages, including Arabic, Urdu, Pashto, and Sindhi.
- We provide a comprehensive evaluation across five languages and six datasets using mBERT, XLM-RoBERTa-large, and LLaMA 3.2 (1B), showing that our method achieves performance comparable to fully supervised fine-tuning while using only 4 training examples per class.

2. Literature Review

As social media and news platforms increasingly feature Arabic, Urdu, Pashto, and Sindhi, automated text classification becomes crucial for extracting actionable insights. Recent advances highlight prompt-based fine-tuning as a more efficient alternative for few-shot learning, where classification is reformulated as an MLM task using prefix and cloze-style templates and predefined label words (Gao et al., 2021; Ullah et al., 2024). Building on this, several studies have shown that prompt-based fine-tuning consistently outperforms standard fine-tuning for text classification in low-resource languages, including Urdu and Roman Urdu (Ullah et al., 2023).

To reduce dependence on manual template design, automatic template search methods such as LM-BFF (Gao et al., 2021) automatically identify effective prompt templates, improving few-shot performance. Similarly, soft prompt tuning (Lester et al., 2021) learns a small set of continuous prompt embeddings to guide the model, enabling efficient adaptation while keeping the pre-trained architecture intact. Chen and Shu (2023) introduced a label-guided data augmentation framework (PromptDA) that enriches the prompt space using semantic label information, effectively blending label semantics with data augmentation to boost few-shot learning performance.

Earlier work, such as (Schick and Schütze, 2021), introduced Pattern-Exploiting Training (PET), demonstrating how handcrafted prompts and verbalizers can guide language models in few-shot classification tasks. However, manually designing verbalizers remains time-consuming and difficult to scale across domains and multilingual datasets. To address this, recent works have focused on automatic verbalizer generation. Hu et al. (2022) proposed Knowledgeable Prompt-Tuning (KPT), which enriches verbalizers with external knowledge bases to reduce bias and improve stability in few-shot classification. Likewise, Cui et al. (2022) introduced Prototypical Verbalizer (ProtoVerb), which learns class-level prototype vectors from training data through contrastive learning, capturing richer semantic information and improving performance in few-shot scenarios. Unlike prior verbalizer-generation approaches that rely on manual curation or external resources, our MaskedVerbalizer leverages the model's own masked predictions to iteratively refine label words in a fully self-supervised manner. This enables efficient, language-agnostic adaptation with minimal data in low-resource text classification.

3. Methodology

Prompt-based fine-tuning of Large Language Models (LLMs) involves three key components: (1) prompt selection, (2) template design, and (3) verbalizer construction (Gao et al., 2021). The prompt selection part focuses on choosing input

examples that best represent the dataset's diversity, while template design aims to provide the best template that can transform input instances into open/cloze style prompts through handcrafted templates (e.g., "یہ {mask} ہے۔") (e.g "This sentence is [mask]."). The verbalizer construction part maps model-predicted vocabulary tokens to task-specific class labels, and remains a critical bottleneck. Poor alignment between verbalizer tokens and label semantics often limits performance, especially in morphologically rich and low-resource languages such as Arabic and Urdu. Formally, given an input space X, a label set Y, a vocabulary V of the PLM, a template function $T : X \rightarrow Prompt(x)$, and a verbalizer $V_b : Y \rightarrow V$, the language model M predicts labels by maximizing the probability of the verbalizer token for the masked prompt:

$$\hat{y} = \arg\max_{y \in Y} P_{\mathcal{M}}\big(V_b(y) \mid T(x)\big) \quad (1)$$

While significant work has been done across all three parts, existing verbalizers often fail to capture the semantic complexity of class labels in low-resource right-to-left languages, thereby limiting overall classification accuracy. To address this issue, our proposed MaskedVerbalizer will generate verbalizers that better align with label semantics, thereby maximizing classification performance across these languages.

3.1. MaskedVerbalizer

The proposed MaskedVerbalizer operates by leveraging a model's own masked token predictions, extracted from a held-out portion of the dataset, to iteratively refine class-indicative label words. This data-driven refinement process eliminates reliance on external knowledge sources or complex manual tuning, instead aligning the verbalizer with the model's contextual understanding and the dataset's intrinsic semantics check. For each language, the dataset is divided into $70/30$ train/test split. The 70% training portion is further subdivided into two disjoint subsets: 20% of the full dataset is used to select few-shot examples and to generate an initial verbalizer. In contrast, the remaining 50% of the full dataset is reserved for model evaluation, during which masked token predictions are collected for MaskedVerbalizer. The final evaluation is performed on the 30% test split. From the 20% dataset, we select four examples per class to form a 4-shot configuration. For each class, two initial verbalizer words are produced either manually (based on linguistic and dataset intuition) or automatically by prompting GPT using only the 4-shot samples collected from the 20% subset as shown in Figure 1. A uniform manual template is used in all experiments, which is translated for each language. The template is as follows:

```
This sentence is [MASK].
```

Using the above template and the initial verbalizer words, we perform prompt-based fine-tuning of the chosen PLM (XLM-RoBERTa and mBERT) on the 4-shot samples. After the few-shot fine-tuning, the fine-tuned model is evaluated on the 50% dataset split. During this evaluation, we record only the correctly classified instances and extract the word predicted at the `[MASK]` position for each such instance. For each class, we compute the frequency distribution of these predicted words. The MaskedVerbalizer for each class is then constructed by selecting the top two or three most frequent words according to the following rule:

- If the top three words exhibit comparable frequencies (for example, $130, 120, 110$), all three words are retained.
- If the third word's frequency is substantially lower than the top two, only the top two words are retained.

This procedure ensures that the resulting verbalizer is optimally aligned with both the dataset's characteristics and the model's representational behavior. To increase robustness in realistic low-resource settings, the following heuristics are applied when necessary:

- **Label noise:** If the dataset exhibits labeling errors (e.g., frequent label flips), the verbalizer may be expanded to include up to five high-frequency words per class to reduce sensitivity to noisy annotations.
- **Class imbalance and word overlap:** When class imbalance causes the same high-frequency word to appear as a top candidate for multiple classes, the following procedure is adopted:
  1. Compare the word frequency across classes; the class with the higher frequency retains the word.
  2. For the smaller class, promote its second-best word to first if its frequency is close to the original top word's frequency. If the second word exhibits a substantially lower frequency, the smaller class retains its top two or three words as usual.

This strategy mitigates ambiguous word sharing across classes and results in verbalizers that more effectively capture discriminative semantics, particularly for underrepresented classes.

The finalized, dataset-specific MASKEDVERBALIZER is integrated into the same few-shot prompt setup, utilizing the same 4-shot examples and template. The base model is fine-tuned with these refined verbalizers following the standard training procedure, and subsequently evaluated on the held-out 30% test split. The performance of the MASKEDVERBALIZER is compared against standard OpenPrompt verbalizers (Automatic and Soft) (Ding et al., 2022) as well as the fully supervised $70/30$ baseline. Empirically, the MASKEDVERBALIZER yields substantial improvements over competitive verbalizers in 4-shot settings and, in several cases, approaches the performance of the fully supervised model.

4. Evaluation

In this section, we describe the datasets and PLMs used in our experiments.

4.1. Datasets

We utilized multiple languages and datasets, Arabic (one dataset), Urdu (two datasets), Pashto (one dataset), Sindhi (one dataset), and English (one dataset) to evaluate our approach MASKEDVERBALIZER using prompt-based fine-tuning. The Arabic Ham and Spam Tweets dataset (Kaddoura and Henno, 2024) comprises $13,241$ Arabic tweets collected via Twitter API between January and March 2021. Each tweet is labeled as either Ham or Spam. The iNLTK Urdu News Dataset (Arora, 2020) consists of approximately $4,500$ Urdu news headlines curated from various news websites, enabling multiclass classification. The Urdu Sentiment Corpus (Urdu Corpus v1) (Khan et al., 2017; Khan and Nizami, 2020) has positive and negative sentiment classes. We consider a single "O" class as an outlier in the Urdu Corpus v1 and have removed it. Although labeled as positive or negative, our analysis revealed that many samples were semantically closer to neutral, reducing the clarity of binary sentiment boundaries. The Pashto Text Sentiment Analysis Dataset (Ali, 2024) consists of approximately $21,800$ sentences collected from Wahdat Newspaper issues and articles by professors, annotated into three sentiment classes: positive, negative, and neutral. The Safali-S/Sindhi-News-Headline-Dataset-for-Category-Based-Sentiment-Analysis (Soomro et al., 2025) comprises $30,462$ headlines collected from various online Sindhi news platforms. Each headline is annotated with both its category and sentiment polarity; for these experiments, we focused on the sentiment polarity labels, which are classified as positive, negative, or neutral. Finally, we utilized SST-2 (Stanford Sentiment Treebank v2) dataset (Socher et al., 2013), an English sentiment dataset consisting of sentences from movie reviews, each labeled as either positive or negative.

4.2. Pretrained Language Models

We utilized two PLMs, i.e., (1) Multilingual BERT (mBERT) (Devlin et al., 2019), introduced by Google, and (2) XLM-RoBERTa-large (XLM-R) (Conneau et al., 2020), which is a large-scale multilingual transformer model developed by Facebook AI. The mBERT model has shown competitive performance on a variety of multilingual Natural Language Processing (NLP) tasks. The XLM-R consistently outperforms mBERT in multilingual benchmarks and is particularly effective for cross-lingual understanding. We also utilized the Llama 3.2 (1B) (AI, 2024), an open-weight generative model from Meta's Llama 3 series optimized for efficient multilingual reasoning. For reproducibility, all models were fine-tuned using a learning rate of 2×10^{-5} and trained for 5 epochs using a fixed random seed of 20. The optimization method is AdamW (an Adam optimizer variant).

5. Results and Discussion

We evaluated the performance of the MASKEDVERBALIZER approach across five languages and six datasets using three different models, including mBERT, XLM, and Llama 3.2(1B). We compare our results with other proposed techniques such as Soft Verbalizer (SV) (Ding et al., 2022), Automatic Verbalizer (AV) (Schick et al., 2020), Knowledge Verbalizer (KV) (Hu et al., 2022), GPT/Manual Verbalizer (GPT) (OpenAI, 2025), and Knowledge Verbalizer (KV) combined with GPT Verbalizer (GPT) and Knowledge Verbalizer (KV) combined with our proposed technique MASKEDVERBALIZER (MV). Additionally, we include zero-shot and 70–30 standard supervised fine-tuning for comparison. We also analyzed LLaMA 3.2(1B) and followed a prompt-based approach similar to GPT-style inference. Figure 2 illustrates the LLaMA prompt template employed in our experiments. This comparison helps us analyze the difference between fine-tuned PLMs (mBERT and XLM-R) models and LLaMA.

5.1. mBERT Performance

As illustrated in Table 1, the mBERT model (utilizing MASKEDVERBALIZER) shows noticeable improvements over both SV and AV across almost all datasets. For example, on the Arabic Spam and Ham Tweets dataset, MV achieves 90.43% accuracy and 77.16% macro-F1 score, substantially higher than SV, which is 86.23% accuracy,

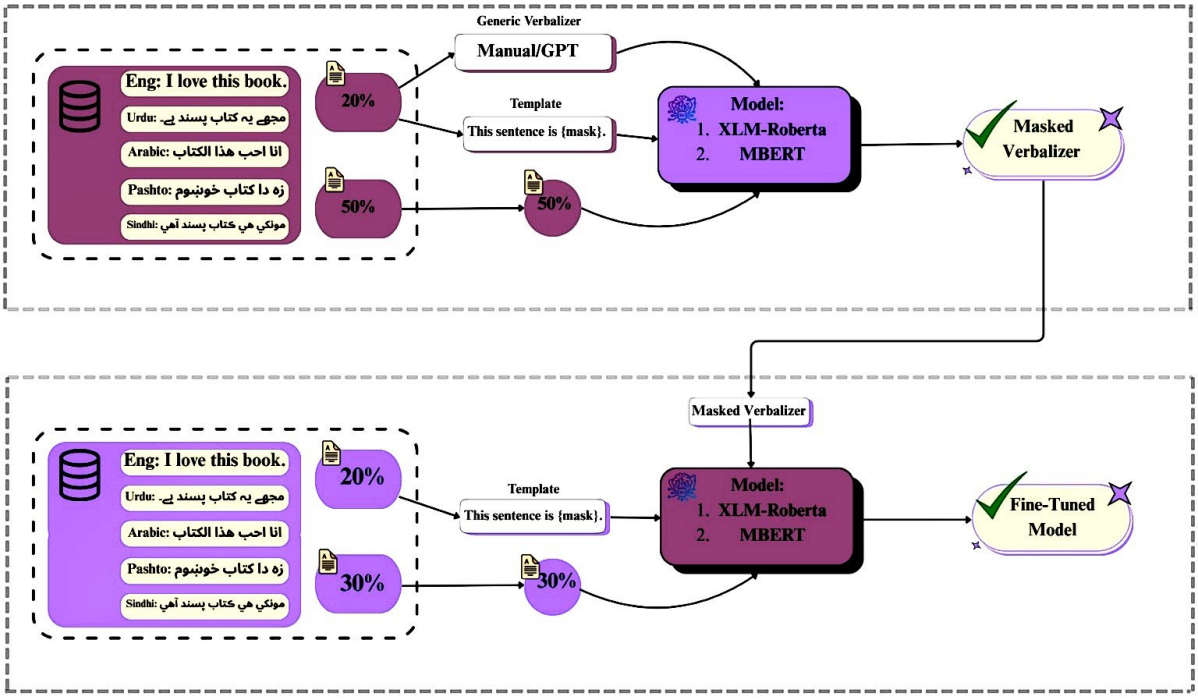

Figure 1: The MASKEDVERBALIZER framework. The top section shows verbalizer construction: 4 examples per class are selected from 20% of the dataset to create an initial verbalizer (manual or GPT-based) with the template `This sentence is [MASK]` (translated for each language). The model is fine-tuned and evaluated on 50% of data to extract masked predictions that form the MASKEDVERBALIZER. The bottom section shows the final training phase using the constructed verbalizer, with evaluation on the remaining 30% of the test set.

Lang.	0-shot	4-shot						70/30
		AV	SV	MV	MV+KV	GPT	GPT+KV	
				mBERT				
Arabic	34.0/33.2	48.9/41.8	86.2/53.0	90.4/77.1	**90.4/77.2**	78.9/67.6	79.4/67.7	**99.5/99.0**
Urdu(iNLTK)	43.7/42.0	34.6/32.0	47.6/39.1	72.8/68.9	**74.0/70.2**	51.7/50.5	51.9/50.6	**93.3/92.2**
Urdu v1	45.5/33.3	45.2/45.2	48.3/36.6	48.3/47.6	41.8/39.5	49.3/40.5	**51.3/47.9**	**64.6/64.2**
				XLM-RoBERTa				
Pashto	30.6/25.9	32.7/**32.5**	38.5/32.0	39.2/25.1	**39.8**/25.7	34.1/32.3	34.2/32.4	**40.6**/19.2
Sindhi	44.4/32.1	32.9/32.3	52.3/44.1	84.2/83.5	**84.5/83.9**	71.1/66.3	71.0/66.1	**93.4/93.5**
Arabic	70.5/52.6	51.1/43.7	88.2/80.5	**92.7/84.2**	90.9/78.6	88.2/69.4	88.3/69.8	85.3/46.0
Urdu(iNLTK)	35.0/27.7	33.1/30.5	55.6/37.0	**89.7/87.3**	89.5/87.0	88.7/84.8	88.9/84.9	**92.1/91.2**
Urdu v1	51.0/50.7	52.3/52.3	52.7/48.4	**64.9/63.7**	**64.9/63.7**	53.4/45.6	54.0/47.0	51.0/33.7
SST-2	53.4/50.8	51.4/51.4	72.1/71.7	86.2/86.2	86.1/86.1	72.7/70.9	74.2/72.7	**91.5/91.4**

Table 1: Performance comparison of MASKEDVERBALIZER against existing verbalizer methods. Results are reported as Accuracy/Macro-F1. Abbreviations: Automatic Verbalizer (AV), Soft Verbalizer (SV), MASKEDVERBALIZER (MV), Knowledge Verbalizer (KV), GPT Verbalizer (GPT), MV+KV (combined approach), and GPT+KV (combined approach). The 70/30 baseline represents standard fine-tuning with 70% training and 30% test data.

53.05% macro-F1 score, and AV (48.94% accuracy, 41.82% macro-F1 score. The MV also outperforms GPT verbalizer across all datasets. However, there is one exception on the Urdu Corpus v1 dataset, where MV performs slightly worse than GPT. A closer look reveals that the dataset's quality is a limiting factor. Although it is labeled as binary (positive vs. negative), many samples have a neutral tone, making it difficult for both the model and human annotators to clearly distinguish be-

Lang.	0-shot	4-shot	70/30
		LLaMA 3.2 (1B)	
Pashto	32.96/30.19	34.72/**34.22**	**42.55**/33.01
Sindhi	31.77/19.18	59.39/45.70	**93.22**/**93.29**
Arabic	86.08/51.41	72.83/59.40	**99.62**/**99.25**
Urdu	41.84/31.54	14.29/17.72	**72.11**/**72.11**

Table 2: Performance comparison of LLaMA 3.2 (1B) in 0-shot, 4-shot, and fully supervised ($70/30$) settings. Results are reported as Accuracy/Macro-F1.

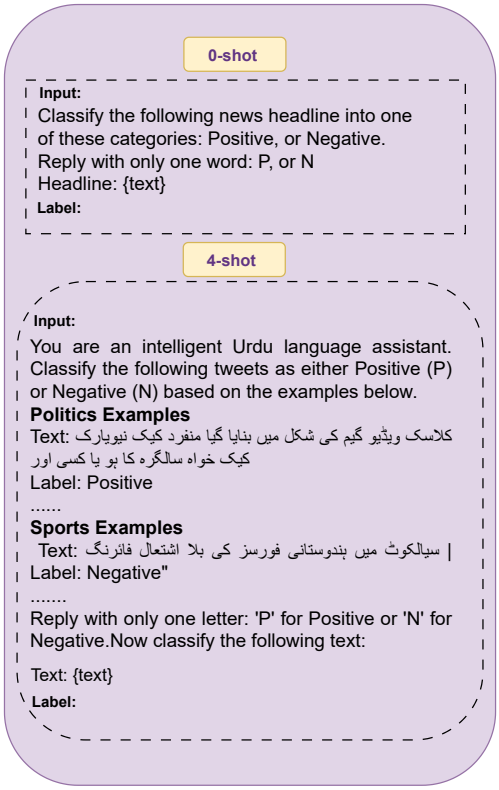

Figure 2: Llama Template for 0-shot and 4-shot

tween the classes.

Nevertheless, MV proves highly competitive with fully supervised fine-tuning. For instance, on the Arabic Spam and Ham Tweets dataset, MV reaches 90.43% accuracy with just 4-shot, approaching the standard fine-tuning performance of 99.50% accuracy. MV combined with KV shows better performance on Arabic and Urdu (iNLTK News) datasets, surpassing nearly all zero-shot and 4-shot baselines. Specifically, on the Urdu iNLTK News dataset, it achieves 74.01% accuracy and 70.22% macro-F1, representing an absolute increase of 59.27 percentage points in accuracy and 28.16 percentage points in macro-F1 over the zero-shot performance (14.74% accuracy, 42.06% macro-F1). This notable improvement is attributed to the synergy between MV and KV: MV provides high-quality discriminative label words, while KV intelligently processes and aggregates predictions across them. Hence, leading to more precise and reliable classification outcomes. Overall, these results highlight that MV combined with KV excels at few-shot text classification in low-resource languages.

5.2. XLM-R Performance

The XLM-R model exhibits substantial performance improvements with the proposed MASKEDVERBALIZER as seen in Table 1. Moreover, across nearly all datasets and languages, MV consistently outperforms alternative verbalizers, including AV, SV, GPT, and GPT+KV. For example, on the Urdu (iNLTK) News dataset, XLM-R achieves 89.77% accuracy and 87.32% macro-F1, surpassing all zero-shot and 4-shot baseline configurations. Remarkably, MV also outperforms standard supervised fine-tuning across several datasets, including Arabic Spam and Ham Tweets, Urdu Corpus v1, and SST-2. Specifically, on the Arabic Spam and Ham Tweets dataset, the model attains 92.72% accuracy and 84.21% macro-F1, outperforming supervised fine-tuning, which yields 85.36% accuracy and 46.05% macro-F1.

Even on noisy datasets, such as Urdu Corpus v1, where labels are frequently misassigned, both MV and MV+KV outperform all baseline approaches. For instance, on the Pashto dataset, MV combined with KV achieves 39.81% accuracy, approaching the standard supervised fine-tuning performance of 40.64% accuracy. Additionally, on the Sindhi dataset, MV combined with KV surpasses all zero-shot and 4-shot baselines, achieving 84.58% accuracy and 83.97% macro-F1, corresponding to absolute improvements of 47.11% in accuracy and 51.82% in macro-F1 over zero-shot performance. These results demonstrate that combining the KV with the MASKEDVERBALIZER is highly effective even under label noise and extremely limited data. In contrast, KV combined with GPT never surpasses any 4-shot baseline in our experiments. See Appendix A for qualitative analysis of the verbalizer words generated by mBERT and XLM-R.

Furthermore, to evaluate the generalizability of our proposed framework to left-to-right languages, we tested it on a subset of the English SST-2 dataset. The MASKEDVERBALIZER achieved 86.25% accuracy and 86.23% macro-F1, surpassing all zero-shot and 4-shot baselines. Overall, these results demonstrate that the MASKEDVERBALIZER is not only effective in low-resource right-to-left languages but can also be successfully scaled to other left-to-right languages.

5.3. LLama 3.2 (1B)

The experimental results in Table 2 demonstrate a clear distinction between the performance of MaskedVerbalizer models and instruction-tuned LLaMA models across multiple low-resource languages. For the Arabic *Ham–Spam* dataset, MV achieved better results with XLM-RoBERTa, obtaining an accuracy and macro-F1 score of 92.72% and 84.21%, respectively, outperforming mBERT with 90.43% and 77.16%, respectively. In contrast, LLaMA 3.2 (1B) exhibited competitive performance only in the $70/30$ fine-tuning setup (99.62% / 99.25%), whereas its 0-shot and 4-shot settings lagged behind (86.08% / 51.41% and 72.83% / 59.40%)(2). This performance gap underscores that LLaMA, being an autoregressive model rather than an MLM, struggles in few-shot or zero-shot scenarios where the model must infer task semantics purely from the prompt structure without supervised adaptation.

Figure 2 presents the prompt format used for LLaMA-based evaluations. In the 0-shot and 4-shot settings, LLaMA relies solely on its instruction-following capability to interpret such prompts, following a GPT-style autoregressive decoding strategy rather than masked words prediction as used in MLM-based models. Consequently, its predictions often reflect broad linguistic priors rather than precise task-specific distinctions, leading to greater variability and lower F1 scores across languages. In contrast, using MV with prompt-based fine-tuning with mBERT and XLM-RoBERTa directly exploits the masked-prediction objective to acquire class-indicative lexical cues from few-shot examples, thereby achieving more stable and generalizable performance even under limited data availability.

For the other low-resource datasets (Urdu, Sindhi, and Pashto), MV models displayed a similar robustness trend, though performance varied with language resource availability. XLM-R consistently outperformed mBERT, achieving 64.97% / 63.70% on Urdu, 84.28% / 83.57% on Sindhi, and 39.29% / 25.10% on Pashto. In comparison, LLaMA 3.2 1B's performance improved substantially with full fine-tuning (e.g., 93.22% / 93.29% on Sindhi and 72.11% / 72.11% on Urdu) but remained weak in few-shot and zero-shot setups, particularly for morphologically rich and underrepresented languages such as Pashto. These findings collectively highlight that while MaskedVerbalizer effectively leverages contextual token prediction even in few-shot scenarios, instruction-based generative models such as LLaMA require significant supervised adaptation to achieve comparable results, especially in low-resource multilingual contexts.

6. Conclusion

In this paper, we introduced MaskedVerbalizer, a novel verbalizer construction technique designed to enhance prompt-based fine-tuning for low-resource right-to-left languages. Despite the progress of soft and automatic verbalizers, these methods often fail to fully capture class-specific nuances, particularly in multilingual or low-resource scenarios. MaskedVerbalizer consistently outperforms soft and automatic verbalizers across multiple datasets and models. Furthermore, both mBERT and XLM-R, when utilized with our MaskedVerbalizer technique combined with the KV, deliver consistently better performance across diverse languages and datasets. These findings underscore the crucial role of verbalizer quality in prompt-based learning, especially within low-resource and multilingual contexts. Our results demonstrate that even with limited data, a well-constructed verbalizer can significantly improve model performance and overall effectiveness.

In future work, we aim to evaluate MaskedVerbalizer on a broader range of multilingual and domain-specific datasets, further refine its construction process, and enhance its 4-shot performance by jointly optimizing template–verbalizer pairs for each dataset. We believe these directions will lead to more efficient and scalable approaches for prompt-based fine-tuning in low-resource NLP scenarios.

7. Limitations

The limitations of the proposed technique (MaskedVerbalizer) can be observed from the results, especially with the Urdu v1 (Khan et al., 2017; Khan and Nizami, 2020) corpus. When the dataset is noisy or poorly labeled, it can hinder the technique by preventing the model from predicting strong, distinctive words for each class. Also, there is a challenge related to the model dependence of the technique itself, which acts as a double-edged sword. As seen in the results from both mBERT and XLM-R, the performance varies between the two. This is because each model behaves differently and predicts different words. While one could argue that the words predicted by a stronger model, such as XLM-R, could be reused for mBERT to potentially improve its accuracy, this would violate the core principle of MaskedVerbalizer. The technique is fundamentally based on the idea that each model should use the words it predicts in correctly classified instances, as those words are the ones the model associates with the respective classes. Therefore, a model's own capabilities can limit the

performance of MaskedVerbalizer.

8. Ethical Considerations

We acknowledge several ethical considerations in this work. The datasets from social media and news platforms may contain inherent biases that trained models could perpetuate, and users should consider these limitations when deploying our methods in real-world applications. While our approach is intended for beneficial classification tasks such as spam detection and sentiment analysis, we recognize that this technology could potentially be misused for surveillance or censorship purposes. We strongly advocate for responsible use with appropriate safeguards in place. Finally, our work is limited to a subset of low-resource right-to-left languages due to data availability constraints, which does not reflect the relative importance of excluded languages.

References

Meta AI. 2024. LLaMA 3.2: A New Generation of Multilingual Foundation Models. https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/. Model release; 1B and 3B parameter versions.

Aizaz Ali. 2024. Pashtu Text Sentiment Analysis Dataset. https://doi.org/10.17632/s9k7dk9sc6.1. Mendeley Data, V1.

Gaurav Arora. 2020. iNLTK: Natural Language Toolkit for Indic Languages. In *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*, pages 66–71, Online. Association for Computational Linguistics.

Canyu Chen and Kai Shu. 2023. PromptDA: Label-guided Data Augmentation for Prompt-based Few Shot Learners. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 562–574, Dubrovnik, Croatia. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Ganqu Cui, Shengding Hu, Ning Ding, Longtao Huang, and Zhiyuan Liu. 2022. Prototypical Verbalizer for Prompt-based Few-shot Tuning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7014–7024, Dublin, Ireland. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ning Ding, Shengding Hu, Weilin Zhao, Yulin Chen, Zhiyuan Liu, Haitao Zheng, and Maosong Sun. 2022. OpenPrompt: An Open-source Framework for Prompt-learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 105–113, Dublin, Ireland. Association for Computational Linguistics.

Ali Faheem, Faizad Ullah, Muhammad Hammad, Ahmed Hassan, Muhammad Sohaib Ayub, and Asim Karim. 2026. U-MIRAGE: Benchmarking Chain-of-Thought Reasoning for Urdu Medical QA. In *The 2nd Workshop on NLP for Languages Using Arabic Script (AbjadNLP @ EACL 2026)*.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making Pre-trained Language Models Better Few-shot Learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830, Online. Association for Computational Linguistics.

Shengding Hu, Ning Ding, Huadong Wang, Zhiyuan Liu, Jingang Wang, Juanzi Li, Wei Wu, and Maosong Sun. 2022. Knowledgeable Prompt-tuning: Incorporating Knowledge into Prompt Verbalizer for Text Classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2225–2240, Dublin, Ireland. Association for Computational Linguistics.

Sanaa Kaddoura and Safaa Henno. 2024. Dataset of Arabic spam and ham tweets. *Data in Brief*, 52:109904.

Muhammad Yaseen Khan, Shah Muhammad Emaduddin, and Khurum Nazir Junejo. 2017.

Harnessing English Sentiment Lexicons for Polarity Detection in Urdu Tweets: A Baseline Approach. In *2017 IEEE 11th International Conference on Semantic Computing (ICSC)*, pages 242–249.

Muhammad Yaseen Khan and Muhammad Suffian Nizami. 2020. Urdu Sentiment Corpus (v1.0): Linguistic Exploration and Visualization of Labeled Dataset for Urdu Sentiment Analysis. In *2020 International Conference on Information Science and Communication Technology (ICISCT)*, pages 1–15.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The Power of Scale for Parameter-Efficient Prompt Tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

OpenAI. 2025. GPT-5. `https://openai.com/`. Large language model.

Timo Schick, Helmut Schmid, and Hinrich Schütze. 2020. Automatically Identifying Words That Can Serve as Labels for Few-Shot Text Classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5569–5578, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021. Exploiting Cloze-Questions for Few-Shot Text Classification and Natural Language Inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Safdar Soomro, Siti Yuhaniz, Mazhar Dootio, Dr Mujtaba, and Jawaid Siddiqui. 2025. Category-Based Sentiment Analysis of Sindhi News Headlines Using Machine Learning, Deep Learning, and Transformer Models. *IEEE Access*, PP:1–1.

Faizad Ullah, Ubaid Azam, Ali Faheem, Faisal Kamiran, and Asim Karim. 2023. Comparing Prompt-Based and Standard Fine-Tuning for Urdu Text Classification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6747–6754, Singapore. Association for Computational Linguistics.

Faizad Ullah, Ali Faheem, Ubaid Azam, Muhammad Sohaib Ayub, Faisal Kamiran, and Asim Karim. 2024. Detecting Cybercrimes in Accordance with Pakistani Law: Dataset and Evaluation Using PLMs. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4717–4728, Torino, Italia. ELRA and ICCL.

Faizad Ullah, Safiullah Faizullah, Imdad Ullah Khan, Turki Alghamdi, Toqeer Ali Syed, Ahmad B Alkhodre, Muhammad Sohaib Ayub, and Asim Karim. 2025. Prompt-based fine-tuning with multilingual transformers for language-independent sentiment analysis. *Scientific reports*, 15(1):20834.

A. Qualitative Analysis of MaskedVerbalizer

Table A1 presents a qualitative look at the verbalizer words produced by GPT, mBERT, and XLM-RoBERTa. A few patterns stand out. GPT words are generally intuitive, for example, اعلان، ترویج (announcement, promotion) for Arabic spam and اداکارہ، ویب سیریز (actress, web series) for Urdu entertainment. These words reflect human intuition rather than model behavior. In contrast, mBERT often selects function words or generic tokens such as بھی، کیا (also, what) for Urdu sentiment. This suggests that mBERT's masked predictions are less discriminative for morphologically rich languages. XLM-R, on the other hand, tends to produce more semantically grounded words, as seen in کرکٹ، میچ (cricket, match) for the cricket class and *horrible, ridiculous* for SST-2 negative sentiment. However, for Pashto and Sindhi, both models predict overlapping tokens across classes. For instance, زور، پې appears under all three Pashto sentiment classes. This reflects the limited multilingual coverage of these models for Pashto. The Sindhi results are particularly interesting. The GPT verbalizer assigns ڪيل، ميچ (played, match) to the negative class, which aligns poorly with sentiment. This is consistent with our earlier observation that the Sindhi dataset contains labeling noise. This further explains why MaskedVerbalizer's gains are more modest on that dataset compared to Arabic.

Language	Dataset	Class	GPT Words	mBERT Words	XLM-RoBERTa Words
Arabic	Spam-Ham Tweets	Spam	اعلان، ترويج	است، العربية	فقط، من
		Ham	خبر، تقرير	التالية، العسكرية	خبر، من
Urdu	iNLTK News	Entertainment	اداکارہ، ویب سیریز	فلم، کیا	فلم، اداکارہ
		Cricket	کھلاڑی، ٹیم	منتخب، بازی	کرکٹ، میچ
		Crime	تشدد، آبروریزی	بھی، قتل	جرم، قتل
	Corpus v1	Positive	تفریح، معاشرت	بھی، کیا	میں، مبارک
		Negative	سیاست، تنقید	نہیں، میں	غلط، جرم
Pashto	Sentiment	Positive	ښه، مثبت	–	پې، زور، مشهور
		Negative	بد، منفي	–	دا، زور، پې
		Neutral	بې‌طرف، عادي	–	نه، زور، پې
Sindhi	Sentiment	Positive	سياست، حڪومت	–	سياست، غلط
		Negative	ڪِل، ميچ	–	بهترين، صحيح
		Neutral	جُرم، قتل	–	قتل، پوليس
English	SST-2	Positive	Good, Excellent	–	very, beautiful, great
		Negative	Bad, Terrible	–	horrible, ridiculous, bad

Table A1: Qualitative analysis of verbalizer words. GPT Words refer to the initial seed words, while mBERT and XLM-RoBERTa Words refer to the final words selected by MaskedVerbalizer using the respective models. Words are shown in their original script.