scientific reports

Check for updates

OPEN

Dataset creation and benchmarking for Kashmiri news snippet classification using fine-tuned transformer and LLM models in a low resource setting

Deheem U Deyar, Anirud Ramani, Deepa Gupta✉, Priyanka C. Nair & Manju Venugopalan✉

Kashmiri language, recognized as one of the low-resource languages, has rich cultural heritage but remains underexplored in NLP due to lack of resources and datasets. The proposed research addresses this gap by creating a dataset of 15,036 news snippets for the task of Kashmiri news snippets classification, created through the translation of English news snippets into Kashmiri using the Microsoft Bing translation tool. These snippets are manually refined to ensure domain specificity, covering ten categories: Medical, Politics, Sports, Tourism, Education, Art and Craft, Environment, Entertainment, Technology, and Culture. Various machine learning, deep learning, transformer-models, and LLMs are explored for text classification. Among the models experimented for classification, fine-tuned ParsBERT-Uncased emerged as the best-performing transformer model, achieving an F1 score of 0.98. This work not only contributes a valuable dataset for Kashmiri but also identifies effective methodologies for accurate news snippet classification in the Kashmiri language. This research developed an essential dataset, which to our best belief, is the first attempt at creating a manually labelled corpus for the Kashmiri language and also devised an architecture using the best combination of embeddings, algorithms, and transformer-models for accurate text classification. It contributes significantly to the field of NLP for this underrepresented language.

Keywords Low-resource languages, Kashmiri NLP, Dataset creation, Kashmiri news snippet classification, Large language models, Transformers

The Kashmiri language, or Koshur, is an Indo-Aryan spoken by nearly 7 million native speakers in the Indian-administered union territory of Jammu and Kashmir[1,2]. The language is one of the major Dardic languages in the Indo-Iranian section of the Indo-European family of languages[3,4], Kashmiri holds a unique position with its use of both Perso-Arabic and Devanagari scripts[5]. Kashmiri is second fastest growing language in India[6], and is recognised as official language of Jammu and Kashmir by Official Languages Act, 2020, it is being promoted in education and official uses, with plans to make it a compulsory subject in schools[7,8]. The language is rich in cultural, historical, and linguistic importance, Kashmiri is relatively less explored in Natural Language Processing (NLP). The lack of standardized linguistic materials, large-scale data, and thin databases from single sources are critical issues[9,10]. Indian languages like Hindi also suffer from similar issues, and hence strong linguistic resources are required for text classification and processing[11]. Studies are also exploring ways to address resource scarcity in Indian languages and are trying to leverage lexical similarity to exploit vocabulary overlap among language datasets[12]. Identifying Kashmiri as a cultural treasure and an economic growth and tourism driver, this work aims to develop NLP tools and resources for the language. Such innovation holds the potential to enhance education through the availability of native-language materials, facilitate wide-ranging and reasonable quality education, and promote continuous learning among the region's linguistic communities.

Since there is limited work done on Kashmiri language and has no comprehensive dataset for Kashmiri language processing, it is necessary to collect and curate the dataset manually. The process of manually creating a dataset for the Kashmiri language poses several significant challenges. The major limitation is the scarcity of

Department of Computer Science and EngineeringAmrita School of Computing, Amrita Vishwa Vidyapeetham, Bengaluru, India. ✉email: g_deepa@blr.amrita.edu; v_manju@blr.amrita.edu

resources, there are not enough comprehensive linguistic materials, datasets, and tools available for Kashmiri when compared to more widely spoken languages. This lack of resources complicates the processes of data collection, translation and analysis; however, the entire undertaking of data creation manually also involves other difficulties. Specifically, it requires extensive hours spent gathering and refining texts, ensuring translation accuracy and navigating variations in sentence structures and dialects.

Although the Kashmiri language possesses intricate grammatical and phonetic features, this complexity raises numerous issues in the development of consistent, high-quality datasets with adequate recall for tasks associated with natural language processing. To create a meaningful and relevant dataset, English news snippets are collected from ten domains: Medical, Politics, Sports, Tourism, Education, Art and Craft, Environment, Entertainment, Technology, and Culture. These news snippets are manually refined wherein each news snippet is carefully analysed and modified to fit the specific characteristics and requirements of its corresponding domain. This iterative refinement ensured that the dataset had highly relevant and domain-specific content that is suitable for the task at hand. Further, the refined news snippets are translated into Kashmiri using the Microsoft Bing translation tool. This resulted in 15,036 sentences for all ten domains, such that the dataset is diverse and representative. By adapting the dataset to the linguistic features of the specific domains, Kashmiri news snippets can be classified with more relevant and accurate data. This data creation process not only increases the dataset but also the reliability and precision of the models, so the proposed work aims to analyse the impact of various factors on the classification of Kashmiri news snippets and provide insights into the best strategies for processing and categorizing text data in low-resource languages like Kashmiri.

The research work's distinguishing novelty lies in the following aspects:

- This proposed work contributes the first ever labelled dataset for Kashmiri text classification for the task of news snippet classification from 10 different domains which include education, culture, art and craft politics etc. helping to bridge the gap in NLP infrastructure for low-resource languages. The earlier efforts like AI-4Bharat's Kashmiri-English parallel corpora have focused only on machine translation tasks. A recent survey in 2024[13] investigating linguistic resources in the Kashmiri language found that the only available resources include a few Kashmiri-English dictionaries, Kashmiri Wordnet, a monolingual corpus under the EMILLE project, parallel Kashmiri corpora such as NLLB-200 by Meta AI and BPCC by AI4Bharat, as well as a spell checker and a speech tool for Kashmiri. These findings highlight the scarcity of linguistic resources for NLP in Kashmiri where the proposed work introduces a novel contribution by developing a multiclass text classification dataset for Kashmiri news snippet classification, addressing a critical gap in resources for this low-resource language.
- This work also provides a foundation for wider tasks and establishes first benchmark for Kashmiri text classification by adapting ML, DL, transformers, and LLMs to a low-resource language, thus laying the foundation for future work in Kashmiri NLP.

The key contributions of this research work include:

- The work involves manually creating a labelled dataset for multi-class classification of news snippets, focusing on 10 diverse news domains within the Kashmiri language.
- A thorough comparative analysis of Machine Learning classifiers and Deep Learning architectures, both using pre-trained embeddings, along with fine-tuned Transformers models and LLMs, is conducted to evaluate their effectiveness and assess their suitability for classifying Kashmiri news snippets, implementing comprehensive experimentation with various hyperparameters and training methodologies to enhance performance across models.
- The work contributes to preservation of endangered languages, promoting digital inclusion and empowering Kashmiri speakers through educational and practical applications.

Related works

The nature of the Kashmiri language and other low-resource Indic languages poses challenges for NLP. Recent studies highlight efforts to compile datasets for such low resource languages (LRL). These include techniques for developing dialogue systems without using pre-existing data[14,15], the creation of tools for Peru's endangered languages to automatically gather and process language data from the web to produce monolingual datasets[16,17].

Qumar et al.[13] reported that the only available Kashmiri linguistic resources include a few Kashmiri-English dictionaries, Kashmiri Wordnet, a monolingual corpus under the EMILLE project, parallel Kashmiri corpora such as NLLB-200 by Meta AI and BPCC by AI4Bharat, as well as a spell checker and a speech tool for Kashmiri underscoring resource creation's role for advancing NLP tools and preserving the linguistic diversity of Kashmiri language. The authors emphasized that most of these resources exhibited low quality, redundancy, and weak cultural-linguistic representation, underscoring the urgent need for high-quality Kashmiri language resources. Fesseha et al.[18] demonstrated the feasibility of CNNs and word embeddings for Tigrinya text, a low-resource (Afro-Asiatic) language. It provided hope to many resource-poor languages. Their work offered the vector for later advancement. Yu et al.[19] recently proposed quantum-based update of the basic RNNs to aid in boosting the performance of the classification exercises under limited resource constraints. However, other works proposed by Cruz & Cheng established baseline studies focussed on languages with low resources for text classification[20]. Marivate et al. stressed on creating a good dataset and ventured into data collection in low-resource languages like Setswana and Sepedi providing a roadmap for future advancements[21]. Similar efforts are further enhanced by the proposed domain-adversarial learning to improve text classification for low-resource languages described by Grießhaber[22]. As well as cross-lingual model fine-tuning as proposed by Li et al.[23] Adelani et al. [24] worked on Masakhanews, a news classification in African languages, highlighting challenges and solutions. Similarly, the

research hints the use of pre-trained embeddings and machine learning models for news topic categorization and for low-resource languages like Kashmiri[25–27].

In the field of text classification, the transformer model plays the role of an imperative one, especially in low-resource conditions. Similar to the finding of Agbesi et al.[28] fine-tuning pre-trained transformers is effective at text classification in Ewe, providing promising results for African languages. Such a type of success is achieved by Alam et al.[29] who used transformer models on Bangla text, achieving accuracy improvements in resource-scarce environments. The works by Aggarwal et al.[12] further developed the scalability of transformers, providing multilingual classification approaches for Indian languages. On the other hand, Mirashi et al.[30] also more formally introduced the L3Cube-IndicNews dataset as valuable for the further fine-tuning of transformers in various news classification tasks regarding Indic languages. To support such multilingual and low-resource contexts, Multilingual BERT (mBERT) Woo et al. [31] was introduced as a single transformer model trained on 104 languages, enabling cross-lingual transfer learning. mBERT builds upon the architecture of BERT Devlin et al.[32], which introduced deep bidirectional representations by jointly conditioning on both left and right context across all layers. In contrast, ParsBERT Farahani et al.[33], a monolingual Persian model, achieved strong results on Persian classification tasks, often outperforming mBERT in language-specific scenarios. Wertz[34] developed few-shot learning for improved transformer performance, addressing limitations and proposing improved strategies. Arora[35] proposed the inltk tool for Indic language classification, aiming to offer scalable multilingual classification algorithms for Indian languages. And other studies focus on classification using DistilBERT[36], LSTM[37] and other DL and transformer-based models for similar works such as fake news detection[38] or text classification[39]. Taken together, such efforts highlight the growing importance of transformers and fine-tuning strategies in text classification, particularly in low-resource languages. In the last few years, LLMs have proved to be a great instrument for solving text classification tasks, including those in low-resource scenarios. Sun et al. continue to highlight LLMs regarding their effectiveness of the model in a limited resource environment and possibility to deliver high accuracy despite the unavailability of large labelled data[40], or with minimal instances[41]. To refine the outcomes even more, Patwa et al.[42] utilized PEFT and SD augmentation, which are found to advance LLMs on low-resource classification. Joshi et al.[43] fine-tuned the LLMs to the particular task and demonstrated that such strategies can lead to massive spikes in classification effectiveness in LRL's. However, even in the case of LLMs as annotators for low-resource languages, some hurdles are at work, and that is why Jadhav et al.[44] proposed ways to enhance their annotation functions and strategies to improve their application for resource-scarce languages. Finally, Boyina et al.[45] proved the effectiveness of zero-shot and few-shot learning using LLMs in Telugu news classification and stated possibilities of LLMs in performing text classification tasks for LRL's. Altogether, all these works hint at the potential that LLMs can help develop the task of text classification for LRLs.

Proposed methodology

The proposed work has been carried out in two phases: dataset creation for Kashmiri news snippet classification followed by an extensive exploration of different models for news snippet classification. The detailed framework is presented in Fig. 1 and explained in detail in the following subsections.

Manual construction of the Kashmiri news dataset and data description

To build the Kashmiri news dataset, news snippets and articles in English were collected from websites, blogs, and online newspapers such as Greater Kashmir, Rising Kashmir, Gaatha, Caper Travel India, and Incredible India. The links to these sources are listed in the supplementary. The collected content was grouped into ten different domains: Medical, Politics, Sports, Tourism, Education, Art and Craft, Environment, Entertainment, Technology, and Culture. For example, Greater Kashmir and Rising Kashmir provided news related to Politics, Technology, Education, and Medical domains. Gaatha was used to obtain articles from two categories: Art & Craft and Culture. Caper Travel India contributed content on Culture, including festivals and traditional Practices. Incredible India was referred to for Tourism-related information. The English news snippets were initially translated into Kashmiri using Microsoft Bing Translator, as at the time it was the most reliable tool supporting Kashmiri and produced translations that were more semantically coherent and natural compared to alternatives such as OpenL and Rephrasely. The translations served only as draft material to facilitate faster processing, while the final dataset quality was established through systematic human verification. To ensure linguistic accuracy, contextual relevance, and cultural authenticity, a two-step human refinement process was applied, first, the first author, a native Kashmiri speaker, reviewed each snippet, removed ambiguous and noisy entries, and corrected minor lexical and grammatical issues, second, a Kashmiri language expert independently verified all translations for cultural and contextual appropriateness and fluency. This sequential review ensured that the dataset is both linguistically reliable and culturally authentic. This layered validation also helped mitigate potential translation noise and errors that might arise from relying on raw MT output. In addition to human validation, we performed a quantitative evaluation on a representative subset of the data. Using back-translation (Kashmiri → English via Microsoft Bing Translator) and comparison with the original English snippets, we obtained a SacreBLEU score of 44.85, a Corpus BLEU score of 42.89, and an average BERT similarity of 0.9735, demonstrating strong alignment at the domain and semantic level. For annotations, each snippet was initially categorized according to the section of the source website and news outlet (e.g., Politics, Sports, Tourism) as the first step in labeling. This initial assignment was not blindly accepted, both annotators the first author and the Kashmiri language expert carefully reviewed each snippet to verify domain fidelity. In cases where content overlapped multiple domains, the snippet's central theme determined its final label.

The annotation process followed a sequential two-pass validation, where the first author and a Kashmiri language expert independently reviewed and refined after the initial category assignments. This dual-review strategy ensured that every snippet was checked with complementary expertise, resulting in a carefully refined

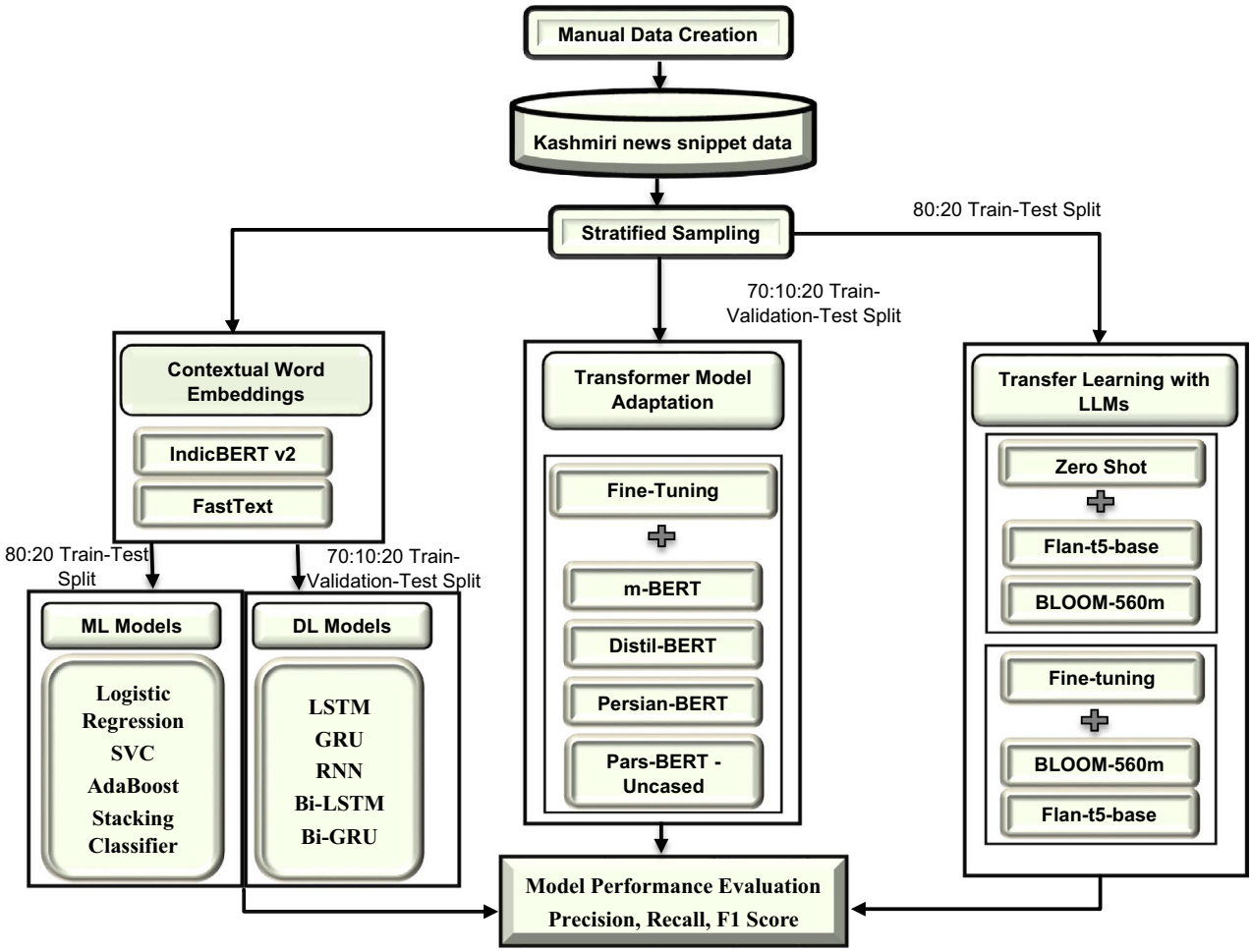


Fig. 1. Proposed framework for news snippet classification in Kashmiri.

dataset rather than raw machine-translated text. Sample Kashmiri news snippets, along with their English translations and domain classification, are shown in Supplementary Table S1. Since the focus is news snippet categorization, minor inconsistencies in spelling and punctuation were considered negligible, as they do not affect the learning of meaningful domain-level patterns. The categorization process emphasized preserving the semantic meaning of each snippet in relation to its domain, ensuring reliable domain distinctions even when small wording differences were present. The distribution plots of the Kashmiri news snippets in Fig. 2 highlights that the topics are well represented across diverse domains. The slight imbalance in the dataset adds complexity and reflects real-world scenarios, making it suitable for training classification models. Figure 3. depicts the sentence length distribution in all domains, and it can be seen that the maximum concentration of sentences is between a range of 25 and 35 words, ensuring consistency and considering the need for a proper amount of textual content for domain-specific classification tasks. The words and news snippet length statistics across the 10 domains is presented in Table 1.

As observed in Table 1, the variation in total and maximum word counts across domains reflects the natural differences in how news is reported- detailed descriptions in Culture and Entertainment leads to longer snippets, while factual updates in Tourism or Education remain shorter. Despite this, the average news snippet word length is more or less consistent across the ten domains which would help prevent model bias toward text length, ensures fair feature representation, and improves generalization by encouraging the model to learn semantic differences rather than relying on surface patterns like input size. The minimum word count differences show that factual updates like Medical, Tourism or Sports need fewer words, while Art & Craft requires more detail even in short snippets to convey meaning. This realistic distribution added complexity and prepared models to handle varied news styles during classification. The word cloud representations for a few domains from the dataset are showcased in Supplementary Fig. S1.

To measure the quality of the dataset for the classification tasks, the similarity between various domains is calculated using Cosine and Jaccard Similarity measures which measure semantic similarity and literal word overlap respectively. The similarity scores across domains remain low, with the highest being just 0.26 between Education and Technology, and the lowest around 0.13 between Medical and Entertainment, showing a good diversity in vocabulary. Additionally, the Cosine similarity of 0.0358 and Jaccard similarity of 0.07 between the training and testing sets indicate a random and balanced split, and confirm that the dataset is not overly curated

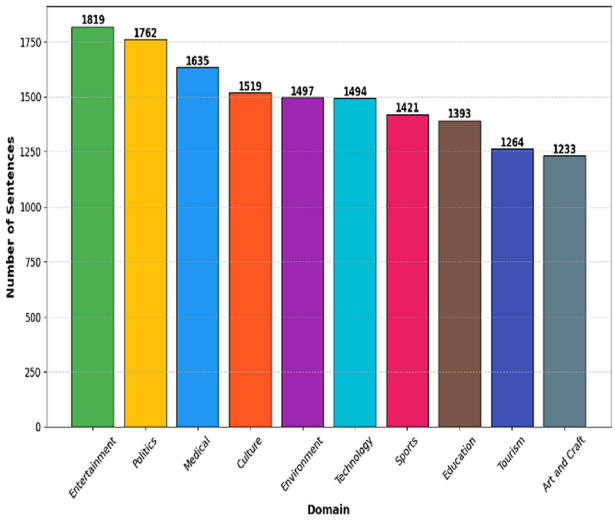


Fig. 2. Distribution of Kashmiri news snippets by domain and number of sentences.

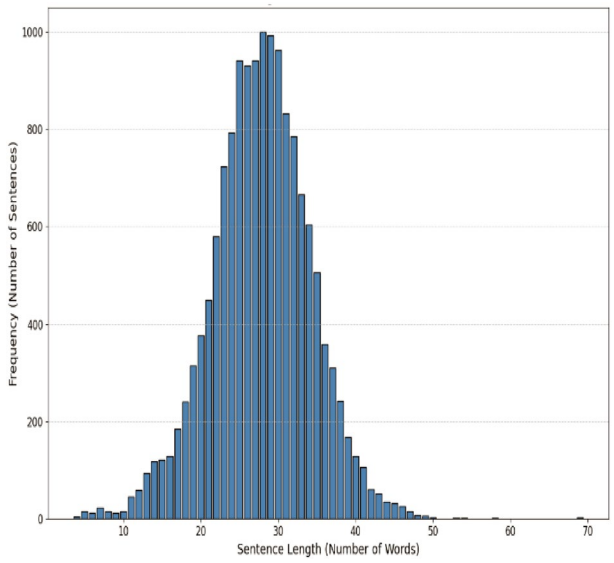


Fig. 3. Distribution of sentence lengths and their frequency across all domains.

Domain	Total words per domain	Maximum no of words per snippet	Minimum no of words per snippet	Average no of words per snippet
Art and Craft	34,483	42	16	27.97
Culture	44,524	69	11	29.31
Education	35,957	42	11	25.81
Entertainment	52,575	53	9	28.90
Environment	41,250	47	14	27.55
Medical	47,019	45	4	28.25
Politics	49,775	44	8	28.25
Sports	39,454	44	4	27.76
Technology	39,983	46	11	26.76
Tourism	32,656	42	4	25.83

Table 1. Word distribution and news snippet length statistics across 10 Domains.

to ease classification. This randomness helps the models face diverse samples during training and perform reliably on unseen data.

Feature representation using contextual word embeddings

The text embeddings of the Kashmiri news snippets are generated using IndicBERT v2 and FastText embedding models. IndicBERT v2 particularly, is useful since it is specially trained on Indian languages, giving it an edge in terms of capturing the nuances of Kashmiri text due to its multilingual pre-training. This makes it effective in handling grammatical and lexical richness. In contrast, FastText learns word-level semantics including sub-word information, which is essential to represent morphologically rich languages like Kashmiri. This is because most words in Kashmiri are obtained from the root form by affixation. These embeddings act as strong input features for subsequent classification models experimented in the proposed approach.

Machine learning and deep learning models

The study employs Logistic Regression, Support Vector Classifier (SVC), AdaBoost Classifier, and a Stacking Classifier. Logistic regression provides a simple and interpretable baseline, while SVC handles high-dimensional feature spaces effectively, making it suitable for embedding-based representations. AdaBoost combines multiple weak learners to improve classification performance. The Stacking Classifier integrates outputs from different models to enhance accuracy. The models are trained and tested using an 80:20 stratified split with IndicBERT v2 and FastText embeddings.

Deep learning models such as LSTM, GRU, RNN, Bi-GRU, and Bi-LSTM are used to capture sequential dependencies in text. These models effectively learn contextual relationships within sentences. Despite the relatively small dataset size, LSTM[35] and GRU handle long-range dependencies well and suit the complex nature of news snippets. Their bidirectional versions process both past and future contexts to strengthen sentence understanding.

Transformer-based models

The approach of using transformer-based models such as mBERT[31], DistilBERT[36], ParsBERT[33], and BERT-Base ParsBERT-Uncased to utilize their pre-trained powers and transformers models have demonstrated superior performance, especially when dealing with low-resource languages such as Kashmiri. The exploration of transformer-based models-mBERT and ParsBERT, are of special relevance because of their pre-training on huge datasets including languages like Urdu, Persian, and Arabic, which share many linguistic similarities with Kashmiri. This shared linguistic heritage can allow a better adaptation to the Kashmiri text of such models for the task of Kashmiri news snippet classification. These pre-trained transformer models are fine-tuned on the training data to adapt them to the specific task of Kashmiri news snippet classification.

Large Language models

Two large language models, BLOOM-560 M and Flan-T5-Base, were experimented with in two setups. Initially, the models were evaluated on the test data of the Kashmiri dataset in a zero-shot scenario. In the second setup, the models were fine-tuned using the ADAM optimizer with the training data and subsequently assessed for their performance on the test data. BLOOM-560 m is a multilingual model pre-trained on the Roots corpus, which contains 4.6% Arabic, 1% Urdu, and Persian and hence it is expected to perform well in classifying Kashmiri text by leveraging patterns from related languages. The discussed models mark an important step to show the capability of advanced language modeling for low-resource languages such as Kashmiri and provide relevant insight into the same or similar tasks in Kashmiri NLP.

Experimental setup and evaluation metrics

Table 2 summarizes the tuned hyperparameters for the different machine learning models, while Table 3 provides a summary of the hyperparameter tuning that is done for the deep learning architectures. Adam optimizer and sparse categorical cross-entropy loss are used to handle multi-class classification efficiently. Tables 4 and 5 give insights on the hyperparameters used for the Transformer and Large Language models, respectively. Transformer models use AutoTokenizer with a smaller batch size and fine-tuned learning rates to manage memory usage while ensuring effective training. LLMs adopt BLOOM and T5 tokenizers with a lower batch size and higher learning rate to balance computational feasibility given the model size and GPU constraints. The hyperparameter values such as batch size, learning rate, and model depth are chosen based on multiple experimental evaluations to achieve a fair balance between performance and computational efficiency. These values such as tokenizer, loss function and optimizer primarly selected from standard settings with adjustments made where appropriate

Model	Hyperparameter	Value/shape
Logistic regression	Penalty	I2
	Max iteration	1000
AdaBoost classifier	Number of estimators	100
SVC	Kernel	Linear
Stacking	Base estimator	AdaBoost, SVC, logistic regression
	Final estimator	Logistic regression

Table 2. Hyperparameters for machine learning Models.

Hyperparameter	Value/shape
LSTM units	128,64,32
GRU units	128,64,32
Simple RNN units	128,64,32
Bi-GRU units	128,64,32
Bi-LSTM units	128,64,32
Loss	Sparse categorical crossentropy
Optimizer	Adam
Epochs	10
Batch size	32

Table 3. Hyperparameters for deep learning Model.

Hyperparameter/setting	Value/description
Tokenizer	AutoTokenizer with max length of 128 tokens
Learning rate	2e-5
Batch size	16
Epochs	10
Evaluation metric	Accuracy, F1-Score
Loss function	Cross-entropy (used by trainer)
Optimizer	AdamW (default in trainer)

Table 4. Hyperparameters for transformer Models.

Hyperparameter/setting	Value/description
Tokenizer	BloomTokenizerFast and T5Tokenizer with max length of 128 tokens
Learning rate	5e-5
Train batch size	4
Epochs	10
Evaluation metric	Accuracy, F1-score
Loss function	Cross-entropy (used by trainer)
Optimizer	AdamW (default in trainer)

Table 5. Hyperparameters for large language models.

based on their suitability for the task. These choices are replicated as they also align with best practices reported in foundational transformer work[32] and recent efficient fine-tuning strategies[46], supporting their effectiveness across various NLP tasks. CUDA usage constraints and available hardware resources guided the selection of lighter configurations to optimize training time without compromising the model's capability.

Model performance evaluation

The study evaluates the models using accuracy, precision, recall, F1-score, and confusion matrices to ensure a thorough performance assessment. The work provides a foundational exploration into the field of text classification for the Kashmiri language and paves the way for future research in this field.

Results

A detailed assessment of a set of different embedding and classifiers on the manually created Kashmiri Language News Snippet dataset, gives a thorough understanding on how effective they are for text classification tasks.

The section is divided into the following subsections:

- Performance analysis of machine learning classifiers with various embeddings for Kashmiri news snippet classification.
- Performance analysis of deep learning classifiers with various embeddings for Kashmiri news snippet classification.
- Comprehensive analysis of transformer models and their performance in Kashmiri news snippet classification task.
- Comparative analysis of large language models using zero-shot prompting and fine-tuning on the Kashmiri News snippet dataset.

Embedding	Classifiers	Precision	Recall	F1	Accuracy
IndicBERT v2	Logistic regression	0.94	0.94	0.94	0.94
	SVC	0.95	0.96	0.95	0.96
	AdaBoost	0.75	0.75	0.75	0.75
	Stacking classifier	**0.96**	**0.95**	**0.95**	**0.96**
FastText	Logistic regression	0.84	0.84	0.84	0.84
	SVC	0.87	0.87	0.87	0.87
	AdaBoost	0.48	0.47	0.44	0.46
	Stacking classifier	0.87	0.87	0.87	0.87

Table 6. Performance comparison of machine learning classifiers with various embeddings for Kashmiri news snippet Classification. Bold values indicate the best-performing model for each embedding or experimental setting

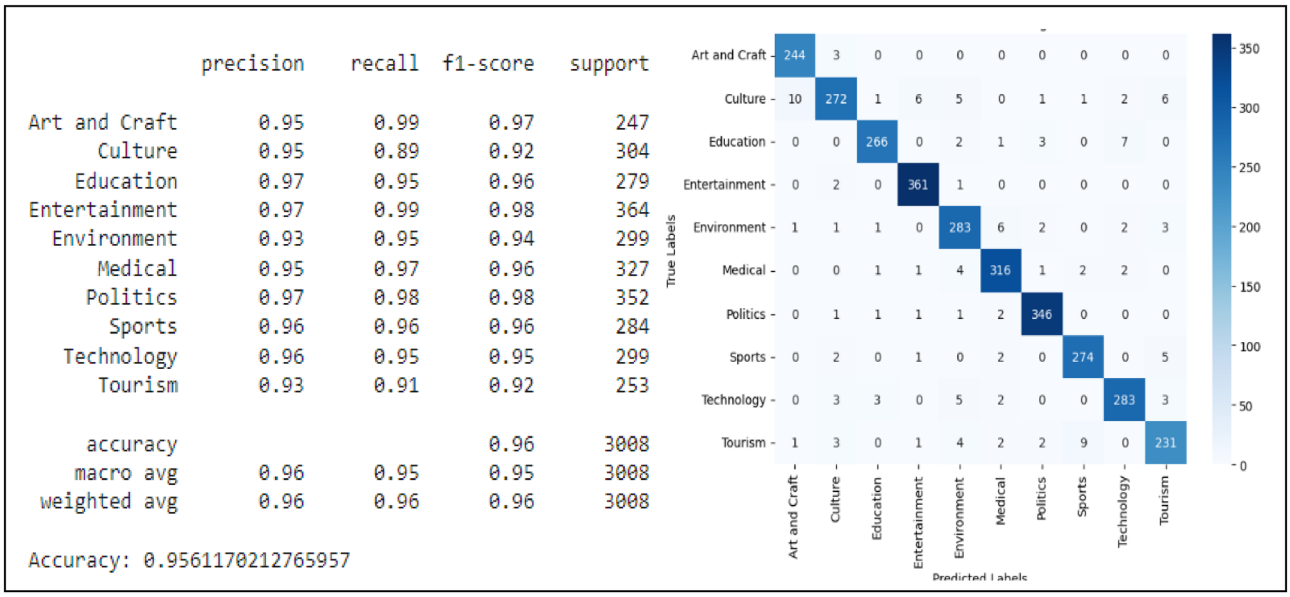


Fig. 4. Classification report and confusion matrix for best performing machine learning framework-stacking classifier using the IndicBERT v2 embeddings on Kashmiri news snippet classification.

Performance analysis of machine learning classifiers with various embeddings for Kashmiri news snippet classification

The machine learning models used in this work are Logistic Regression Classifier, Support Vector Classifier (SVC), AdaBoost Classifier, and a Stacking Classifier combining the strengths of these three classifiers-Logistic Regression, SVC, and AdaBoost. The word embeddings IndicBERTv2 and FastText are used in combination with these models. Logistic Regression and SVC were used with scikit-learn's built-in multi-class support, using softmax and One-vs-Rest strategies respectively.

Performance analysis of deep learning classifiers with various embeddings for Kashmiri news snippet classification

Table 6 shows the performance of these models with the two embeddings - IndicBERTv2 and FastText. The results show the effectiveness of combining word embeddings and classifiers for the multi-class classification task. The Stacking Classifier with IndicBERT v2 embeddings achieves the highest performance, with F1-score of 0.95 and an accuracy of 95.6%. The ML Stacking Classifier with IndicBERT v2 embeddings in Fig. 4 reveal strong performance across all categories, with high F1-scores on domains Politics and Entertainment (0.98 each). Macro and weighted averages for recall, precision, and F1-score all exceed 0.90, demonstrating well-balanced accuracy across domains. High diagonal values on the confusion matrix confirm the efficacy of the model, demonstrating the usefulness of employing ensemble methods with IndicBERT v2 embeddings in text classification on multiple domains in low-resource languages like Kashmiri.

The performance comparison of the various deep learning models—LSTM, GRU, RNN, Bi-GRU, Bi-LSTM in combination with the two word embeddings—IndicBERTv2 and FastText, is given in detail in Table 7.

The classification report and confusion matrix for the GRU classifier using the IndicBERT v2 word embeddings, depicted in Fig. 5, reveal relatively balanced performance across all domains, with precision, recall,

Embedding	Classifiers	Precision	Recall	F1	Accuracy
IndicBERT v2	LSTM	0.92	0.91	0.91	0.91
	GRU	**0.92**	**0.92**	**0.92**	**0.92**
	RNN	**0.92**	**0.92**	**0.92**	**0.92**
	Bi-GRU	0.92	0.91	0.91	0.91
	Bi-LSTM	0.91	0.90	0.90	0.90
FastText	LSTM	0.85	0.84	0.84	0.85
	GRU	0.86	0.86	0.86	0.86
	RNN	0.88	0.88	0.88	0.88
	Bi-GRU	0.88	0.88	0.88	0.88
	Bi-LSTM	0.86	0.86	0.86	0.86

Table 7. Performance comparison of deep learning classifiers with various embeddings for Kashmiri news snippet classification. Bold values indicate the best-performing model for each embedding or experimental setting

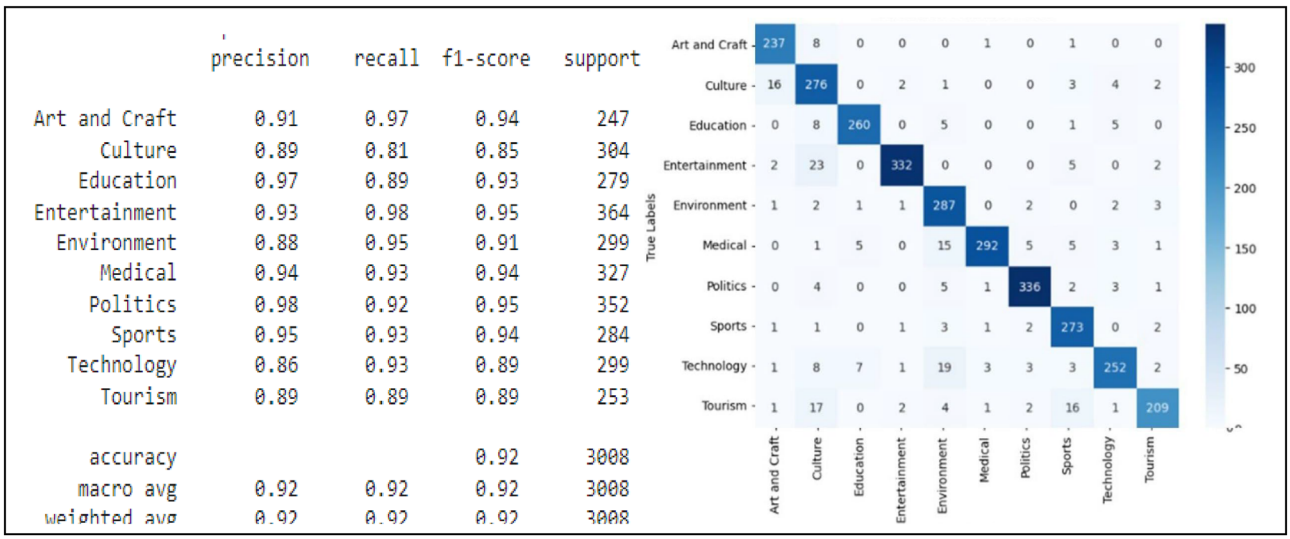


Fig. 5. Classification Report and Confusion Matrix for the best performing DL framework—GRU classifier using the IndicBERT v2 Embeddings on Kashmiri News Snippet Classification.

and F1-scores above 0.80. The GRU model performs well in domains like Politics, Medical, and Sports, showing strong metrics.

The loss curves for the GRU model are depicted in Supplementary Fig. S2.

Overall, the outcomes illustrate the importance of combining GRU with robust contextual embeddings like IndicBERT v2 to achieve high accuracy and well-balanced classification of Kashmiri multi-domain news snippets. The outcomes reflect that deep learning models perform well, but machine learning models, specifically the Stacking Classifier with IndicBERT v2 embeddings, outperform the deep learning models.

Comprehensive analysis of transformer models and their performance in Kashmiri news snippet classification task

The performance comparison of the various transformer models- mBERT Multilingual-BERT-Cased, DistilBERT-Base-Uncased, ParsBERT-v3.0, and BERT-Base-ParsBERT-Uncased, is given in detail in Table 8. The transformer models achieve exceptional performance on the Kashmiri news classification task, with BERT-Base-ParsBERT-Uncased achieving the highest F1-score of 0.98 among all transformer models, and performing strongly across all domains.

The classification report and confusion matrix for the Bert-Base-PARSBERT-Uncased on Kashmiri News Snippet Classification are depicted in Fig. 6. The model achieves perfect precision and recall (1.00) on domains like Politics and Entertainment with near-perfect classification accuracy. It performs reasonably well on other domains, i.e., Art and Craft (0.99), Culture (0.97), and Education (0.98), reflecting the effectiveness of the model for Kashmiri news classification. There are minimal confusions in classes with contextual or lexical overlapping attributes, which the model is not able to fully distinguish. The loss curve for the BERT-Base-PARSBERT-Uncased model are given in Supplementary Fig. S3.

Approach	Model	Precision	Recall	F1	Accuracy
Fine-tuning	Multilingual-BERT-cased	0.97	0.97	0.97	0.97
Fine-tuning	Distil_BERT-base-uncased	0.93	0.93	0.93	0.93
Fine-tuning	ParsBERT (v3.0)	0.94	0.94	0.94	0.94
Fine-tuning	BERT-base-ParsBERT-uncased	**0.98**	**0.98**	**0.98**	**0.98**

Table 8. Performance comparison of transformer models for Kashmiri news snippet Classification. Bold values indicate the best-performing model for each embedding or experimental setting

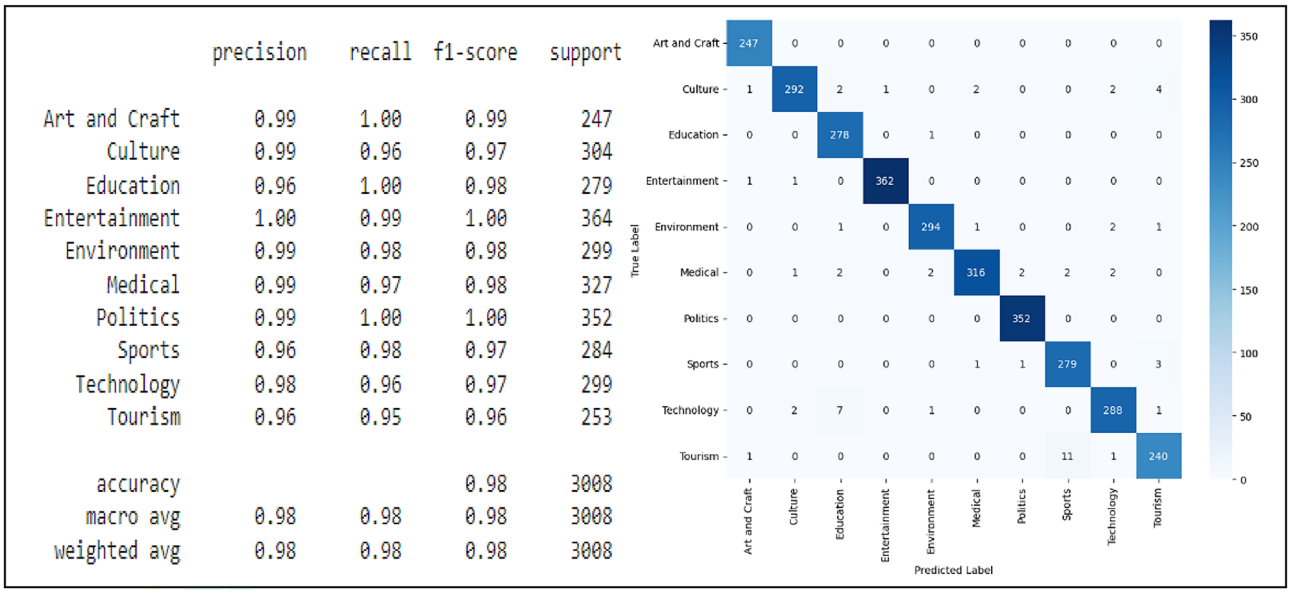


Fig. 6. Classification report and confusion matrix for the best performing transformers model BERT-Base-PARSBERT-Uncased on Kashmiri news snippet classification.

LLM model	Approach	Precision	Recall	F1	Accuracy
Flan-T5-base	Zero-shot prompting	0.01	0.10	0.02	0.09
	Fine-tuning	0.56	0.26	0.25	0.28
BLOOM-560 m	Zero-shot prompting	0.05	0.09	0.04	0.09
	Fine-tuning	**0.97**	**0.97**	**0.97**	**0.97**

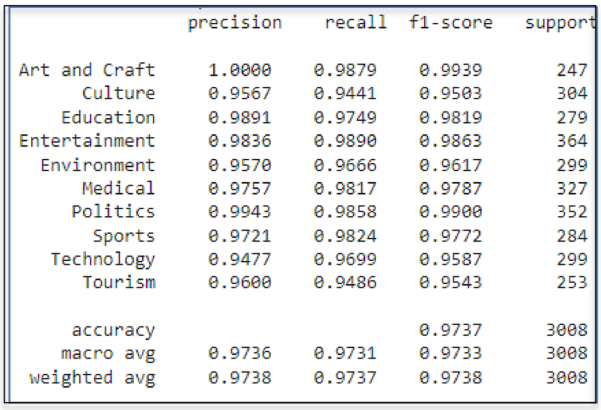
Table 9. Performance comparison of large language models using zero-shot and fine-tuning for Kashmiri news snippet classification. Bold values indicate the best-performing model for each embedding or experimental setting

Multilingual-BERT-Cased attains 0.97 accuracy, with excellent recall and precision (0.97), followed by Distil_BERT-Base-Uncased and ParsBERT (v3.0) with accuracies of 0.930 and 0.940, respectively. They perform well but fall short of BERT-Base-ParsBERT-Uncased.

Comparative analysis of large language models using zero-shot prompting and fine-tuning on the Kashmiri news snippet dataset

The study employs the Large Language Models (LLMs) BLOOM-560 m and Flan-T5-Base for both zero-shot prompting and fine-tuning on the Kashmiri news snippet classification task, and presents their performance metrics in Table 9.

Flan-T5-Base achieves very low accuracy and F1 scores in both fine-tuning and zero-shot settings. In contrast, BLOOM-560 m achieves better performance with an accuracy of 0.973 and consistently high precision, recall, and F1-score of 0.97. However, in zero-shot settings, BLOOM-560 m shows poor performance with a precision of 0.10, recall of 0.08, and F1-score of 0.04, indicating its limited ability to classify Kashmiri news without task-specific fine-tuning. While the model's multilingual pre-training provided general capabilities, the absence of Kashmiri-specific data limited its performance in Zero-shot settings. Fine-tuning BLOOM-560 m on the Kashmiri news dataset significantly improved its F1-score from 0.04 in zero-shot prompting to 0.97.



	precision	recall	f1-score	support
Art and Craft	1.0000	0.9879	0.9939	247
Culture	0.9567	0.9441	0.9503	304
Education	0.9891	0.9749	0.9819	279
Entertainment	0.9836	0.9890	0.9863	364
Environment	0.9570	0.9666	0.9617	299
Medical	0.9757	0.9817	0.9787	327
Politics	0.9943	0.9858	0.9900	352
Sports	0.9721	0.9824	0.9772	284
Technology	0.9477	0.9699	0.9587	299
Tourism	0.9600	0.9486	0.9543	253
accuracy			0.9737	3008
macro avg	0.9736	0.9731	0.9733	3008
weighted avg	0.9738	0.9737	0.9738	3008

Fig. 7. Classification report for the fine-tuned LLM Model-BLOOM-560 m on Kashmiri news snippet classification.

Model	Precision	Recall	F1	Accuracy
ML stacking classifier with IndicBERT v2	0.96	0.95	0.95	0.96
GRU with IndicBERT v2	0.92	0.92	0.92	0.92
BERT-Base-ParsBERT-uncased (finetuned)	0.98	0.98	0.98	0.98
BLOOM-560 m (finetuned)	0.97	0.97	0.97	0.97

Table 10. Performance comparison of best models metrics for Kashmiri news snippet classification.

The classification report for the BLOOM-560 m model depicted in Fig. 7, shows robust and balanced performance across all 10 domains of Kashmiri news snippets. High precision, recall, and F1-scores are observed consistently, with macro and weighted averages of 0.974, highlighting the model's ability to handle multi-class classification effectively. Specific domains, such as Art and Craft, Politics, and Education, achieve near-perfect precision and F1-scores, indicating the model's strong capability to capture distinctive linguistic patterns in these domains. The loss curve for the LLM Model-BLOOM-560 m model is given in Supplementary Fig. S4.

Discussion

The summary of the results given in Table 10, demonstrate the performance of the transformer-based and large language models over general machine learning and deep learning-based approaches on Kashmir news classification.

The strong performance of the Stacking Classifier with IndicBERT v2 embeddings is due to robustness of the Stacking Classifier, which combines the strengths of multiple base models, such as Logistic Regression, SVC, and AdaBoost, into a combined framework. By combining the predictions of these different models, the Stacking Classifier minimizes individual biases or errors, improving overall accuracy and generalization. This makes it highly effective across different domains, even with variation in data characteristics.

The model with IndicBERT v2 embedding performs better in this case, since it is specifically designed and pre-trained on a diverse corpus of Indian languages, which enables it to capture linguistic nuances such as complex grammar, contextual variations even for a low-resource language like Kashmiri. FastText relies on sub-word representations which lack contextual understanding, while IndicBERT uses transformers to create contextualized embeddings, capturing word relationships and meanings effectively for multi-domain text classification.

The results of the various DL classifiers, as given in Table 7, indicate that the GRU and RNN classifiers with IndicBERT v2 obtain the best results, with an F1-score of 0.92 on the test set, showing that these models are good at capturing the contextual nuances and complexities within the dataset, enabling them to classify the text with a high level of precision and recall. GRU and RNN outperform LSTM, Bi-GRU, and Bi-LSTM in Kashmiri classification because Kashmiri text data typically involves shorter sequences and straightforward dependencies. The additional complexity of bidirectional or LSTM models introduces unnecessary overhead, reducing their effectiveness for this specific task. The models combined with FastText embeddings result in noticeably lower performance, with the highest F1-scores being 0.88 for the Bi-GRU and RNN classifiers, due to limited domain-specific contextual understanding, while IndicBERT v2 outperforms with richer, contextualized representations tailored for Indian languages.

However, the GRU model struggles slightly with domains like Culture, Tourism, and Technology, likely due to overlapping or ambiguous content. For example, Culture is often confused with Art and Craft (16 instances), and Tourism overlaps with Culture (17 instances), highlighting challenges in distinguishing similar linguistic patterns.

The results of the various transformer models for Kashmiri News Snippet classification, as given in Table 8, indicate the model's ability to handle the complexities of Kashmiri text, relying on its deeper contextual understanding and pre-training. Kashmiri, while an Indo-Aryan language, has been profoundly shaped by centuries of contact with Persian, Arabic, and Urdu, leading to lexical borrowing, phonological changes, and grammatical borrowing[47]. This historical and linguistic overlap makes models trained on Persian, Arabic or similar language corpora more capable of generalising to Kashmiri. The combination of domain-specific training and high-dimensional embeddings enables BERT-Base-PARSBERT-Uncased to maintain consistent performance across all metrics, making it the most suitable model for this task among all the transformer-based models used in this work.

The performance comparison of LLMs revealed that Flan-T5-Base performs poorly due to its limited domain-specific knowledge. Flan-T5 primarily focuses on solving instruction-based tasks[48] in high-resource languages, which does not align well with the requirements of low-resource language classification. This weakness is attributed to Flan-T5's pretraining and fine-tuning data. The original T5 was trained primarily on the Colossal Clean Crawled Corpus (C4), which is composed almost entirely of English web text[49]. Subsequent instruction-tuning of Flan-T5 relied heavily on English datasets, limiting its capacity to transfer to low-resource languages. Since the model lacks training on linguistically relevant languages like Persian and Arabic, which share closer ties to Kashmiri, it struggles to model Kashmiri text. Furthermore, Flan-T5's SentencePiece tokenizer struggles with Unicode Perso-Arabic scripts, often producing < unk > tokens for Kashmiri characters. As a result, the embeddings do not accurately capture the input text, leading to degraded classification performance.

In comparison, BLOOM benefits from its multilingual pretraining on the ROOTS corpus, which includes 4.6% Arabic, 1% Urdu, and several Indic languages[50]. Since Kashmiri shares lexical and grammatical features with these languages, BLOOM already has prior exposure to linguistically relevant data. Its byte-level BPE tokenizer preserves all Unicode characters accurately, allowing embeddings to faithfully represent the Kashmiri text. This enables the model to learn meaningful patterns during fine-tuning, directly contributing to its superior classification performance, as reported in Table 9. This demonstrates the importance of task-specific adaptation and fine-tuning in low-resource language tasks.

Overall, the analysis of the study's results shows that stacking offers consistent improvements by utilizing the complementary strengths of multiple classifiers. Deep learning approaches demonstrate the potential to capture complex patterns but require larger datasets to reach their full effectiveness, in low-resource settings like Kashmiri, they tend to underperform. In contrast, simpler machine learning methods remain more stable under such conditions. Transformer-based models, BERT-Base-PARSBERT-Uncased performs strongly due to its pre-training on linguistically relevant corpora for Kashmiri, which enables the model to capture the intricate linguistic features of the language effectively. Fine-tuning BLOOM-560 m on this dataset allows it to making use of its multilingual background, enabling it to handle Kashmiri text more effectively. BLOOM-560 m performs well for Kashmiri classification because of its exposure to a wide variety of morphologically rich and linguistically related languages during pre-training on the ROOTS corpus. Additionally, the tokenizer's ability to preserve all Unicode characters ensures that embeddings accurately represent the text, enabling BLOOM to recognize shared linguistic patterns from related languages and process Kashmiri more effectively.

The analysis across the 10 domains highlights domain-specific performance trends for the Kashmiri news snippet classification task as shown in Supplementary Fig. S5. Domain-wise analysis reveals that categories like Art and Craft and Entertainment are easier to classify because their content is more distinct. In contrast, Culture and Tourism pose challenges due to overlapping features, making it harder for models like GRU to perform well. In Education and Politics, BLOOM-560 m and transformer-based models like BERT-Base-PARSBERT-Uncased outperform others, showcasing their strength in handling detailed contextual information, while GRU trails in precision. For Technology, the Stacking Classifier and BLOOM-560 m excel, while GRU struggles with lower precision, likely due to more technical and domain-specific terms. Lastly, the Medical and Sports domains see strong performance across all models, but transformer models lead slightly, reinforcing their ability to generalize across diverse contexts. Overall, transformer models generalize well across different domains, showing that their design and pre-training make them better suited for handling linguistic diversity in Kashmiri news classification.

Conclusion and future scope

This work emphasizes overcoming data scarcity challenges in text analytics for the low-resource language, Kashmiri, through a manual approach to generate the Kashmiri dataset, ensuring its quality and authenticity. To the best of the authors' knowledge, this is the first dataset for Kashmiri text classification, marking it a significant contribution to NLP research for low-resource languages. By focusing on different domains like Medical, Politics, Sports, Tourism, Education, Art and Craft, Environment, Entertainment, Technology, and Culture, this work helps to understand the efficacy of the classification models in these areas and usefulness for this specific task of Kashmiri news snippet classification. This study highlights the superiority of IndicBERTv2 embedding when paired with both traditional machine learning as well as deep learning models, showing the effectiveness of various embedding techniques and classification models for a largely under-resourced language in text analysis, like Kashmiri. Additionally, fine-tuned ParsBERT Uncased emerges as the best transformer model, achieving remarkable accuracy and F1 score of 0.98, while the fine-tuned BLOOM-560 m model demonstrates exceptional performance among Large Language Models, highlighting their potential to significantly advance NLP applications for the Kashmiri language. The proposed model can be utilised for developing educational materials and resources by identifying key topics and organizing content in a way that supports structured learning in low-resource languages which contributes towards the goal of quality education.

The study faced several limitations due to the scarcity of digital resources for the Kashmiri language, which demands significant manual effort in data creation, translation, and labeling. Most available resources remain confined to old books, libraries, and archives, while some newspaper data exists only in image formats, making

extraction difficult without reliable OCR tools unfortunately, no standard OCR models currently exist for Kashmiri. Additionally, the absence of pre-trained models limits the scope for directly fine-tuning existing NLP architectures.

Future work can focus on expanding the dataset, automating data collection, applying advanced NLP techniques to improve model performance and exploring cross-domain transfer learning to achieve accurate and resilient models, especially for data collection and classification tasks. Extending these approaches to other low-resource languages will contribute towards building more inclusive multilingual NLP systems, ultimately reducing technological bias.

Data availability

The datasets generated and analysed during the current study are available in the GitHub repository, [https://github.com/DeheemBhat/Multiclass-Classification-of-Kashmiri-News-Snippets-Dataset-Creation-and-Comparative-Evaluations.git](https:/github.com/DeheemBhat/Multiclass-Classification-of-Kashmiri-News-Snippets-Dataset-Creation-and-Comparative-Evaluations.git)The dataset shared here is our complete dataset, consisting of 15,036 snippets, and the code is provided for reference. The full dataset is publicly available on GitHub.

Received: 24 January 2025; Accepted: 14 October 2025
Published online: 19 November 2025

References

1. Wikipedia. Kashmiri language. Accessed (2025). https://en.wikipedia.org/wiki/Kashmiri_language
2. Koshur.org. Accessed (2025). https://www.koshur.org/.
3. Wikipedia. Dardic languages. Accessed (2025). https://en.wikipedia.org/wiki/Dardic_languages
4. Omniglot Accessed. Kashmiri Alphabet, Pronunciation & Language. (2025). https://omniglot.com/writing/kashmiri.htm
5. Snedden, C. Understanding Kashmir and Kashmiris. *Choice Rev. Online.* **53**, 8. https://doi.org/10.5860/choice.195226 (2016).
6. Wikipedia Accessed. Scheduled Languages of India. (2025). https://en.wikipedia.org/wiki/Languages_of_India#Scheduled_languages
7. Ministry of Law and Justice (Legislative Department). The Jammu and Kashmir Official Languages Act. (2020). https://prsindia.org/files/bills_acts/bills_parliament/2020/Jammu%20and%20Kashmir%20Official%20Languages%20Act,%202020.pdf (2020).
8. OneIndia. Kashmiri made compulsory subject in schools. (2008). https://www.oneindia.com/2008/11/01/kashmiri-made-compulsory-subject-in-schools-1225558278.html
9. Lone, N. A., Giri, K. J. & Bashir, R. Natural language processing resources for the Kashmiri language. *Indian J. Sci. Technol.* **15**, 2275–2281 (2022).
10. Thukroo, I. A. & Bashir, R. Spoken language identification system for Kashmiri and related languages using mel-spectrograms and deep learning approach. In *Proc. 7th Int. Conf. Signal Process. Commun. (ICSC)* 250–255 (IEEE 2021).
11. Mehta, M. et al. Hindi text classification: a review. In *Proc. 3rd Int. Conf. Adv. Comput. Commun. Control Netw. (ICAC3N)* 839–843 (IEEE 2021).
12. Aggarwal, S., Kumar, S. & Mamidi, R. Efficient multilingual text classification for Indian languages. In *Proc. Int. Conf. Recent Adv. Nat. Lang. Process. (RANLP)* 19–25 (2021).
13. Qumar, S. M. U., Azim, M. & Quadri, S. M. K. Emerging resources, enduring challenges: a comprehensive study of Kashmiri parallel corpus. *AI Soc.* **40**, 2385–2403 (2024).
14. Li, L. et al. Zero-resource knowledge-grounded dialogue generation. *Adv. Neural Inf. Process. Syst.* **33**, 8475–8485 (2020).
15. Hudeček, V. Low-resource methods for dialogue systems applications (2024).
16. Bustamante, G., Oncevay, A. & Zariquiey, R. No data to crawl? Monolingual corpus creation from PDF files of truly low-resource languages in Peru. In *Proc. 12th Lang. Resour. Eval. Conf.* 2914–2923 (2020).
17. King, B. P. Practical natural language processing for low-resource languages. Ph.D. thesis (2015).
18. Fesseha, A. et al. Text classification based on convolutional neural networks and word embedding for low-resource languages. *Tigrinya Inf.* **12**, 52 (2021).
19. Yu, W. et al. Application of quantum recurrent neural network in low resource Language text classification. *IEEE Trans. Quantum Eng.* **5**, 1–15(2024).
20. Cruz, J. C. B. & Cheng, C. Establishing baselines for text classification in low-resource languages. (2020). arXiv:2005.02068.
21. Marivate, V. et al. Investigating an approach for low resource Language dataset creation, curation, and classification: Setswana and sepedi. (2020). arXiv:2003.04986.
22. Grießhaber, D., Vu, N. T. & Maucher, J. Low-resource text classification using domain-adversarial learning. *Comput. Speech Lang.* **62**, 101056 (2020).
23. Li, X. et al. Springer,. Low-resource text classification via cross-lingual language model fine-tuning. In *China Natl. Conf. Chinese Comput. Linguist.* 231–246 (2020).
24. Adelani, D. I. et al. Masakhanews: news topic classification for African languages. (2023). arXiv:2304.09972.
25. Santhanalakshmi, S. News article topic classification using embeddings. In *Proc. 14th Int. Conf. Comput. Commun. Netw. Technol. (ICCCNT)* 1–7 (IEEE 2023).
26. Jahnavi, M. et al. Classification of news category using contextual features. In *Proc. Int. Conf. Knowl. Eng. Commun. Syst. (ICKECS)* 1–7 (IEEE 2024).
27. Sheshadri, S. K., Gupta, D. & Costa-Jussá, M. R. Neural machine translation for Kashmiri to English and Hindi using pre-trained embeddings. In *Proc. 2022 OITS Int. Conf. Inf. Technol. (OCIT)* 238–243 (IEEE, 2022).
28. Agbesi, V. K. et al. Pre-trained transformer-based models for text classification using low-resourced Ewe language. *Systems* **12**, 1 (2023).
29. Alam, T., Khan, A. & Alam, F. Bangla text classification using transformers. arXiv:2011.04446 (2020).
30. Mirashi, A. et al. L3Cube-IndicNews: news-based short text and long document classification datasets in indic languages. (2024). arXiv 2401.02254.
31. Wu, S. & Dredze, M. Are all languages created equal in multilingual BERT? In *Proc. 5th Workshop Represent. Learn. NLP* 120–130 (Association for Computational Linguistics, 2020).
32. Devlin, J., Chang, M. W., Lee, K. & Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proc. 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol.* 1, 4171–4186 (2019).
33. Farahani, M., Gharachorloo, M., Farahani, M., Manthouri, M. ParsBERT Transformer-based model for Persian Language Understanding. *Neural Process. Lett.* **53**, 3831–3847 (2021).
34. Wertz, L. When few-shot fails: Low-resource, domain-specific text classification with transformers (2024).
35. Arora, G. inltk: Natural language toolkit for Indic languages. (2020). arXiv:2009.12534.

36. Nair, A. R. et al. Evaluating the impact of text data augmentation on text classification tasks using distilbert. *Procedia Comput. Sci.* **235**, 102–111 (2024).
37. Khuntia, M. & Gupta, D. Indian news headlines classification using word embedding techniques and LSTM model. *Procedia Comput. Sci.* **218**, 899–907 (2023).
38. Subhash, P. M. et al. Fake news detection using deep learning and transformer-based model. In *Proc. 14th Int. Conf. Comput. Commun. Netw. Technol. (ICCCNT)* 1–6 (IEEE 2023).
39. Nair, A. R. et al. Comparative analysis of word embeddings for text classification in Spark NLP. In *Proc. IEEE Int. Conf. Cloud Comput. Emerg. Mark. (CCEM)* 130–136 (IEEE, 2023).
40. Sun, X. et al. Text classification via large Language models. (2023). arXiv:2305.08377.
41. Cahyawijaya, S. et al. LLMs are few-shot in-context low-resource language learners. (2024). arXiv:2403.16512.
42. Patwa, P. et al. Enhancing low-resource LLMs classification with PEFT and synthetic data. (2024). arXiv:2404.02422.
43. Joshi, S. et al. Fine tuning LLMs for low resource languages. In *Proc. 5th Int. Conf. Image Process. Capsule Netw. (ICIPCN)* 511–519 (IEEE 2024).
44. Jadhav, S. et al. On limitations of LLM as annotator for low resource languages. arXiv:2411.17637 (2024).
45. Boyina, K. et al. Zero-shot and few-shot learning for Telugu news classification: A large language model approach. In *Proc. 15th Int. Conf. Comput. Commun. Netw. Technol. (ICCCNT)* 1–7 (IEEE 2024).
46. Simoulin, A., Park, N., Liu, X. & Yang, G. Memory-efficient fine-tuning of transformers via token selection. In *Proc. 2024 Conf. Empirical Methods Nat. Lang. Process.* 21565–21580 (Association for Computational Linguistics, Miami, 2024).
47. Modares Journal of Humanities. Language contact between Kashmiri, Persian, and Arabic. https://lrr.modares.ac.ir/article-14-4094-en.html (Accessed 2025).
48. Longpre, S. et al. The flan collection: designing data and methods for effective instruction tuning. In *Proc. Int. Conf. Mach. Learn.* 22631–22648 (PMLR, 2023).
49. Raffel, C. et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* **21**, 1–67 (2020).
50. Laurençon, H. et al. The bigscience roots corpus: a 1.6 TB composite multilingual dataset. *Adv. Neural Inf. Process. Syst.* **35**, 31809–31826 (2022).

Acknowledgements
The authors would like to acknowledge the support of Tabasum, a former teacher, for her contributions to data creation and manual verification of Kashmiri translations.

Author contributions
Conceptualization by D.G. M.V, P.C.N, Methodology by D.U.D, A.R, D.G, M.V, P.C.N, Model Implemented by D.U.D, A.RManuscript drafted by D.U.D, A.R, Critically reviewed by, D.G. M.V, P.C.N and Supervised by D.G, M.V.

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-24451-4.

Correspondence and requests for materials should be addressed to D.G. or M.V.