Physical Education, Health and Social Sciences
https://journal-of-social-education.org
E-ISSN: 2958-5996
P-ISSN: 2958-5988

Text Classification Performance Analysis through Machine Learning and Deep Learning on the Low-Resourced Balochi Language

Muhammad Ameen Chhajro [1], Sohaib Raheem [2], Farhan Bashir Shaikh [3], Muhammad Hibatullah Channa[4], Fakhira Tabassum[5], Adnan Jahangir Panhwar [6]

[1] Department of Software Engineering, Sindh Maressatul University, Karachi. Email: ameen.chhajro@smiu.edu.pk
[2] Sindh Maressatul University,Karachi. Email: sohaibraheem27@gmail.com
[3] Department of Computer Science, The university of Larkano. Email: Farhan@uolrk.edu.pk
[4] Department of Computer Sciences & Related Studies Hyderabad Institute for Technology & Management Sciences, Hyderabad. Email: hibatullah@hitms.edu.pk
[5] Department of Computer Sciences NUML University Hyderabad Campus. Email: fakhira.tabassum@numl.edu.pk
[6] Department of Computer Science the University of Sindh, Jamshoro. Email: adnanpanhwar1@gmail.com

DOI: https://doi.org/10.63163/jpehss.v4i1.1260

Abstract:
Text classification is a crucial task in Natural Language Processing (NLP). The purpose of text classification research is to classify the text into pre-defined classes automatically. Low-resource languages still receive less attention in NLP tasks due to the scarcity of publicly annotated datasets and computational resources. Similarly, Balochi, a low-resource language with a 2500-year history and cultural significance, has not been considered much for the development of NLP applications. This research study implements a text classification task in Balochi and compares machine learning, Deep Learning, and Transformer-based models. Balochi-language's unlabelled dataset of approximately 5.5k sentences was collected, and various pre-processing techniques, including tokenization, stop words removal, and text normalization, were applied. The experimental results of this research conclude that, among machine learning models, the SGD classifier achieved the highest accuracy of 98.83%. Among Deep Learning models, the BiLSTM achieved the highest accuracy of 98%. However, the Transformer-based model, the pre-trained XLM-RoBERTa, performed exceptionally well, achieving 99% accuracy on the Balochi classification task. These research findings provide a foundation for future multilingual pre-trained models for low-resource languages and aim to develop consistent Balochi language models for NLP applications.

Keywords: Text Classification, Machine learning, Deep Learning, XLM-RoBERTa, Balochi Language, NLP, Low-Resource Languages

Introduction

Natural Language Processing (NLP) has witnessed significant advancements, specifically in high-resource languages with plentiful annotated datasets, aiding accurate text classification, sentiment analysis, and other subsequent tasks. Text classification has been adopted in my applications, including news categorization, spam detection, content moderation, sentiment analysis, and more [1]. On the other hand, low -resource languages are still understudied in NLP research because of

limited datasets, a shortage of validated tools, and insufficient linguistic resources, which hinder the performance of standard machine learning and deep learning approaches. These low-resource languages have complex grammatical structures, diverse and vast vocabularies, and the connected social contexts are unique, which creates further complexities and challenges for reliable and robust models [2]. Recurrent Neural Networks, along with Long Short Term Memory (LSTM) has drawn attention in many language tasks nowadays, which range from language modelling, text sequence modeling, and translation [3]. The Balochi language, with a history of 2500 years, is a prominent example of a low-resource language. Despite its cultural and historical significance, Balochi language datasets are still publicly unavailable, which limits and restricts the development, evaluation, and analysis of NLP applications. The language pre-trained has been witnessed to improve the many language tasks in the field of natural language processing [4] . The Large Language models (LLMs) have recently proven significant performance in high-resource settings, yet their performance in low-resource languages like Balochi is substantially limited by dataset paucity, computational costs, and negligible research focus. Current studies often focus on specific tasks like binary classification, leaving many regional languages, including South Asian and Middle Eastern languages, largely unexplored [5] . Recent approaches and strategies, such as dataset translation, zero-shot learning, and multilingual model adaptation such as BERT, and XLM-RoBERTa offer auspicious avenues to overcome these restrictions, limitations, and perform well on low-resource languages [6]. Developing comprehensive NLP applications for these languages is critical not only for technological incorporation but also for retaining linguistic diversity and supporting digitally underrepresented language communities. To address these challenges faced by low-resource languages, this research paper investigates and analyses the performance of various traditional Machine Learning, Deep Learning, and Transformer-based models on Balochi language for the text classification task.

Problem statement

Text classification is a basic and important task of Natural Language Processing (NLP), which is largely used in applications like information retrieval, sentiment analysis, and automated content organization. While text classification has been studied largely in English and other high-resource languages. Low-resource languages such as Balochi are still underexplored in the field of computational linguistics. However, the history of Balochi is approximately 2500 years old, and it's a mother tongue of millions of people. Still, there is very limited work in modern computational research for Balochi.

Literature Review

A recent study has been conducted on sentiment analysis using machine learning and deep learning techniques on Balochi text. The one of the challenge has been addressed by author is the lack of availability of a cleaned dataset in Balochi language, which is more challenging to create advance and robust Balochi language applications in the field of natural language processing [7].The low resource languages in South Asia has always been facing the challenges of computational resources and data, which results the lack of work and development in the field of natural language processing in these langauges.However, the researchers are trying to contribute and highlight the these issues frequently in their studies, like wise a research study has been implemented on Bengali language on text sentiment analysis, this study is based on quantum classical hybrid model where the multilingual BERT model is combined with quantum recurrent neural networks [8]. The experiment utilizes two Bengali text classification datasets. The book-reviews dataset contains 2000 reviews gathered from the internet, and the other dataset is Youtube-B. It contains 11807 comments collected from the YouTube site. The result shows that the BUQRNN achieves a maximum accuracy improvement of 0.993% while reducing average model complexity by 12%. On the other hand, PN-BUQRNN structure surpassed the BUQRNN structure and outperformed classical approach structures. One of the researches implemented on low resource Sindhi language

text data on classification and data was particularly collected from the newspaper articles. The study was performed on a custom dataset of 6156 Sindhi newspaper articles. The data was divided into three categories: entertainment, sports, and technology. The CNN, RNN, LSTM and hybrid CNN-LSTM models were trained; the final results revealed that the CNN and hybrid CNN-LSTM models secured the highest accuracy of 96%, LSTM: 95.85%, and RNN lagged at 67% accuracy only [9]. Furthermore, no detailed analysis on class imbalance, real-world noisy text, and cross-domain generalization was mentioned. Another research study was conducted on a large and well-annotated dataset for Urdu toxic comment classification, like the PURUTT corpus, consisting of 72,771 labelled comments of both Urdu and Roman Urdu scripts. Transfer learning, multilingual large language models, and ensemble techniques were used to get strong results. These studies demonstrate that if adequate datasets and linguistic resources are available, then the NLP models could be effective for low-resource languages [10]. This literature also exhibits that multilingual embedding and transfer learning based approaches could improve the performance in text classification tasks. This work provides a strong foundation for future research. especially for low-resource languages where standardized datasets and NLP tools are limited. The absence of large, annotated datasets makes text classification in low-resource languages a demanding NLP task. Especially for regional languages like Sindhi, where limited digital resources and linguistic complexities exist. The author performed text classification and proposed a modified TF-IDF term weighting scheme on the Sindhi articles collected from Awami Awaz. This study modified the TF-IDF term weighting scheme to provide better discriminative features compared to traditional TF-IDF. Classical machine learning models, especially Random Forest and Logistic regression, achieve up to 97.92% accuracy with modified TF-IDF [11].This study achieved high accuracy with modified TF-IDF, but this approach is limited to word-level-statistical features and does not capture contextual semantics. In morphologically rich languages like Sindhi, only frequency-based weighting can't handle semantic ambiguity properly. Having status of a low-resource language, Amharic faces so many complexities in text classification, such as morphological complexity, limited datasets, and contextual ambiguity. The previous studies have shown that, when text classification models were applied to the Amharic language, the evaluation results showed in [12] that traditional machine learning (SVM), and different deep learning models (CNN, BiLSTM) in [13] achieved classification accuracy results, which ranges from 71-94% .The previous study applied on the integrating semantic understanding and contextual embedding's from the Amharic language. The Transformer-based models were trained especially BERT, improve accuracy by capturing word-level and context-level semantic information. In the result the Fine-tuned BERT models achieve up to 87–97% accuracy in morphologically complex and low-resource languages like Amharic [14].The previous research study has applied the multilingual transformer models like BERT, XLM-RoBERTa-base, and XLM-RoBERTa-large for Bangla text classification. In that study the author experienced with the different datasets including the YouTube sentiment datasets, Bangla news sentiments dataset, and the authorship attribution dataset were evaluated. The research results concluded that XLM-RoBERTa-large model showed superior performance on all datasets, where 60.4% to 93.41% accuracy were achieved. These findings highlighted that in even low-resource language scenarios, large multilingual transformers provide state-of-the-art performance and provide much better results than classical deep learning models [15].Text pre-processing is the important step for any task performed on text data in low resource languages. The previous research has revealed that the pre-processing in Sindhi language is a critical step due to data scarcity and limited tools availability, which directly affect the performance of downstream tasks such as text classification, sentiment analysis, and information retrieval. The study identified the high-frequency stopwords from the language dataset, and the TF-IDF approach was integrated with a rule-based system for POS tagging and used Text Pre-processsing Techniques for Sindhi language (TPTS). The evaluation results exhibited 89% accuracy, which proves TPTS is an

effective and reliable tool for Sindhi text pre-processing and NLP applications [16]. However, it faced limitations and restrictions due to the small dataset, dependency on rule-based methods, and cross-domain evaluation. These limitations highlighted the need to explore big datasets and advanced deep learning or transformer-based approaches. The literature review showed that a research study implemented a multi-text classification approach for Urdu and Roman Urdu news text data, in which 10500 online news articles were divided into 6 categories, including the Accidental, Education, Entertainment, International, Sports, and Weather. That research adopted the different NLP pre-processing techniques, such as data cleaning, balancing, and stop-word removal, and for feature extraction, Count Vector, TF-IDF, and Chi2 methods were applied. For classification, Naive Bayes, Logistic Regression, Random Forest, Linear SVC, and K-Neighbors Classifier were implemented. The comparative analysis showed that Linear SVC achieved the highest accuracy of 96% and performed better than other models [17]. However, this study only used traditional machine learning models, and the dataset was limited to online news. The research on semantic-level context, cross-domain generalization, and multi-language or mixed text data is still a gap in low-resource languages including Sindhi.

A significant research has already been conducted on sentiment analysis and emotional tone, and public opinion from the text data on widely spoken languages like English and Arabic. However, in low-resource languages like Urdu, it is again challenging due to data scarcity and lack cleaned corpus. In spite of this, authors conducted sentiment analysis research on Urdu dataset and trained the Convolutional Neural Network and Recurrent Neural Network, and these models were tuned on various network hypermeters. In result, the RNN surpassed CNN with 91% accuracy and performed well. The research showed that RNN is a reliable option for Urdu sentiment analysis, even the dataset was limited to newspapers and social media comments [18]. There is still a research gap for informal and noisy texts, cross-domain evaluation, transformer-based context-aware models to develop a robust and scalable solutions for Urdu sentiment analysis. The first research on multi-class classification of Uzbek online news articles focused on organizing news content into appropriate categories. For that, data was collected from the Daryo online news platform and then divided into 10 different categories. So that automatically classifies news articles into appropriate classes. For text classification, classical machine learning models like Support Vector Machine (SVM), Random Forest, Logistic Regression, Decision Tree, and Multinomial Naïve Bayes were applied. And for feature extraction, TF-IDF, word-level, and character-level n-gram techniques were used. Which converts meaningful patterns and relationships of text into numerical features. When defining hyper parameters for text classification, 5-fold cross-validation was used. Experimental results showed that support vector machine (SVM) performed well with character-level four-gram features and secured the highest accuracy of 86.88%, which demonstrates that traditional machine learning approaches can work with low-resource languages [19]. This study provides a strong base for future research and also provides guidance for multilingual text classification. In [20] the authors conduct the research on automatic text classification for the Pashto language on a limited dataset. In this study, 800 Pashto documents were collected and labelled into 8 categories (history, technology, sport, cultural, economic, health, politics, and scientific) manually. The spelling, grammar correction, noisy symbol removal, and tokenization were applied in pre-processing. The Unigram and TF-IDF feature extraction techniques were applied, and multiple classifiers were evaluated, including MLP, SVM, KNN, Decision Tree, Naïve Bayes, Random Forest and Logistic Regression. The research study revealed that the MLP classifier gives super performance with TF-IDF features, where average testing accuracy was 94%. Text classification is one of the important tasks of NLP, where text data is classified into predefined categories. A recent research study investigated multi-label news classification for the Uzbek language by using a large language dataset of 513K articles collected from Daryo and different news websites. In this research, different classical bag-of-words models

like Logistic Regression with word/character n-grams, and the deep learning architectures, including Recurrent Neural Network (RNN), Convolutional Neural Network(CNN), and some hybrid models, were evaluated [21]. Moreover, a pre-trained embedding and transformer-based models such as mBERT and BERTbek were also evaluated. The experimental results indicated that the BERTbek model performed well, with the highest F1-score of 85.2% and classification accuracy of 93.4% achieved across the categories. Because of the scarcity of labelled training data, text classification tasks face complexities for low-resource languages. The recent research introduced an Amharic text classification dataset containing 50,000+ news articles and covering 6 categories, local news, international news, sports, politics, business, and entertainment. This research adopted Naïve Bayes classifier as a baseline models, with 62.2% and 62.3% accuracies with count vectorizer and TF-IDF features, respectively. The scope of this study was only limited to classical machine learning models. Whereas, the deep learning and transformer-based architectures were not explored, which are necessary for context-aware and semantically rich text representation [22]. Besides this, the dataset was only limited to formal news text. Due to which model's robustness, scalability, and performance on real-world applications were not evaluated. These limitations highlight the urgent need for advanced, domain-generalizable, and high-performance classification models for low-resource languages.

Document-level Urdu sentiment analysis is a challenging task of NLP. Due to being a low-resource language, diverse viewpoints and the availability of limited annotated data is a challenge in large documents. In recent research, deep learning architectures like BiLSTM, CNN, CNN-BiLSTM and BERT were evaluated for Urdu sentiment analysis. This study proposed a hybrid model, BiLSTM-SLMFCNN, which integrates single-layer multi-filter CNN and BiLSTM. Models were trained and tested on the VTC customer support dataset which was a small dataset, 405 satisfied and 100 unsatisfied calls, and the IMDB Urdu movie reviews dataset (translated from English, 50,000 reviews). The Experimental results obtained from this research concluded that the proposed model gives better performance, whether the data size is small or large. On the other hand, the IMDB large dataset, the highest accuracy and F1-score were achieved (83–94%) [23]. Furthermore, results also showed that deep learning models work effectively on low-resource languages and small datasets, and BiLSTM-SLMFCNN is a robust and reliable option for document-level sentiment analysis.

The Multilingual text classification has achieved significant progress in NLP due to the increasing volume of non-English content on the internet. A previous research study has investigated the multilingual text classification on English, Yoruba, and Hausa language datasets. In this study, a comparative analysis was performed between BERT Transformer, LSTM, and Naïve Bayes models. The research results demonstrated that the BERT Transformer performed well for English and Hausa (89% and 88% accuracy), whereas LSTM was comparatively more effective for the Yoruba language dataset and achieved 74% accuracy [24]. This study highlights that the BERT model is exceptional for multilingual classification. However, traditional models like LSTM and Naïve Bayes could also be viable options in some scenarios. Despite this study, conducted in multilingual text classification, the issues like limited datasets, model overfitting, as in the Yoruba BERT case, and cross-lingual generalization are not fully addressed for low-resource languages. Furthermore, there is a need for research on language-specific fine-tuning and resource-efficient models for African and regional languages.

Table 1. Summary of Literature Review of Various Model's Performance Analysis on Different Languages Datasets

Title	Models Applied	Dataset	Accuracy
Sentiment Analysis of Balochi Text Using Deep Learning [7]	LSTM, GRU	Phrases(1,701)	83.57%, 81.23%

Application of Quantum Recurrent Neural Network in Low-Resource Language Text Classification [8]	BUQRNN	Text (2,000)	99%
Multiclass Text Classifications of Sindhi Newspaper Articles [9]	RNN,CNN,CNN-LSTM	Articles (6,156)	67%,96%,96%, 95.85%
Urdu Toxic Comment Classification With PURUTT Corpus Development [10]	Word2Vec and FastText	Comments (72,771)	91.65%
Enhancing Text Classification in Low-Resource Languages: A Modified TF-IDF Approach for Effective Sindhi Text Categorization [11]	LR, DT, MLP, SVC,RF	Articles (1457)	97.33%,94.65% 97.62%,97.47% 97.47%
Acceptance Deep Learning-based Amharic Text Classification with Semantic Understanding [14]	BERT model	Lines (3,124,561)	90.34%
Bangla Text Classification using Transformers [15]	BERT XLM-RoBERTa-base XLM-RoBERTa-large	Bangla datasets (Sentiments,News)	60.4% to 93.41%
TPTS: Text pre-processing Techniques for Sindhi Language [16]	TF-IDF	Sindhi text (1500)	89%
Multi-text classification of Urdu/Roman using machine learning and natural language pre-processing techniques [17]	Naive Bayes Logistic Regression, Random Forest Linear SVC, KNN	News (10500)	Linear SVC 96%
Sentiment Analysis of Low-Resource Language Literature Using Data Processing and Deep Learning [18]	CNN RNN	Sentences (45,000)	RNN surpassed the CNN,Securing 91%
Multi-Class Text Classification of Uzbek News Articles using Machine Learning [19]	SVM,DTC,RF LR,MNB	Articles (2000)	Highest accuracy 86%
Tuning Traditional Language Processing Approaches for Pashto Text Classification [20]	MLP,SVM,KNN DT,GNB,MNB	Documents (200)	Average testing accuracy 94%
Text classification dataset and analysis for Uzbek language [21]	Rule based models BoW,CNN,RNN BERTbek	Articles (513K)	BERTbek model achieved overall best performance
Text classification algorithms: A survey [22]	Naive Bayes using count vectorizer and Tf-idf features	AMHARIC news (50k)	62.2% 62.3%

Sentiment Analysis of Urdu Text Using Deep Learning Techniques [23]	BiLSTM,CNN CNN-BiLSTM Proposed hybrid model: BiLSTM-SMLFCNN	Movie reviews Urdu (50k)	83% 79% 83% 94%
Machine Learning Algorithms for Multilingual Text Classification of Nigerian local Languages [24]	LSTM	Headlines of Yoruba language (1697)	74%

Proposed Research Methodology

This research paper proposes a framework for the Balochi language for text classification by using machine learning, deep learning, and transformer-based techniques. The overall process flow from data acquisition to model deployment is critical for developing the final solution. The figure 1 illustrates the proposed framework for Balochi language classification task.

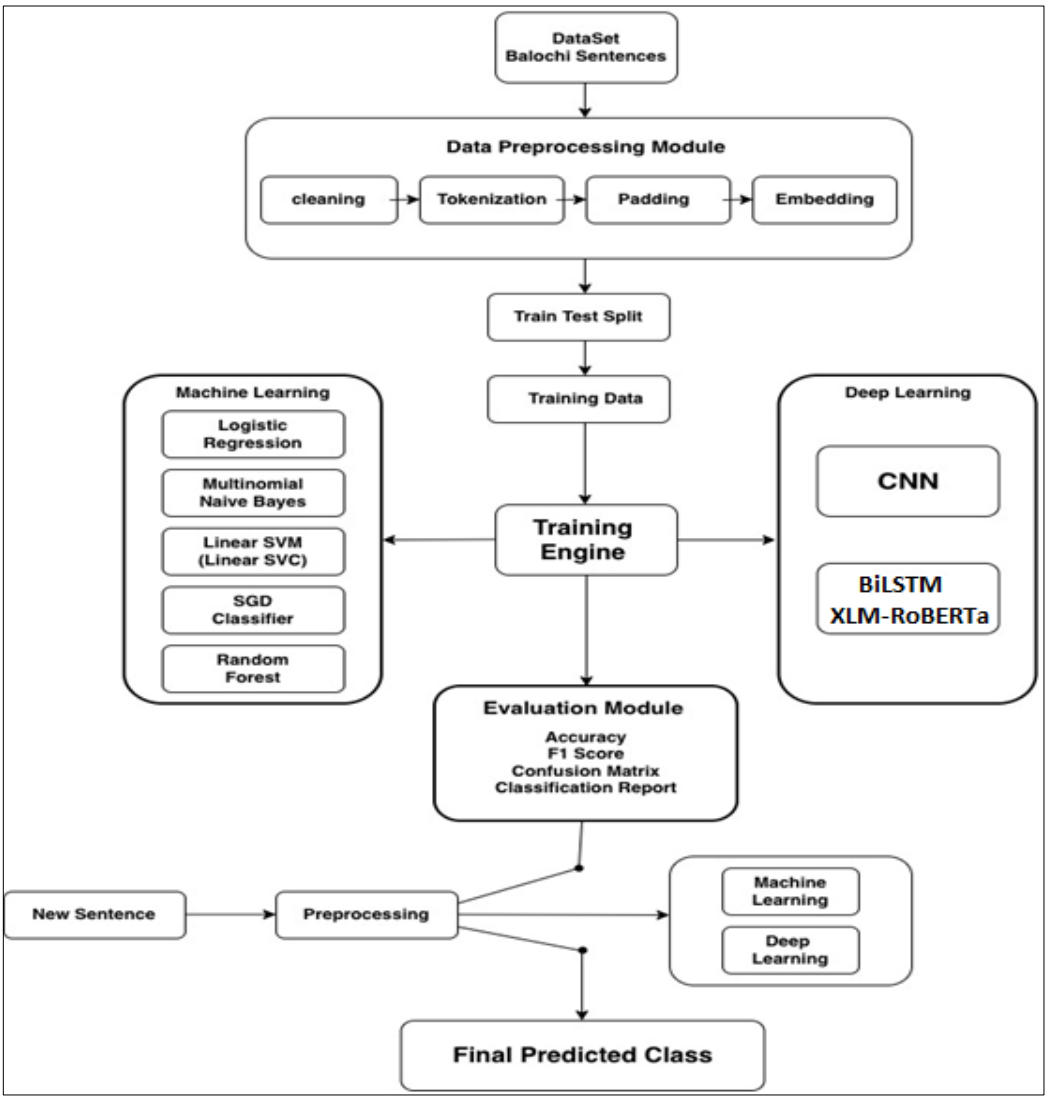

Figure 1. The Proposed Framework for Balochi Language Text Classification

Dataset Collection

Given the scarcity of linguistic resources for Balochi, a proper dataset was searched for

extensively. Among several repositories of data, an unlabeled Balochi text dataset was obtained from the Kaggle platform. This corpus consists of around 5.5k sentences, which is one of the very few publicly available textual resources for Balochi.

Since the dataset did not have labels, the work of manual annotation was pursued for a supervised text classification task. Each sentence was studied carefully and assigned to one of the four predefined categories based on its semantic content: Culture, Social, Religion, and News. The labeling was done manually to make sure that the context is relevant, taking into consideration that in low-resource languages, inconsistencies may occur due to the subjective character of the annotation task. After collection, the Balochi dataset was loaded using Pandas. The dataset consists of text samples along with their class labels. Preliminary exploratory analyses were performed to show insights into the size of the dataset, class distribution, and possible inconsistencies within the entries by checking for missing or noisy entries. That helped to get a grasp on the nature and quality of the data gathered so far.

Text Pre-processing

Text pre-processing is a preparatory step that makes the raw textual data ready for extensive analytics and classification. In the absence of standardized orthography and its digital normalization tools in the Balochi language, the designed pre-processing needed to minimize the noise and maximize the quality of data.To understand the language well we need to mention its alphabets, the foundation Balochi language letters which are available online are in the given figure below

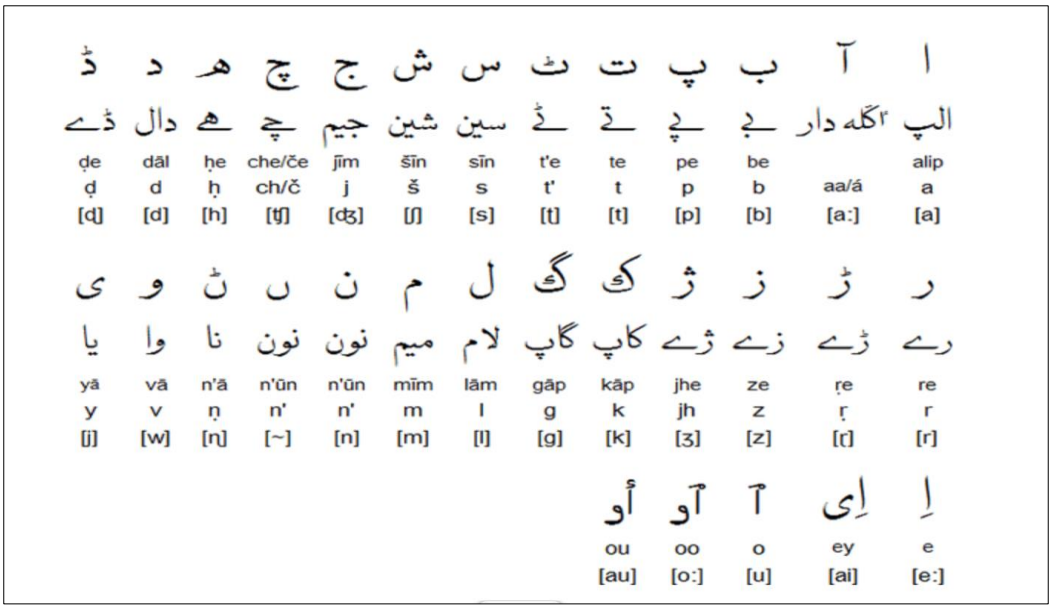

Figure 2..Balochi language alphabets (Source:Omniglot the online encyclopedia of writing systems and languages)

The raw corpus contained lots of irrelevant elements, such as punctuation, special characters, numeric values, URLs, and non-linguistic symbols. These were filtered out by regular expressions while keeping the content featuring linguistically meaningful text only.

Tokenization

Following cleaning, tokenization splits each text sample into its constituent tokens, that is, individual words. Tokenization allows the models to work at a word level; it is important to extract features in most of the machine learning frameworks and to perform sequence modeling in deep learning architectures.

Removing stop words

High-frequency words that do not add much to semantic analysis were filtered out using the NLTK

stop word list. Stop-word removal reduces dimensionality and noise, thereby making the models of classification more efficient as it puts more emphasis on the informational terms. The most common stop words in the Balochi language are listed in the table below

Table 2. Most Common Balochi Language Stop Words

اگر	که	اما	یا	و
در	از	به	پس	چون
من	بر	تا	بی	با
این	شما	ما	او	تو
نه	هم	خود	هر	آن
می	کرد	بود	است	نیست
ء	چرا	کجا	کی	چه
				ئ

In this regard, tokenized text was used to generate sequences of numbers for deep learning models with the Keras Tokenizer. This tokenizer assigns an integer index to each word determined by word frequency in the corpus. Since neural networks take inputs of fixed lengths, the function pad sequences was used to pad the sequences to ensure that input vectors are of the same length.The above steps of pre-processing collectively restructured the unstructured Balochi text into a standardized machine-readable format, apt for machine learning and deep learning–based classification models.

Feature Extraction Technique (TF-IDF)

Feature extraction is an important step in which textual data are transformed into numerical representations that machine learning algorithms can understand. In the present study, meaningful features from preprocessed text were extracted using the TF-IDF technique. TF-IDF performs word weighting, based on the frequency of each word in a document and its inverse frequency throughout the entire corpus. Those words that are contained more frequently within a certain document, yet across other documents rarely, receive higher weights, turning them into more informative terms in classification tasks.

The tokenized documents were then transformed into unigram TF-IDF representations that capture the essential lexical information while controlling much of the impact of commonly occurring words. This resultant TF-IDF feature matrix acts as an input to train and evaluate various machine learning classification models.

TF-IDF enables us to handle high-dimensional textual data; thus far, it has proved to be one of the most reliable feature extraction methods, especially in text classification tasks related to low-resource languages.

Data Splitting

The labeled dataset was divided into training and testing subsets to assess the performance of the machine learning and deep learning models proposed in this paper. The goal of splitting data is to allow the model to learn from one part while assessing their generalization capability on samples that they have not seen. In this research, the applied data was split using the train-test split approach: 80% for training and 20% for test purposes. The splitting was done randomly to keep the proper distribution of all four classes, namely Culture, Social, Religion, and News, in both subsets.That way, there will be less model bias and overfitting, thereby giving a realistic estimate of the classification performance. Although the model parameters have been learned from the training set, the evaluation of performance is done exclusively on the test set using accuracy, precision, recall, and F1-score as standard metrics.

Results and Discussion

After pre-processing, feature extraction, and splitting, a set of classic machine learning models was applied to perform a multi-class text classification task. The models were selected for their strong track record in text classification, especially for low-resource languages. Feature vectors that were computed with TF-IDF on the training data were then used as input for the following classifiers:

As a baseline classifier, logistic regression with TF-IDF features for multi-class classification achieved a test accuracy of 97%. A closer examination indicates that this classifier performed very well on the majority classes, and achieved a macro precision = 0.49, macro recall = 0.28, macro F1 = 0.30, and a weighted F1-score of 0.96. However, the Logistic Regression suggests that it performs well for the majority of the classes, but faces issues in the classification of minor classes due to the unbalancing issue in the Balochi language dataset

We applied a Multinomial Naïve Bayes Classifier with TF-IDF features to classify texts. Our test accuracy achieved 97.2% accuracy. The classifier achieved a precision of 0.97, recall of 1.00, and an F1 score of 0.99. The experimental results clearly states that the overall efficiency of the model for classification task is good with respect to the major classes in the low-resourced dataset.

It used multi-class text classification, employing a Linear Support Vector Machine with TF-IDF features. It reported an overall accuracy of 98.74%. The Linear (SVC) performed well on the classification tasks performed on low resourced Balochi language, and yields a precision of 0.99, recall of 1.00, and an F1-score of 0.99. The model has also outperformed with respect to all classes even on minor categorical data of low-resourced Balochi language. Overall, these results place LinearSVC as the most robust traditional machine learning model in this study and provide strong generalization and effectiveness for text classification in low-resource languages.

We used an SGD classifier with TF-IDF features to classify text into multiple classes. This model achieved a global accuracy of 98.83% and proved to be quite robust and consistent in performance. The overall performance of the SGD classifier is competitive and efficient for low-resource language text classification, comparable to Linear SVM but with significantly faster training and much better scalability.Using the TF-IDF features, the Random Forest Classifier for multi-class text classification achieved an overall accuracy of 98.3%. The overall performance was remarkably good and reached a precision of 0.98, a recall of 1.00, and an F1 of 0.99. The macro-average F1 of 0.71 indicates a decent performance on classes of low-resourced Balochi language dataset, and results altogether indicate that Random Forest is useful for classification tasks in low-resourced languages.

Table 3. Comparison Analysis of Machine Learning Models

Model	Accuracy	Macro Precision	Macro Recall	Macro F1-score	Weighted F1-score
Logistic Regression	0.970	0.49	0.28	0.30	0.96
Multinomial Naïve Bayes	0.972	0.24	0.25	0.25	0.96
Linear SVM (LinearSVC)	0.987	1.00	0.72	0.80	0.99
SGD Classifier	0.988	1.00	0.75	0.83	0.99
Random Forest	0.983	1.00	0.60	0.71	0.98

We implemented models to learn the subtle feature sets and sequence correlations in Balochi texts. Three deep learning models were implemented including, the Convolutional Neural Network (CNN), a Recurrent Neural Network (RNN), specifically a BiLSTM, and the XLM-RoBERTa. We trained these models on tokenized and padded sequences in a labelled dataset, mapping text to numerical representations using word embedding's. we then trained the models on the dataset of

switching epochs number between 10 to 30 due to resources limitations and checked models performances via accuracy, f1-scores and loss results. The experimental results achieved from these models are discussed in the following sections of this research paper. The CNN utilized a multi-channel architecture with kernels of different sizes 2x2, 3x3, and 5x5. All categories went through max-pooling, and the outputs were concatenated into a global feature representation. Classification of the outputs into four categories including, culture, news, religion, and social was performed through fully connected dense layers with SoftMax activation function. The CNN models achieved with an o accuracy of 98%, the accuracy and loss results of CNN model are mentioned in the figure 3 below.

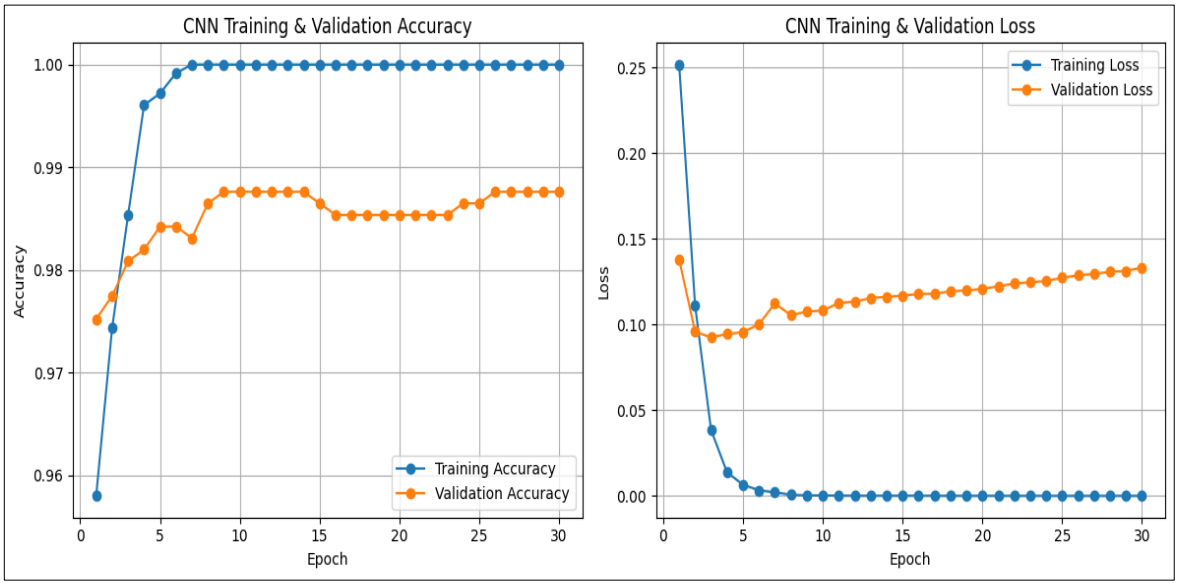

Figure 3. Training and Validation Accuracy and Loss Results of CNN on Balochi Language

The RNN consists of a bidirectional LSTM layer with 100 units. The output of the RNN was combined using both the max pooling and average pooling techniques, followed by dense layers with a SoftMax activation function for the final classification task. The overall accuracy for CNN stood at 98%. The accuracy and loss results of BiLTSM model are mentioned in the figure 4 & 5 below.

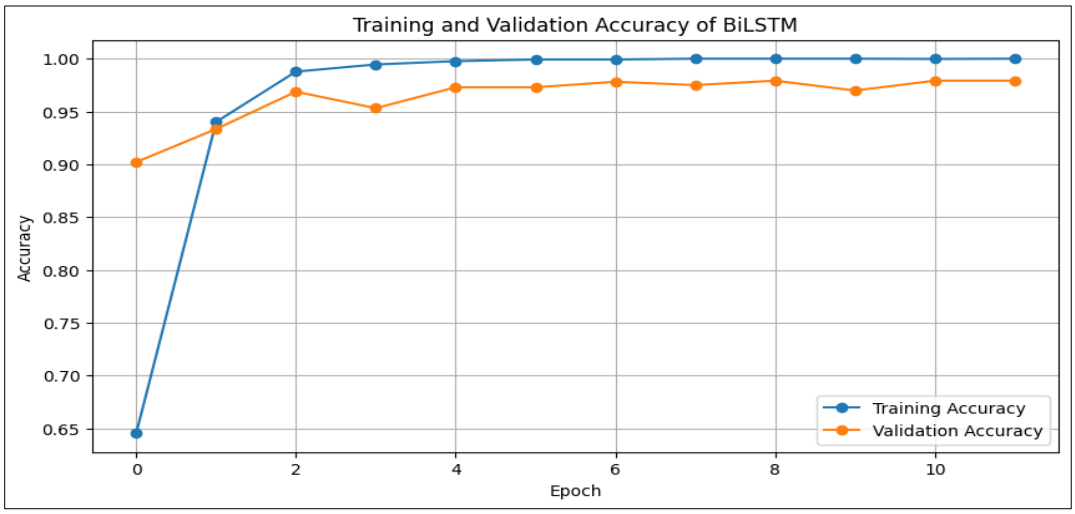

Figure 4. Training and Validation Accuracy Results of BiLSTM on Balochi Language

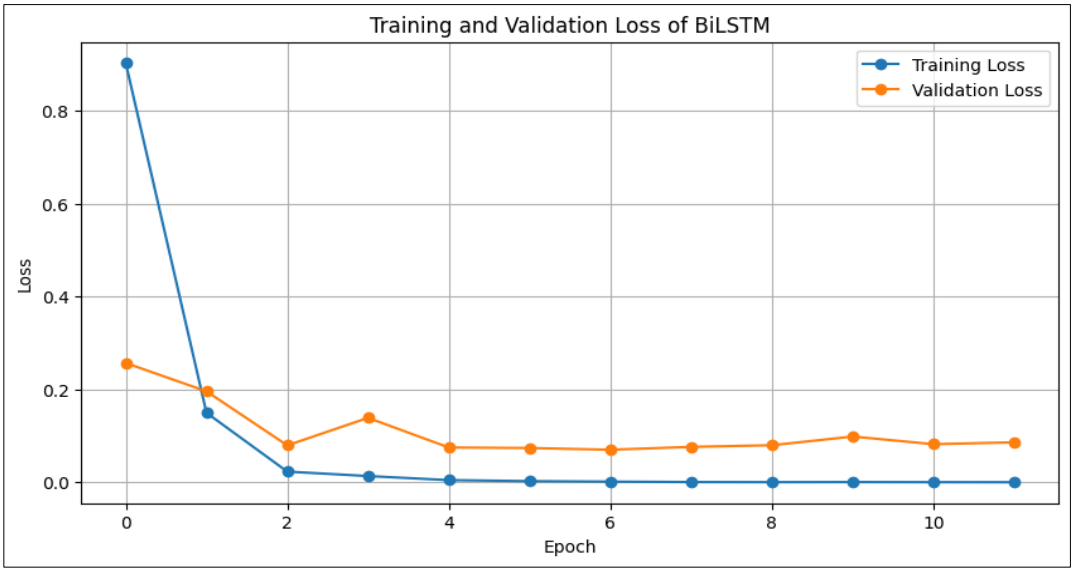

Figure 5. Training and Validation Loss Results of BiLSTM on Balochi Language

Accuracy curves above indicate that the CNN converges rapidly; the training accuracy reaches almost 100% within the first couple of epochs, which means that this model is strongly capable of learning discriminative features from the training data. On the contrary, the validation accuracy gets stabilized at around 98-98.3% with minor fluctuations, reflecting that performance on unseen data improves initially but does not follow the continued rise in training accuracy. This gap in training and validation accuracy reflects the influence of data imbalance and limits generalization gains beyond early epochs.

Loss curves further prove this observation. Training loss decreases sharply and approaches zero, confirming effective optimization on the training set. However, validation loss decreases only during the initial few epochs and then gradually increases, clearly indicating overfitting. Despite achieving very low training loss, the model does not result in corresponding gains in validation performance. These trends underline the fact that while the CNN exhibits very good learning ability, careful regularization together with imbalance-handling strategies will be required to enhance generalization and ensure robust performance on imbalanced text classification datasets.

Although CNN achieves high accuracy due to class imbalance, its macro-F1 score is significantly lower. BiLSTM improves minority class performance, while XLM-RoBERTa achieves the highest accuracy and macro-F1, demonstrating the benefit of pertained multilingual transformers for Balochi text classification.

Table 4. The comparison analysis of Accuracy and Macro F1 of Deep Learning Models

Model	Accuracy	Macro F1
CNN	0.97	0.65
BiLSTM	0.98	0.87
XLM-R	0.99	0.92

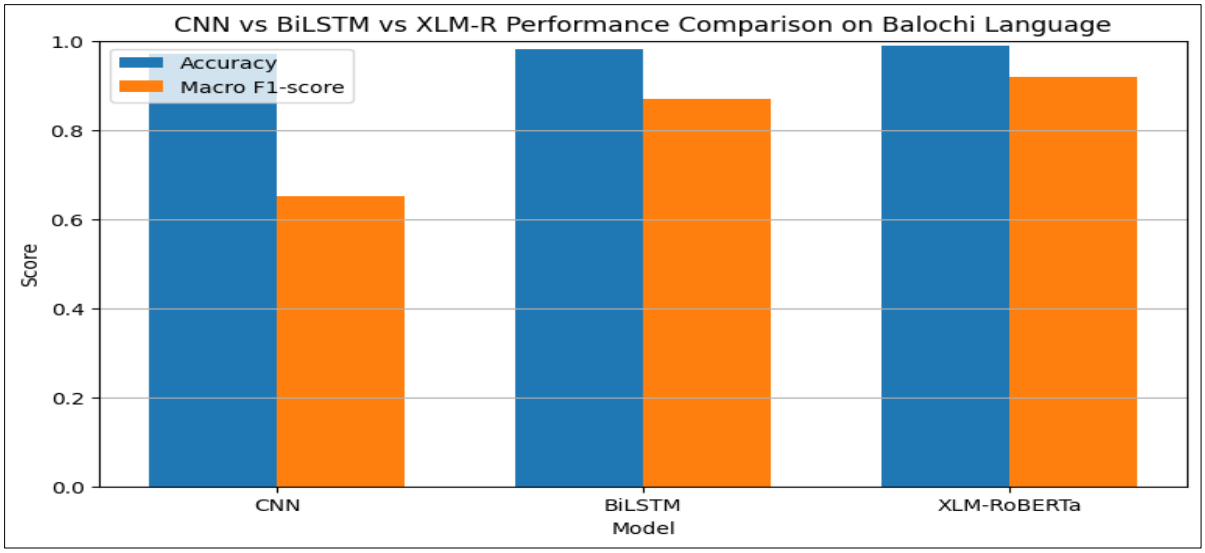

Figure 6. CNN vs BiLSTM vs XLM-R Performance comparison on Balochi Language

The social class prediction retained by CNN model with an F1 score of 0.98, but the other minority classes culture, news, religion was not impressive, with F1 scores per class are 0.65,0.35,0.40 as mentioned in table 4. However, in comparison BiLSTM performed better then CNN in per class F1 scores comparison. Finally, the XLM-RoBERTa performed exceptionally as mentioned in the table 4 and figure 7 below.

Table 5. The Comparison Analysis of Per class F1-Scores of Deep Learning Models

	Per Class F1-scores Comparison		
Class	**CNN**	**BiLSTM**	**XLM-R**
Culture	0.67	0.88	0.92
News	0.35	0.69	0.82
Religion	0.40	0.95	0.97
Social	0.98	0.98	0.99

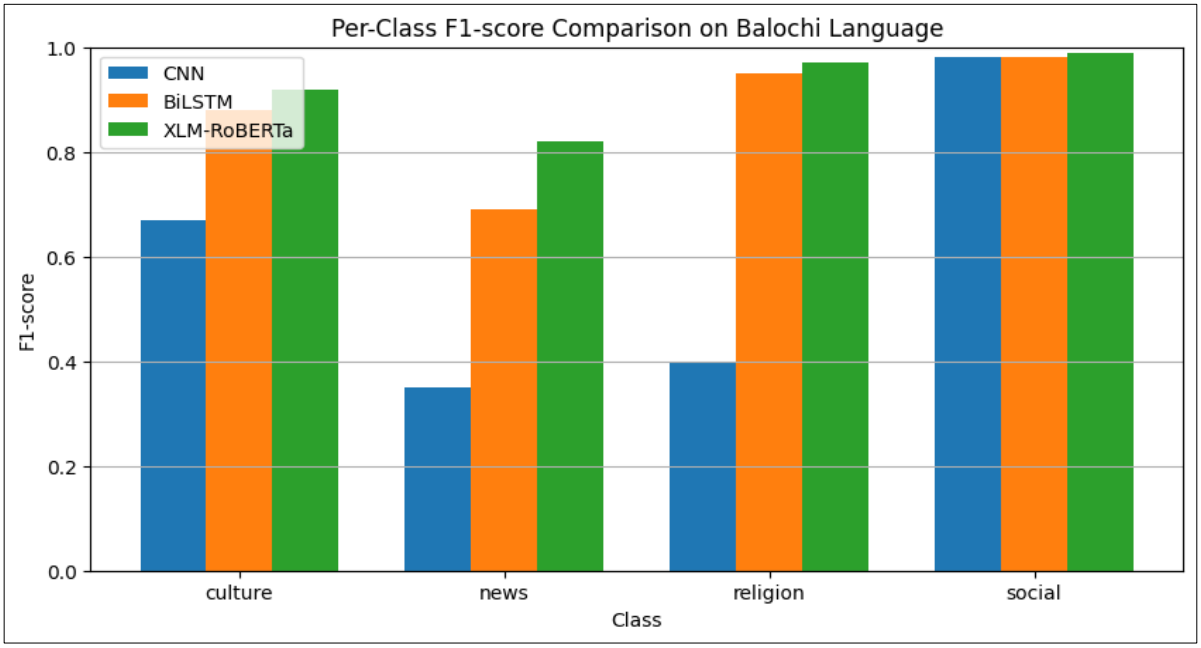

Figure 7. The Comparison Analysis of Per class F1-Scores of Deep Learning Models

Testing Results

In the testing phase of this research study, the overall performance of the five machine learning models was considered. The Machine learning models performed well and predicted the correct label for input sentence given in the Balochi language. The machine learning algorithms testing result has been mention in the figure below.

```
Prediction 1 of 5
Enter a Balochi sentence: اے لگت ہر کجام وقتءَ خیری مضمون مصنوعی ذہانتءَ...
Predicted Label: social
--------------------------------------
Prediction 2 of 5
Enter a Balochi sentence: نیلم سب توں قیمتی بیش بہاءَ نیلا جواہر انت۔
Predicted Label: social
--------------------------------------
Prediction 3 of 5
Enter a Balochi sentence: ءِ شماریاتءَ مطابق ءَ، دلءِ بیماری اوبائیو ویس...
Predicted Label: social
--------------------------------------
Prediction 4 of 5
Enter a Balochi sentence: جوں یہودی تاریخءَ گوں سب توں زلوم عالمگیر نفرت...
Predicted Label: culture
--------------------------------------
Prediction 5 of 5
Enter a Balochi sentence: ایمان دارانی کمیونٹی غیر شاگرداں تک پہنچنت، شا...
Predicted Label: religion
```

Figure 8. Prediction results of Machine Learning on Balochi Sentences

These testing results shows the classification outputs generated by machine learning models on diverse text inputs and the model predictions were observed during this testing phase.

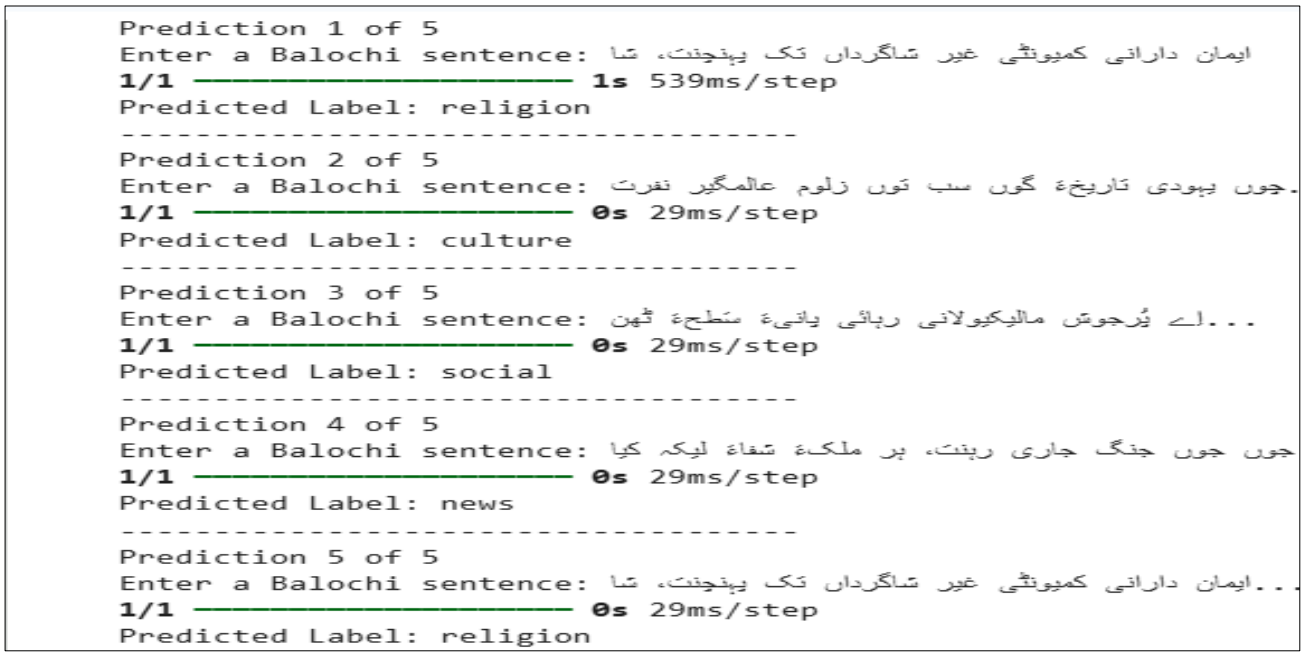

Figure 9. Prediction results of Deep Learning on Balochi Sentence

The Figure 9 above shows another testing example, in which a sentence from the dataset is classified using all the integrated classification models. Just as before, the machine learning classifier and deep learning model have been able to predict the correct class. A testing and practical evaluation interface was implemented using Gradio, to which all chosen models were integrated to facilitate testing. The experimental results from this interface have shown that the ML and DL Classifiers performed correct text classification in the most cases.

Sample test results are shown below as images to demonstrate the practical performance of the proposed models. These images show the classification outputs generated for diverse text inputs by the integrated Gradio interface and serve as a comparison visualization of model predictions during the testing phase.

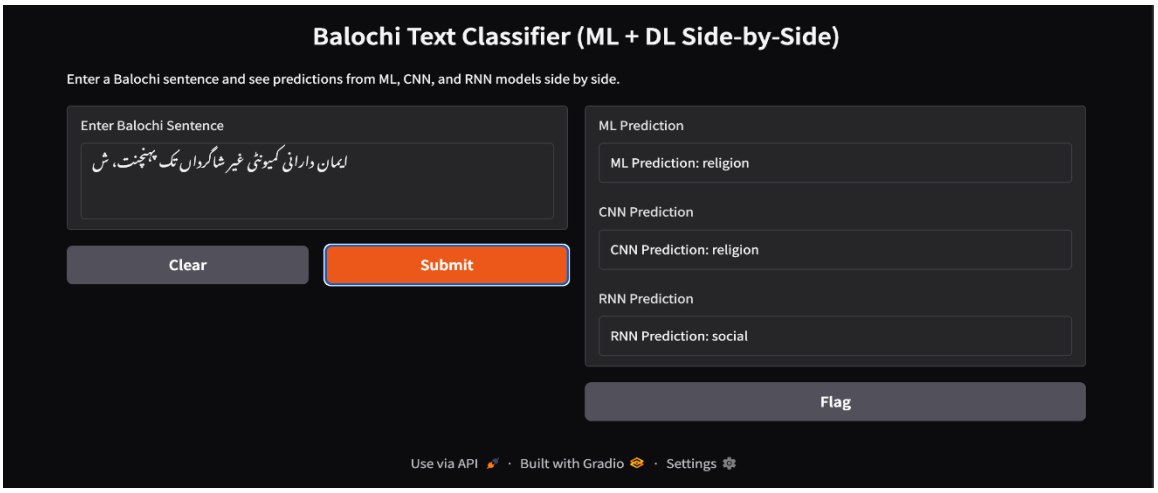

Figure 10. Testing Results of Various Machine Learning and Deep Learning Models

Conclusion

This paper presented a text classification system for low-resource language dataset collections with four classes: culture, social, religion, and news. First, manual annotation of an unlabelled dataset was performed in order to enable subsequent supervised learning. Later, standard text pre-processing was conducted along with TF-IDF-based feature extraction. Several Machine Learning and Deep Learning models were evaluated to determine the most successful approach. The experimental results indicated that, among the ML models, the best overall performance was obtained by the SGD classifier, whereas among the different deep learning methods, the XLM-RoBERTa achieved higher performance than the BiLSTM and CNN models.

Future Work

In the future, more work can be done on dataset expansion and balancing for better prediction of minority classes. It can also be extended using more transformer-based models for better contextual understanding. Advanced pre-processing and data augmentation are other ways in which the system can be developed and deployed as a real-world application.

References

S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep Learning Based Text Classification: A Comprehensive Review," vol. 1, no. 1, pp. 1–43, 2021, [Online]. Available: http://arxiv.org/abs/2004.03705

"Natural language processing applications for low-resource languages,".

K. Mohiuddin et al., "Retention Is All You Need," Int. Conf. Inf. Knowl. Manag. Proc., no. Nips, pp. 4752–4758, 2023, doi: 10.1145/3583780.3615497.

J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," NAACL HLT 2019 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. Proc. Conf., vol. 1, no. Mlm, pp. 4171–4186, 2019.

M. A. Hasan, P. Tarannum, K. Dey, I. Razzak, and U. Naseem, "Do Large Language Models Speak All Languages Equally? A Comparative Study in Low-Resource Settings," 2024, [Online]. Available: http://arxiv.org/abs/2408.02237

A. Conneau et al., "Unsupervised cross-lingual representation learning at scale," Proc. Annu. Meet. Assoc. Comput. Linguist., pp. 8440–8451, 2020, doi: 10.18653/v1/2020.acl-main.747.

Shumaila Hussain, Sibghat Ullah Bazaib, Shahab Qadir, Shah Marjan, Muhammad Imran ghafoor, and Paras Pervaiz, "Sentiment Analysis of Balochi Text Using Deep Learning," VAWKUM Trans. Comput. Sci., vol. 13, no. 1, pp. 190–200, 2025, doi: 10.21015/vtcs.v13i1.2081.

W. Yu, L. Yin, C. Zhang, Y. Chen, and A. X. Liu, "Application of Quantum Recurrent Neural Network in Low-Resource Language Text Classification," IEEE Trans. Quantum Eng., vol. 5, pp. 1–13, 2024, doi: 10.1109/TQE.2024.3373903.

S. Kumar and R. Vavekanand, "Multiclass Text Classifications of Sindhi Newspaper Articles," pp. 0–9, 2025, doi: 10.20944/preprints202501.1580.v1.

H. H. Saeed, T. Khalil, and F. Kamiran, "Urdu Toxic Comment Classification With PURUTT Corpus Development," IEEE Access, vol. 13, no. December 2024, pp. 21635–21651, 2025, doi: 10.1109/ACCESS.2025.3535862.

"In the digital era , Natural Language Processing (NLP) is vital as more communities and

languages go online . While European and East Asian languages have advanced in computational processing , many South Asian languages , including Sindhi , face challe," pp. 29–30, 2022.

G. Nigusie, "Supervised Machine Learning based Amharic Text Complexity Classification Using Automatic Annotator Tool," no. AfricaNLP, pp. 7–14, 2025, doi: 10.18653/v1/2025.africanlp-1.2.

A. A. Ayalew et al., "Amharic event text classification from social media using hybrid deep learning," Int. J. Electr. Comput. Eng., vol. 15, no. 2, p. 2264, 2025, doi: 10.11591/ijece.v15i2.pp2264-2270.

T. Yeshambel, J. Mothe, and Y. Assabie, "Learned Text Representation for Amharic Information Retrieval and Natural Language Processing," Inf., vol. 14, no. 3, 2023, doi: 10.3390/info14030195.

T. Alam, A. Khan, and F. Alam, "Bangla Text Classification using Transformers," 2020, [Online]. Available: http://arxiv.org/abs/2011.04446

A. Nawaz, M. Nawaz, N. A. Shaikh, S. Rajper, J. Baber, and M. Khalid, "TPTS: Text pre-processing Techniques for Sindhi Language," Pakistan J. Emerg. Sci. Technol., vol. 4, no. 3, pp. 1–12, 2023, doi: 10.58619/pjest.v4i3.89.

M. A. Chhajro, "Multi-text classification of Urdu/Roman using machine learning and natural language preprocessing techniques," Indian J. Sci. Technol., vol. 13, no. 19, pp. 1890–1900, 2020, doi: 10.17485/ijst/v13i19.230.

A. Ali, M. Khan, K. Khan, R. U. Khan, and A. Aloraini, "Sentiment Analysis of Low-Resource Language Literature Using Data Processing and Deep Learning," Comput. Mater. Contin., vol. 79, no. 1, pp. 713–733, 2024, doi: 10.32604/cmc.2024.048712.

K. Madatov, S. Sattarova, and J. Vičič, "TF-IDF-Based Classification of Uzbek Educational Texts," Appl. Sci., vol. 15, no. 19, p. 10808, 2025, doi: 10.3390/app151910808.

J. A. Baktash, M. Dawodi, M. Zarif, and N. Hassanzada, "Tuning Traditional Language Processing Approaches for Pashto Text Classification," Int. J. Nat. Lang. Comput., vol. 12, no. 2, pp. 53–66, 2023, doi: 10.5121/ijnlc.2023.12205.

E. Kuriyozov, U. Salaev, S. Matlatipov, and G. Matlatipov, "Text classification dataset and analysis for Uzbek language," no. Section 4, 2023, [Online]. Available: http://arxiv.org/abs/2302.14494

K. Kowsari, K. J. Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, "Text classification algorithms: A survey," Inf., vol. 10, no. 4, pp. 1–68, 2019, doi: 10.3390/info10040150.

A. Irum and M. A. Tahir, "Document-Level Sentiment Analysis of Urdu Text Using Deep Learning Techniques," vol. 2776, no. 2015, pp. 0–2, 2025, [Online]. Available: http://arxiv.org/abs/2501.17175

A. Olalekan, O. Daniel, and O. N. Emuoyibofarhe, "MACHINE LEARNING ALGORITHMS FOR MULTILINGUAL TEXT," vol. 2, no. 2, pp. 25–30, 2022.