L3Cube-IndicNews: News-based Short Text and Long Document Classification Datasets in Indic Languages

Aishwarya Mirashi[1,3], **Srushti Sonavane**[1,3], **Purva Lingayat**[1,3], **Tejas Padhiyar**[1,3] and **Raviraj Joshi**[2,3]
Pune Institute of Computer Technology, Pune[1]
Indian Institute of Technology Madras, Chennai[2]
L3Cube Labs, Pune[3]

Abstract

In this work, we introduce L3Cube-IndicNews, a multilingual text classification corpus aimed at curating a high-quality dataset for Indian regional languages, with a specific focus on news headlines and articles. We have centered our work on 10 prominent Indic languages, including Hindi, Bengali, Marathi, Telugu, Tamil, Gujarati, Kannada, Odia, Malayalam, and Punjabi. Each of these news datasets comprises 10 or more classes of news articles. L3Cube-IndicNews offers 3 distinct datasets tailored to handle different document lengths that are classified as: Short Headlines Classification (SHC) dataset containing the news headline and news category, Long Document Classification (LDC) dataset containing the whole news article and the news category, and Long Paragraph Classification (LPC) containing sub-articles of the news and the news category. We maintain consistent labeling across all 3 datasets for in-depth length-based analysis. We evaluate each of these Indic language datasets using 4 different models including monolingual BERT, multilingual Indic Sentence BERT (IndicSBERT), and IndicBERT. This research contributes significantly to expanding the pool of available text classification datasets and also makes it possible to develop topic classification models for Indian regional languages. This also serves as an excellent resource for cross-lingual analysis owing to the high overlap of labels among languages. The datasets and models are shared publicly at https://github.com/l3cube-pune/indic-nlp.

Keywords: Low Resource Languages, Indic Languages, Web Scraping, News Article Datasets, BERT, Short Text Classification, Long Documents.

1 Introduction

India boasts a rich linguistic diversity, with over 700 languages spoken, out of which 22 are officially recognized. Some of the primarily spoken languages include Hindi, Bengali, Marathi, Telugu, Tamil, Gujarati, Kannada, Odia, Malayalam, and Punjabi. Despite their widespread use, there's a notable scarcity of comprehensive Indic language datasets, primarily due to their low-resource status and linguistic complexity (Patil and Patil, 2017).

These Indic languages are widely spoken all over India and have abundant data available on news websites, and social media but it's difficult to find a labeled dataset of news headlines, articles, and their categories for text classification. This disparity has hindered progress in machine learning and Natural Language Processing (NLP) research for these languages.

Most Indic languages share similarities, but they utilize different writing scripts, making it more challenging to accurately predict news categories. Since English is used widely around the world, and many researchers have studied how to classify text in English, it has resulted in a surplus of classification datasets. The same is not the case for Indic languages. Although there are existing datasets for Indian languages, they have some limitations, such as a lesser number of categories or inconsistent classification criteria for news articles, which narrows down the scope of research in this field. While the IndicNLP News Article dataset covers the major languages it is limited by the count of target labels and high accuracy (Kakwani et al., 2020; Kulkarni et al., 2022). This calls for the creation of more intricate datasets to effectively assess the performance of models. In essence, there is a need for complex datasets that can thoroughly evaluate model effectiveness.

Text classification, a critical task in both machine learning and natural language processing (NLP), involves categorizing text documents into predefined classes based on their content. While there are a handful of publicly available datasets related to news, they often lack diversity in categories and sources, potentially leading to biased results.

Transformer-based models, like LongFormer[1] (Beltagy et al., 2020), require datasets with varying sequence lengths due to their sensitivity to text length. This highlights the need for specialized datasets to develop models for these low-resource languages. Hence, we introduce L3Cube-IndicNews, a comprehensive IndicNews Classification Dataset, sourced from diverse news websites specifically targeting low-resource languages.

This dataset encompasses over 3 lakh records, distributed across 12 diverse news categories, offering an extensive resource for supervised text classification. Each language dataset contains more than 26,000 rows, covering at least 10 significant news categories.

L3Cube-IndicNLP[2] presents monolingual and multilingual models tailored to each Indian regional language. The repository includes individual BERT models for the languages focused on in this work. We conduct a comparative analysis of various monolingual and multilingual BERT models, including L3Cube monolingual BERT (Joshi, 2022a,b), monolingual SBERT (Deode et al., 2023; Joshi et al., 2023a), IndicSBERT (multilingual) (Deode et al., 2023) and IndicBERT(Kakwani et al., 2020).

The key contributions of this work are as follows:

- Introduction of L3Cube-IndicNews, an extensive document classification dataset spanning ten significant Indian languages, each consistent with a range of 12 target labels. The dataset will be made publicly available.

- The corpus comprises three sub-datasets (IndicNews- SHC, LPC, and LDC) catering to short, medium, and long documents, each with varying sentence lengths but consistent target labels.

- The datasets are bench-marked using state-of-the-art pre-trained BERT models: L3Cube monolingual BERT, L3Cube monolingual SBERT, L3Cube IndicSBERT (multilingual) and IndicBERT.

[1] https://huggingface.co/docs/transformers/model_doc/longformer

[2] https://github.com/l3cube-pune/indic-nlp

2 Related work

IndicNLP News Article Classification[3] dataset is part of the AI4Bharat-IndicNLP Dataset (Kunchukuttan et al., 2020) that consists of news articles in 10 Indian languages categorized into classes like sports, entertainment, business, politics, and lifestyle. While this dataset contains a substantial number of records, it falls short in terms of the variety of categories available for news articles in each language, thereby limiting its diversity.

Multi Indic Languages News Dataset[4] is a dataset publicly available on Kaggle. It is a multi-language news dataset from Times Internet for various Indian languages. This data contains columns named title, link, description, long_description, id. Despite the extensive size and diversity of this dataset, encompassing a rich collection of records across numerous Indian languages, it lacks language-wise and news category-wise segregation, which hampers clarity and ease of use.

Varta[5] is a multilingual dataset for headline generation. It encompasses 41.8 million news articles in 14 Indic languages and English. This data is sourced from DailyHunt. This dataset is well-organized and includes a large number of records for each language, covering all the minor details. However, this dataset contains only those articles written by DailyHunt's partner publishers resulting in a biased nature towards a particular narrative or ideology that can affect the representativeness and diversity of the dataset. (Aralikatte et al., 2023) From every language, they randomly sample 10,000 articles each for validation and testing. On average, Varta articles have 17 sentences, and the headlines have just over one sentence. A typical article sentence contains about 18 words, and a headline sentence contains 11 words. While the dataset is a large-scale, high-quality dataset for Indic languages, the headlines in this dataset are 39% smaller than the average sentence in an article.

iNLTK[6] is an openly accessible dataset primarily comprising data for 13 Indic languages, sourced from Wikipedia articles. It encompasses over 12,000 cleaned rows for each of these languages (Arora, 2020). This dataset comprises publicly

[3] https://github.com/AI4Bharat/indicnlp_corpus#indicnlp-news-article-classification-dataset

[4] https://www.kaggle.com/datasets/shaz13/multi-indic-languages-news-datasethttps://www.kaggle.com/datasets/shaz13/multi-indic-languages-news-dataset

[5] https://huggingface.co/datasets/rahular/varta

[6] https://github.com/goru001/inltk

accessible data for languages like Hindi, Bengali, Punjabi, Kannada, and Oriya. For languages such as Gujarati, Malayalam, Marathi, and Tamil, they have created their dataset by extracting information from Wikipedia articles. While the dataset boasts a substantial volume of records for each language, it falls short in terms of categorizing the data into specific news categories. Furthermore, in some languages, it includes not only news-related articles but also other types of content, leading to inefficiency and inconsistency in the dataset's content and structure.

ACTSA(Mukku and Mamidi, 2017) focuses on building a gold-standard annotated corpus of Telugu sentences to support Telugu Sentiment Analysis. The raw data is scraped from five different Telugu news websites viz. Andhrabhoomi, Andhrajyothi, Eenadu, Kridajyothi and Sakshi. In total, they collected over 453 news articles which were then filtered down to 321 articles relevant to their work.

Language	Datasets available	Categories present	Articles
Hindi	BBC Articles BBC Hindi News Articles	India, International, Entertainment, Sports, Others India, Pakistan, News, International, Sntertainment, port, Science, China, Learningenglish, Social, Southasia, Business, Institutional, Multimedia	172K 4335
Bengali	AI4Bharat-IndicNLP Dataset Soham Articles	Entertainment, Sports Kolkata, State, National, International, Sports, Entertainment	14K 72K
Marathi	AI4Bharat-IndicNLP Dataset iNLTK Headlines	Entertainment, Lifestyle, Sports State, Entertainment, Sports	4.5K 85K
Telugu	AI4Bharat-IndicNLP Dataset	Entertainment, Business, Sports	24K
Tamil	AI4Bharat-IndicNLP Dataset iNLTK Headlines	Entertainment, Politics, Sports Tamil-cinema, Business, Spirituality	11.7K 127K
Gujarati	AI4Bharat-IndicNLP Dataset iNLTK Headlines	Business, Entertainment, Sports Entertainment, Business, Tech	2K 31K
Kannada	AI4Bharat-IndicNLP Dataset iNLTK-IndicNLP News Category	Entertainment, Lifestyle, Sports Entertainment, Sports, Tech	30K 6.3K
Odia	AI4Bharat-IndicNLP Dataset iNLTK-IndicNLP News Category	Business, Crime, Entertainment, Sports Sports, Business, Entertainment	30K 19K
Malayalam	AI4Bharat-IndicNLP Dataset iNLTK Headlines	Business, Entertainment, Sports, Technology Entertainment, Sports, Business	6K 12K
Punjabi	AI4Bharat-IndicNLP Dataset iNLTK -IndicNLP News Category	Business, Crime, Entertainment, sports Politics, Non-politics	3.1K 800

Table 1: Available Datasets in Indic languages

3 Curating the dataset

We introduce L3Cube-IndicNews, a comprehensive dataset compilation designed to facilitate the classification of both short text and long documents. Within IndicNews, there are three meticulously crafted supervised datasets: Short Headlines Classification (SHC), Long Document Classification (LDC), and Long Paragraph Classification (LPC).

Featuring a broad spectrum of information, the dataset comprises a minimum of 10 distinct categories corresponding to 10 prominent languages spoken in India. The careful organization of both language and newly defined categories enhances overall clarity. Sourced from multiple websites, it ensures a diverse array of content, exclusively from reputable news sources. Notably, there are no constraints on the length of articles and headlines, emphasizing a commitment to quality. In essence, it stands as a high-quality, versatile, and meticulously curated dataset.

3.1 Data collection

For the dataset curation process, we identified several websites to collect a substantial number of articles for each news category. The Hindi language dataset was scraped from Jansatta[7]. The Marathi language dataset was scraped from Lokmat[8]. The Bengali language dataset was scraped from Aajkal[9], Ganashakti[10], BBC[11], Anandabazar[12], abnews24[13], and Sangbadpratidin[14]. The Telugu language dataset was scraped from ABP Telugu[15]. The Tamil Language dataset was scraped from Hindu Tamil[16]. The Gujarati language dataset was scraped from ABP Gujarati[17]. The Kannada language dataset was scraped from Kannada Prabha[18] and PublicTV[19]. The Odia language dataset was scraped from Odisha Bhaskar[20] and Dharitri[21]. The Malayalam language dataset was scraped from Madhyamam[22]. The Punjabi language dataset was scraped from Khabarwaale[23].

The data was gathered through the utilization of the urllib package for managing URL requests, coupled with the BeautifulSoup package for extracting data from the HTML content of the requested URL.

[7] https://www.jansatta.com/
[8] https://www.lokmat.com/
[9] https://www.aajkaal.in/
[10] https://ganashakti.com/
[11] https://www.bbc.com/bengali
[12] https://www.anandabazar.com/
[13] https://www.sangbadpratidin.in/
[14] https://www.abnews24.com/
[15] https://telugu.abplive.com/
[16] https://www.hindutamil.in/
[17] https://gujarati.abplive.com/
[18] https://www.kannadaprabha.com/
[19] https://publictv.in/
[20] https://odishabhaskar.com/
[21] https://www.dharitri.com/https://www.dharitri.com/
[22] https://www.madhyamam.com/
[23] https://www.khabarwaale.com/

Every website had organized its news articles into predefined categories, and during the scraping process, we retained this categorization to utilize it as the target label.

Each dataset was originally scraped to contain 3 columns: Title, Category, and News. Then 35% of the news article from column 'News' was used to create a fourth column 'Sub article', primarily containing a subset of news articles. The final curated dataset underwent shuffling and cleaning. We then divided these datasets into three supervised datasets: Short Headlines Classification (SHC), Long Document Classification (LDC), and Long Paragraph Classification (LPC).

Short Headlines Classification (SHC): This dataset contains the headlines of news articles paired with their respective categorical labels.

***Long Paragraph Classification (LPC)*:** In this dataset, each record contains a sub-article of news with its respective categorical label.

***Long Document Classification (LDC)*:** This dataset contains records having an entire news article with its corresponding categorical label.

3.2 Data Preprocessing

At first, we break down the article into sentences, making sure we handle the punctuation and sentence boundaries in Indic languages correctly. Then, we carefully break each sentence into tokens and filter them for special characters. We selectively retain tokens with an initial character aligned with the specific Indic language character set. We also created a dataset-cleansing function wherein each text is tokenized and then undesired elements, including words with specific characters, substrings, or minimal length are eliminated. Furthermore, regular expression was used to filter out words that don't match the script of the language. This data refinement process ensures the dataset is free from unwanted elements, making it optimal for further use.

3.3 Dataset Statistics

Each dataset consists of more than 26,000 rows and 10-12 categories of news articles per dataset. Each of the 10 datasets was split into the train, test, and validation datasets in the ratio of 80:10:10.

Language	Labels
Hindi (11)	Auto, Business, Crime, Education, Entertainment, Health, International, Nation, Politics, Sports, Technology
Bengali (10)	Sports, National, International, Kolkata, State, Politics, Entertainment, Technology, Editorial, Lifestyle
Marathi (12)	Auto, Bhakti, Crime, Education, Fashion, Health, International, Manoranjan, Politics, Sports, Tech, Travel
Telugu (10)	Business, Crime, Education, Entertainment, Jobs, Lifestyle, Politics, Sports, Technology, World
Tamil (10)	Auto, Business, Crime, Education, Entertainment, Health, India, Lifestyle, Politics, World
Gujarati (10)	Astro, Auto, Business, Crime, Entertainment, International, Nation, Sports, State, Technology
Kannada (10)	Business, Crime, Cuisine, Entertainment, International, Nation, Politics, Sports, State, Technology
Odia (10)	Lifestyle, Entertainment, News, Crime, Business, Health, Politics, Career, Sports, Editorial
Malayalam (10)	Crime, Entertainment, Gulf, International, Kerala, Lifestyle, National, Opinion, Sports, Technology
Punjabi (10)	World, Sports, Religious, Education, Transfer & Appointments, Literature, Health, National, Crime, Lifestyle

Table 2: Languages and their categorical labels

Language	Train	Test	Validation	Total
Hindi	30851	3835	3835	38521
Bengali	23970	2997	2996	29963
Marathi	22014	2761	2750	27525
Telugu	21103	2640	2650	26393
Tamil	25030	3129	3129	31288
Gujarati	26472	3347	3417	33236
Kannada	24642	3058	3058	30758
Odia	27420	3445	3434	34299
Malayalam	28000	3500	3500	35000
Punjabi	25494	3189	3186	31869

Table 3: Distribution of dataset into train, test, and validation in the ratio 80:10:10.

Title	Category
'कांग्रेस को सत्ता या PM पद में दिलचस्पी नहीं', मल्लिकार्जुन खड़गे ने बताया क्या है विपक्ष की बैठक का उद्देश्य'कांग्रेस को सत्ता या PM पद में दिलचस्पी नहीं', मल्लिकार्जुन खड़गे ने बताया क्या है विपक्ष की बैठक का उद्देश्य	National

Figure 1: SHC Dataset Overview

Sub article	Category
उसने सत्ता में आने के लिए अपने सहयोगी दलों के वोट का इस्तेमाल किया और फिर उन्हें छोड़ दिया आज बीजेपी अध्यक्ष और उनके नेता अपने पुराने साथियों को साथ लाने के लिए एक राज्य से दूसरे राज्य भाग रहे हैंमल्लिकार्जुन खड़गे ने कहा कि हर संस्था को विपक्ष के खिलाफ एक हथियार के रूप में इस्तेमाल किया जा रहा है CBI, ED और इनकम टैक्स का लगातार इस्तेमाल हो रहा है कानूनी प्रक्रिया में उलझाने के लिए हमारे नेताओं के खिलाफ फर्जी क्रिमिनल केस फाइस किए जा रहे हैं हमारे सांसदों को निलंबित करने के लिए संवैधानिक अथॉरिटी का इस्तेमाल किया जा रहा है विधायकों को ब्लैकमेल करके या घूस देकर बीजेपी में शामिल किया जा रहा है और सरकारें गिराई जा रही हैं	National

Figure 2: LPC Dataset Overview

News	Category
Lok Sabha Chunav में बीजेपी को हराने के लिए विपक्ष मंथन कर रहा है। मंगलवार को बैठक में कांग्रेस के अध्यक्ष मल्लिकार्जुन खड़गे ने कहा कि पीएम पद में हमारी रुचि नहीं है।CBI और ED को हथियार की तरह इस्तेमाल कर रही है बीजेपीबेंगलुरु में चल रही विपक्ष की मीटिंग में कांग्रेस पार्टी की तरफ से बड़ा बयान सामने आया है। न्यूज एजेंसी ANI द्वारा सूत्रों के हवाले से दी गई जानकारी के अनुसार, कांग्रेस पार्टी के अध्यक्ष मल्लिकार्जुन खड़गे ने कहा कि कांग्रेस को पावर या पीएम पद में कोई इंटरेस्ट नहीं है। सूत्रों ने बताया कि उन्होंने कहा, "मैंने MK स्टालिन के जन्मदिन पर चेन्नई में पहले ही कहा था कि कांग्रेस को सत्ता या प्रधानमंत्री पद में कोई दिलचस्पी नहीं है। इस बैठक में हमारा इरादा अपने लिए सत्ता हासिल करना नहीं है। यह हमारे संविधान, लोकतंत्र, धर्मनिरपेक्षता और सामाजिक न्याय की रक्षा के लिए है।"उन्होंने विपक्ष की मीटिंग में कहा कि हम जानते हैं कि स्टेट लेवल पर हम लोगों के बीच में कुछ मतभेद हैं। ये मतभेद विचारधारा से जुड़े नहीं हैं। ये मतभेद इतने बड़े भी नहीं है कि हम आम आदमी, मीडिल क्लास, युवा, गरीब, दलित, आदिवासी और अल्पसंख्यकों के हक के लिए इन्हें पीछे नहीं छोड़ सकते।सूत्रों के अनुसार, उन्होंने आगे कहा, "यहां पर हम 26 दल हैं। हम सभी मिलाकर 11 राज्यों में सरकार चला रहे हैं। बीजेपी को 303 सीटें अपने दम पर नहीं मिली हैं। उसने सत्ता में आने के लिए अपने सहयोगी दलों के वोट का इस्तेमाल किया और फिर उन्हें छोड़ दिया। आज बीजेपी अध्यक्ष और उनके नेता अपने पुराने साथियों को साथ लाने के लिए एक राज्य से दूसरे राज्य भाग रहे हैं।मल्लिकार्जुन खड़गे ने कहा कि हर संस्था को विपक्ष के खिलाफ एक हथियार के रूप में इस्तेमाल किया जा रहा है। CBI, ED और इनकम टैक्स का लगातार इस्तेमाल हो रहा है। कानूनी प्रक्रिया में उलझाने के लिए हमारे नेताओं के खिलाफ फर्जी क्रिमिनल केस फाइल किए जा रहे हैं। हमारे सांसदों को निलंबित करने के लिए संवैधानिक अथॉरिटी का इस्तेमाल किया जा रहा है। विधायकों को ब्लैकमेल करके या घूस देकर बीजेपी में शामिल किया जा रहा है और सरकारें गिराई जा रही हैं।	National

Figure 3: LDC Dataset Overview

4 Models

4.1 L3Cube Monolingual BERT[24] for Indic languages

We use the monolingual BERT models for the 10 Indic languages, released by L3cube-Pune[25] as the base models. These models are termed as HindBERT[26], BengaliBERT[27], MahaBERT[28], TeluguBERT[29], TamilBERT[30], GujaratiBERT[31], KannadaBERT[32], OdiaBERT[33], MalayalamBERT[34], PunjabiBERT[35]. These models are fine-tuned on the existing multilingual models like MuRIL (Khanuja et al., 2021), xlmRoBERTa (Conneau et al., 2019), and IndicBERT on the monolingual corpus.

4.2 L3Cube Indic Sentence BERT[36] models (Monolingual)

We also evaluate L3Cube monolingual Indic SBERT models that are HindSBERT[37](Joshi et al., 2023b), BengaliSBERT[38], MahaSBERT[39], TeluguSBERT[40], TamilSBERT[41], GujaratiSBERT[42], KannadaSBERT[43], OdiaSBERT[44], MalayalamSBERT[45], and PunjabiSBERT[46].

4.3 L3Cube Indic SBERT[47] (Multilingual)

IndicSBERT is the first multilingual SBERT model trained specifically for Indic languages. Sentence-BERT (SBERT) (Reimers and Gurevych, 2019) is a modified version of the BERT (Devlin et al., 2018) architecture designed to generate sentence representations for the improved semantic similarity between sentences. The SBERT uses a Siamese network (Koch et al., 2015) and is trained using specific datasets like STS, resulting in representations specifically geared for semantic similarity.

4.4 AI4Bharat indicBERT[48]

IndicBERT is a multi-lingual AlBERT model provided by AI4Bharat exclusively pre-trained in 12 Indian languages. It is pre-trained on AI4Bharat IndicNLP Corpora of around 9 billion tokens.

5 Evaluation

We evaluated each Indic dataset using the monolingual BERT model, multilingual BERT and SBERT models provided by L3Cube, and the indicBERT model provided by AI4Bharat. The results of evaluating these models on the curated datasets are shown in Table 4.

[24] https://arxiv.org/abs/2211.11418
[25] https://github.com/l3cube-pune
[26] https://huggingface.co/l3cube-pune/hindi-bert-v2
[27] https://huggingface.co/l3cube-pune/bengali-bert
[28] https://huggingface.co/l3cube-pune/marathi-bert
[29] https://huggingface.co/l3cube-pune/telugu-bert
[30] https://huggingface.co/l3cube-pune/tamil-bert
[31] https://huggingface.co/l3cube-pune/gujarati-bert
[32] https://huggingface.co/l3cube-pune/kannada-bert
[33] https://huggingface.co/l3cube-pune/odia-bert
[34] https://huggingface.co/l3cube-pune/malayalam-bert
[35] https://huggingface.co/l3cube-pune/punjabi-bert
[36] https://arxiv.org/pdf/2304.11434.pdf
[37] https://huggingface.co/l3cube-pune/hindi-sentence-bert-nli
[38] https://huggingface.co/l3cube-pune/bengali-sentence-bert-nli
[39] https://huggingface.co/l3cube-pune/marathi-sentence-bert-nli
[40] https://huggingface.co/l3cube-pune/telugu-sentence-bert-nli
[41] https://huggingface.co/l3cube-pune/tamil-sentence-bert-nli
[42] https://huggingface.co/l3cube-pune/gujarati-sentence-bert-nli
[43] https://huggingface.co/l3cube-pune/kannada-sentence-bert-nli
[44] https://huggingface.co/l3cube-pune/odia-sentence-bert-nli
[45] https://huggingface.co/l3cube-pune/malayalam-sentence-bert-nli
[46] https://huggingface.co/l3cube-pune/punjabi-sentence-bert-nli
[47] https://huggingface.co/l3cube-pune/indic-sentence-bert-nli
[48] https://huggingface.co/ai4bharat/indic-bert

Language	Model	SHC	LDC	LPC
Hindi	HindiBERT	**86.600**	**91.681**	**88.097**
	HindiSBERT	86.518	91.534	87.810
	IndicSBERT	85.870	91.604	86.922
	IndicBERT	83.910	86.809	84.424
Bengali	BengaliBERT	82.549	**95.293**	**85.195**
	BengaliSBERT	**82.749**	94.141	84.470
	IndicSBERT	80.674	92.457	85.232
	IndicBERT	80.007	90.757	82.457
Marathi	MarathiBERT	**91.163**	**94.706**	86.731
	MarathiSBERT	91.017	94.349	**87.439**
	IndicSBERT	90.510	93.987	87.103
	IndicBERT	89.388	92.627	85.222
Telugu	TeluguBERT	89.810	**92.765**	**91.818**
	TeluguSBERT	**90.416**	92.651	91.098
	IndicSBERT	87.916	92.348	91.515
	IndicBERT	88.371	92.943	91.811
Tamil	TamilBERT	**81.785**	84.521	**81.300**
	TamilSBERT	81.720	**86.122**	79.573
	IndicSBERT	81.209	84.227	79.703
	IndicBERT	81.275	83.226	80.571
Gujarati	GujaratiBERT	**89.898**	**95.278**	**90.640**
	GujaratiSBERT	89.808	95.060	90.091
	IndicSBERT	88.613	95.277	89.573
	IndicBERT	87.909	90.854	89.050
Kannada	KannadaBERT	**91.410**	94.706	**87.704**
	KannadaSBERT	89.340	94.539	87.345
	IndicSBERT	90.255	**94.833**	87.835
	IndicBERT	87.965	91.857	87.358
Odia	OdiaBERT	83.399	92.940	**84.557**
	OdiaSBERT	**84.137**	**93.424**	83.918
	IndicSBERT	82.065	93.054	84.441
	IndicBERT	82.661	86.962	82.322
Malayalam	MalayalamBERT	**80.440**	**88.573**	**81.705**
	MalayalamSBERT	80.171	88.201	81.171
	IndicSBERT	77.780	88.029	79.672
	IndicBERT	75.114	83.098	76.457
Punjabi	PunjabiBERT	85.456	90.182	84.163
	PunjabiSBERT	90.363	**94.866**	90.333
	IndicSBERT	89.490	94.224	88.446
	IndicBERT	**91.725**	94.781	**90.358**

Table 4: Accuracies for all the models trained on SHC, LDC, and LPC datasets in percentage (%)

Language	Model	SHC	LDC	LPC
Hindi	HindiSBERT	86.209	88.451	88.279
Bengali	BengaliSBERT	82.615	93.126	89.289
Marathi	MarathiSBERT	91.017	94.349	87.439
Telugu	TeluguSBERT	88.674	92.310	91.174
Tamil	TamilSBERT	81.754	84.902	80.993
Gujarati	GujaratiSBERT	88.792	92.378	90.609
Kannada	KannadaSBERT	89.797	94.212	88.260
Odia	OdiaSBERT	84.137	93.424	83.918
Malayalam	MalayalamSBERT	79.114	87.743	80.428
Punjabi	PunjabiSBERT	94.744	93.614	93.782

Table 5: Accuracies for monolingual SBERT model trained on mixed dataset (SHC + LPC + LDC) in percentage (%)

6 Results

For evaluating the models on these datasets, we use accuracy as our main evaluation metric to understand the performance of our models. Table 4 presents the accuracies obtained by fine-tuning our models on the datasets. The confusion matrices for the Kannada dataset trained and tested on SHC, LDC, and LPC respectively are illustrated in Figures 4, 5, and 6.

Key observations are outlined as follows:

- In most cases, the L3Cube Monolingual model tends to exhibit superior performance in terms of accuracy, as demonstrated by the metrics presented in the table across all corpora.
- Among the Short Headlines Classification (SHC), Long Document Classification (LDC), and Long Paragraph Classification (LPC) datasets, LDC demonstrated the most impressive results after fine-tuning text classification. This aligns with expectations, given that long document data inherently contains more information compared to shorter-length datasets.
- On the other hand, SHC reported comparatively lower accuracy scores across all three document types. This may be attributed to the fact that news headlines can sometimes encompass more generalized information, potentially leading to some degree of confusion for the models.

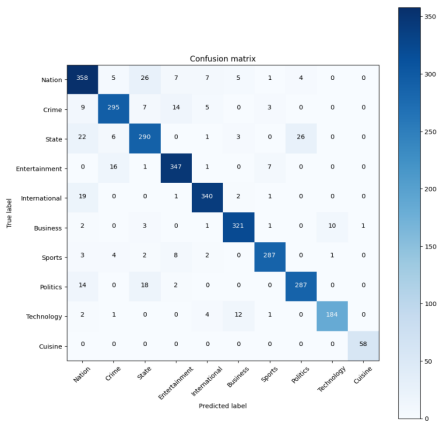

Figure 4: Confusion matrix for the Kannada SHC dataset

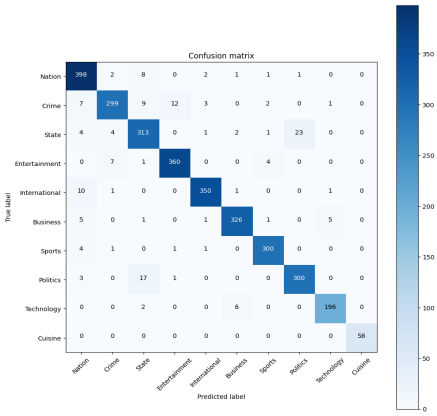

Figure 5: Confusion matrix for the Kannada LDC dataset

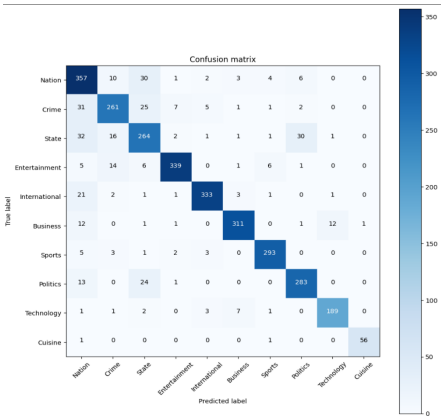

Figure 6: Confusion matrix for the Kannada LPC dataset

Table 5 presents the accuracies obtained by training the L3Cube Monolingual Sentence BERT model on a mix of all the 3 datasets (SHC + LPC + LDC) and then evaluating them on individual datasets. A cross-analysis was conducted by evaluating the performance of L3Cube Monolingual Indic Sentence BERT model on individual test datasets. Upon fine-tuning the Monolingual SBERT model on a mixed dataset, notable results were observed, with the Long Document Classification (LDC) task exhibiting the most impressive performance.

7 Future Work

There is potential for expanding the dataset's labels to achieve broader category coverage or facilitate further category expansion. The dataset creation process has been streamlined through automation, eliminating the need for manual typing. Looking forward, there is a prospect of implementing the curation of a manually typed or verified dataset, ensuring an even higher level of accuracy and reliability for future applications.

Currently, we have achieved proficiency in Long Document Classification (LDC) for a single language, and the performance in Long Paragraph Classification (LPC) is satisfactory. However, we acknowledge that the performance in Short Headlines Classification (SHC) falls short of optimization. To address this, we have implemented a unified model selection, specifically opting for the monolingual SBERT based on its significant performance on individual datasets for each language.

This chosen model exhibits competence in handling mixed datasets, comprising of SHC, LPC, and LDC, and its performance has been thoroughly examined. It's important to note that our current focus remains on a single language. Looking ahead, we recognize the potential for future developments, particularly in the realm of cross-dataset analysis.

8 Conclusion

In this research paper, we introduce L3Cube-IndicNews, a comprehensive collection of three labeled datasets, encompassing over 3 lakh records in ten different Indic languages. These datasets are designed for text classification tasks in the context of Indian languages. Within this paper, we provide an in-depth overview of the creation process, which involves the use of 10 to 12 distinct categorical labels to curate these supervised datasets. To assess the effectiveness of these datasets, we conducted fine-tuning on BERT-based models, serving as a valuable benchmark for future research and development.

Our experiments involved four key models: L3Cube Monolingual BERT, L3Cube Monolingual SBERT, L3Cube-IndicSBERT (multilingual), and AI4BHARAT IndicBERT. Notably, our findings indicate that the BengaliBERT Model achieved the highest accuracy when applied to the LDC dataset. We believe that the availability of our datasets will contribute significantly to the enhancement of NLP support for the Indic languages, promoting its growth and development in this field.

9 Acknowledgements

This work was carried out under the mentorship of L3Cube, Pune. We would like to express our gratitude towards our mentor, for his continuous support and encouragement. This work is a part of the L3Cube-IndicNLP Project.

References

Rahul Aralikatte, Ziling Cheng, Sumanth Doddapaneni, and Jackie Chi Kit Cheung. 2023. V\= arta: A large-scale headline-generation dataset for indic languages. *arXiv preprint arXiv:2305.05858*.

Gaurav Arora. 2020. inltk: Natural language toolkit for indic languages. *arXiv preprint arXiv:2009.12534*.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Samruddhi Deode, Janhavi Gadre, Aditi Kajale, Ananya Joshi, and Raviraj Joshi. 2023. L3cube-indicsbert: A simple approach for learning cross-lingual sentence representations using multilingual bert. *arXiv preprint arXiv:2304.11434*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ananya Joshi, Aditi Kajale, Janhavi Gadre, Samruddhi Deode, and Raviraj Joshi. 2023a. L3cube-mahasbert and hindsbert: Sentence bert models and benchmarking bert sentence representations for hindi and marathi. In *Science and Information Conference*, pages 1184–1199. Springer.

Ananya Joshi, Aditi Kajale, Janhavi Gadre, Samruddhi Deode, and Raviraj Joshi. 2023b. L3cube-mahasbert and hindsbert: Sentence bert models and benchmarking bert sentence representations for hindi and marathi. In *Science and Information Conference*, pages 1184–1199. Springer.

Raviraj Joshi. 2022a. L3cube-hindbert and devbert: Pre-trained bert transformer models for devanagari based hindi and marathi languages. *arXiv preprint arXiv:2211.11418*.

Raviraj Joshi. 2022b. L3cube-mahacorpus and mahabert: Marathi monolingual corpus, marathi bert language models, and resources. In *LREC 2022 Workshop Language Resources and Evaluation Conference 20-25 June 2022*, page 97.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, NC Gokul, Avik Bhattacharyya, Mitesh M Khapra, and Pratyush Kumar. 2020. Indicnlpsuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, et al. 2021. Muril: Multilingual representations for indian languages. *arXiv preprint arXiv:2103.10730*.

Gregory Koch, Richard Zemel, Ruslan Salakhutdinov, et al. 2015. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2. Lille.

Atharva Kulkarni, Meet Mandhane, Manali Likhitkar, Gayatri Kshirsagar, Jayashree Jagdale, and Raviraj Joshi. 2022. Experimental evaluation of deep learning models for marathi text classification. In *Proceedings of the 2nd International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications: ICMISC 2021*, pages 605–613. Springer.

Anoop Kunchukuttan, Divyanshu Kakwani, Satish Golla, Avik Bhattacharyya, Mitesh M Khapra, Pratyush Kumar, et al. 2020. Ai4bharat-indicnlp corpus: Monolingual corpora and word embeddings for indic languages. *arXiv preprint arXiv:2005.00085*.

Sandeep Sricharan Mukku and Radhika Mamidi. 2017. Actsa: Annotated corpus for telugu sentiment analysis. In *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems*, pages 54–58.

Harshali B Patil and Ajay S Patil. 2017. Mars: a rule-based stemmer for morphologically rich language marathi. In *2017 international conference on computer, communications and electronics (Comptelix)*, pages 580–584. IEEE.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.