

CEYNEWS: A Sri Lankan Trilingual News Corpus

Ajinthha Sivakulasingam¹, Kengatharaiyer Sarveswaran¹, Tharindu Ranasinghe²

¹University of Jaffna, Sri Lanka, ²Lancaster University, UK

{sajintha,sarves}@univ.jfn.ac.lk, t.ranasinghe@lancaster.ac.uk

Abstract

Language models have achieved state-of-the-art performance across many natural language processing (NLP) tasks, but low-resource languages remain underrepresented due to limited pre-training data and evaluation benchmarks. While recent efforts have introduced corpora, models, and benchmarks for such languages, most focus on a single language in isolation. This is particularly limiting in multilingual countries such as Sri Lanka, where the lack of comparable multilingual resources hinders the development and fair evaluation of models across languages. To address this gap, we introduce CEYNEWS, the largest trilingual news corpus for Sinhala, Tamil, and English. The dataset contains over 1.02 million articles (136 million tokens) collected between 2013 and 2026 from three news sources, with full article text and rich metadata. Using CEYNEWS, we develop three downstream tasks: news source identification, news category classification, and headline generation, and evaluate multilingual models across different learning settings. Our results demonstrate the challenges of applying multilingual models to these tasks, particularly for the headline generation task. CEYNEWS fills a major gap in Sri Lankan and South Asian language resources and is publicly released to support research in multilingual NLP, LLM adaptation, and digital humanities.

1 Introduction

The rapid advancement of language models (LMs) (Devlin et al., 2019; Brown et al., 2020) has transformed natural language processing (NLP), achieving state-of-the-art performance across a wide range of tasks (Ghaseminejad Raeini, 2025). However, the success of LMs is heavily dependent on the availability of large-scale, high-quality pre-training corpora (de Gibert et al., 2024). As a result, their capabilities remain disproportionately concentrated in high-resource languages.

Low-resource languages, in contrast, continue to lag significantly behind (Joshi et al., 2020). This disparity is further amplified by the limited availability of curated benchmarks for low-resource languages, hindering reliable evaluation (Hu et al., 2020; Shardlow et al., 2024a,b). As a result, extending the benefits of language models to low-resource languages remains an open and pressing challenge (Hettiarachchi et al., 2026, 2025).

Addressing this, recent efforts have sought to develop language resources for low-resource languages, including corpora (Hettiarachchi et al., 2024; Charfi et al., 2019), pre-trained models (Dhananjaya et al., 2022; Evdaimon et al., 2024), and benchmarks (Micallef and Borg, 2025; Urbizu et al., 2022). However, most focus on individual languages in isolation, with comparable multilingual resources that enable cross-lingual pre-training and evaluation within a single linguistic or national context remain scarce. This is particularly problematic in multilingual societies, where unequal language representation can lead to weaker models and unfair evaluation across languages (Blasi et al., 2021).

In this work, we address this gap for Sri Lanka, a setting that presents a compelling case for multilingual resource development. In Sri Lanka, Sinhala and Tamil serve as official languages, with English functioning as a link language (Mendis and Rambukwella, 2020). Despite this official multilingualism, large-scale resources spanning all three languages remain scarce (Ranathunga et al., 2025). Sinhala has benefited from recent efforts to develop large-scale corpora (Hettiarachchi et al., 2024), benchmarks (Ranasinghe et al., 2025a,b,c), and language models (Dhananjaya et al., 2022). However, Tamil resources and models are predominantly focused on Indian Tamil (Unjum et al., 2025), resulting in degraded performance on Sri Lankan Tamil (Jeyarajalingam et al., 2026; Mahaganapathy et al., 2026; Thillainathan et al.,

2026). Although the two varieties are mutually intelligible, they exhibit systematic differences at the phonological, lexical, morphological, and syntactic levels, reflecting their distinct historical development and patterns of language (Unjum et al., 2025; Krishnamurti, 2003; Schiffman, 1999; Kailainathan, 1999).

These challenges highlight a broader limitation in Sri Lankan NLP, as Sinhala primarily suffers from limited data availability, while Sri Lankan Tamil remains underrepresented in existing multilingual resources. Moreover, most datasets, benchmarks, and models have been developed separately for individual languages, resulting in fragmented resources that inadequately support multilingual training and evaluation. This is particularly problematic in Sri Lanka, where public services, education, media, and everyday communication often involve Sinhala, Tamil, and English, with frequent code-switching that monolingual models fail to capture (Ranathunga et al., 2025). Therefore, there is a pressing need for cross-lingual corpora and unified benchmarks that represent all three languages within a shared national context.

In this paper, we introduce CEYNEWS, a unified trilingual news corpus for Sri Lanka, designed to support multilingual and cross-lingual NLP research under realistic national settings. It comprises over 1.02 million articles from three major Sri Lankan outlets—Ada Derana, Hiru, and ITN—spanning 2013–2026. Unlike Sinhala-only corpora such as NSINA (Hettiarachchi et al., 2024) or general-purpose multilingual document collections (Senaratna, 2026), CEYNEWS is purpose-built for benchmarked multilingual and cross-lingual NLP evaluation, providing standardised trilingual splits across Sinhala, Tamil, and English within a single framework. Our **main contributions** are as follows:

- (i) We introduce CEYNEWS, the largest trilingual, processable, news corpus for Sri Lanka to date, comprising over 1.02 million articles and 136 million tokens.¹
- (ii) We construct three CEYNEWS-based benchmarks: (i) news source identification, (ii) news category classification, and (iii) headline generation, with standardised train/validation/test splits covering all three languages.

- (iii) We evaluate state-of-the-art multilingual models, analysing in-language performance and cross-lingual transfer among the three languages, and provide insights into multilingual generalisation in a low-resource setting.

2 CEYNEWS Corpus

We first describe the corpus collection methodology, followed by a statistical analysis of CEYNEWS.

2.1 Data Collection

We selected three major Sri Lankan news outlets: Ada Derana, Hiru News, and ITN News, based on their coverage of both Sinhala and Tamil. Ada Derana and Hiru News also publish English-language news, enabling the construction of a trilingual corpus, while ITN News was used only for Sinhala and Tamil as no corresponding English-language news section was available.

For each available article, we collected the source, publication timestamp, headline, article content, URL, and news category when the category was available. We developed a separate crawling strategy for each outlet because their structures and content delivery mechanisms differed.

2.1.1 Ada Derana

Ada Derana articles were collected using a sequential, identifier-based crawling strategy. The scraper systematically generated article URLs by incrementing this identifier and retrieved each available article page. Direct HTTP requests to the website encountered a JavaScript-based browser validation mechanism. Therefore, Selenium WebDriver with headless Chrome was initially used to load a seed page and establish a valid browser session. The resulting session cookies were transferred to a persistent requests.Session, which was subsequently used to retrieve the public article pages more efficiently. When the session expired, or the browser validation page reappeared, the session-initialisation procedure was repeated. The timestamp, headline, and article content were extracted from each successfully retrieved page using BeautifulSoup. News categories were not collected because this information was not available on the individual article pages. Failed requests were retried using incremental backoff, and a checkpoint containing the most recently pro-

¹Our dataset is available at <https://zenodo.org/records/20579021>.

cessed identifier allowed interrupted crawling runs to resume without restarting the collection.

2.1.2 Hiru News

Two crawling strategies were used for Hiru News because its multilingual content was exposed through different interfaces. Sinhala and English articles were retrieved through paginated API endpoints that returned batches of articles in JSON format. The endpoints were queried using the page number and news category. From each response, we extracted the timestamp, headline, article content, URL, and category. Tamil articles were collected from individual article pages containing sequential numeric identifiers in their URLs. The scraper incremented these identifiers, retrieved each available page through HTTP requests, and extracted the timestamp, headline, article content, and category.

2.1.3 ITN News

ITN News articles were collected using a date-based archive traversal strategy. For every date within the predefined collection period, the scraper generated the corresponding daily archive URL and traversed all available pagination pages. Article links were first extracted from these index listings, after which each individual page was retrieved. The timestamp, headline, article content, category, and URL were extracted from each entry and stored in JSON format.

2.2 Data Preprocessing

After data collection, the extracted articles underwent a multi-stage preprocessing pipeline. Unicode NFC normalisation was applied to the headlines and article content, and selected Unicode artefacts introduced during web extraction were removed. Manually defined rule-based cleaning procedures were applied to remove WordPress shortcodes, embedded-media markup, residual HTML elements, URLs, duplicated timestamp headers, separator lines, source attributions, reporter bylines, and website-navigation remnants. Quotation marks and dashes were standardised, and redundant whitespace was collapsed.²

Records with missing or empty headlines or article content after preprocessing were excluded. Exact duplicates were identified using the normalized headline–content pair, and only one instance of

each duplicate was retained. All available metadata fields were preserved.

2.3 CEYNEWS Corpus Statistics

After preprocessing, the resulting CEYNEWS corpus contains approximately 1.026 million articles and 136 million words. We report corpus size in words to provide a model-independent measure that does not depend on the vocabulary or tokenization scheme of a particular language model. For the subsequent token-length analysis, we used the pretrained XLM-RoBERTa tokenizer as a common multilingual subword-tokenization framework (Conneau et al., 2020) because CEYNEWS is intended to support multilingual NLP and language-modeling applications.

To enhance the historical coverage of CEYNEWS and avoid redundant data collection, we also provide mappings to relevant portions of the existing NSINA corpus (Hettiarachchi et al., 2024). Specifically, Ada Derana Sinhala is mapped to NSINA articles published before 2017, while ITN Sinhala is mapped to articles published before 2023. Therefore, the Ada Derana and ITN Sinhala article counts reported in Table 1 include these incorporated NSINA articles. The table presents the overall distribution of articles across languages and sources, with Ada Derana and Hiru providing substantial coverage across all three languages.

To provide a more detailed characterization of the linguistic structure of the corpus, we analysed a random sample of 100,000 articles from each language. This analysis was performed on article content only, excluding headlines. Sentence segmentation and word-level tokenization were performed using Stanza³ for English and Tamil and Sinling⁴ for Sinhala. In this analysis, a token refers to a word-level token, and a unique token type refers to a distinct surface word form rather than a lemma. Table 2 reports the resulting sentence, token, vocabulary, and length statistics. The sampled English articles contain 18.64 million word-level tokens across 825,785 sentences, while the corresponding Tamil and Sinhala samples contain 10.09 million and 12.07 million tokens, respectively. English articles contain an average of 186.36 tokens and 8.26 sentences per article. The corresponding averages for Tamil are 100.93 tokens and 7.15

²The crawling and preprocessing code is publicly available at <https://github.com/LTG-UoJ/CeyNews>.

³<https://stanfordnlp.github.io/stanza/>

⁴<https://github.com/ysenarath/sinling>

Source	Tamil				Sinhala				English			
	From	To	Articles	Tokens	From	To	Articles	Tokens	From	To	Articles	Tokens
Ada Derana	2013	2025	158,289	21,343,134	2008	2025	193,190	26,752,342	2013	2025	78,745	18,086,084
Hiru	2017	2026	154,013	13,174,057	2013	2026	272,275	32,155,375	2013	2026	150,693	25,803,113
ITN	2018	2026	48,671	4,174,477	2022	2026	60,145	8,137,253				
Total			360,973	38,691,668			525,610	67,044,970			229,438	43,889,197

Table 1: Dataset statistics across news sources and languages. The Ada Derana Sinhala counts include NSINA articles published before 2017, while the ITN Sinhala counts include NSINA articles published before 2023 (Hettiarachchi et al., 2024).

sentences per article, while Sinhala articles contain 120.70 tokens and 5.96 sentences per article. At the sentence level, English has an average of 22.57 tokens per sentence, compared with 14.11 for Tamil and 20.26 for Sinhala. The median sentence lengths are 22 tokens for English, 13 for Tamil, and 18 for Sinhala. The sampled Tamil articles contain 514,069 unique surface token types, whereas the English sample contains 214,032 and the Sinhala sample contains 218,946.

For further analysis, we examined XLM-R subword-token length distributions and year-wise article coverage across Sinhala, Tamil, and English. The article distributions were based on data collected by our crawlers between 2013 and 2025. Sinhala articles from Ada Derana before 2017 and ITN before 2023 are excluded from the year-wise analysis, as these articles were obtained from the NSINA corpus. Figures 1 and 2 present the subword-token-length distributions of the headlines and article content, respectively. As shown, most headlines contain 12–20 tokens, whereas most articles contain approximately 75–225 tokens. Sinhala has the largest annual article volume, reaching around 48,000 articles in 2024, followed by Tamil, which peaks at approximately 42,000 articles in 2020. English contains fewer articles overall, with approximately 12,000–28,000 articles per year. These distributions demonstrate substantial coverage across all three languages while also highlighting differences in the volume and temporal availability of news content.

3 Benchmark Tasks

We derive three benchmark tasks using CEYNEWS that capture complementary aspects of news processing to show the utility of the corpus like CEYNEWS: (i) news source identification, (ii) news category classification, and (iii) news headline generation. The first two are classification tasks that assess whether models can capture

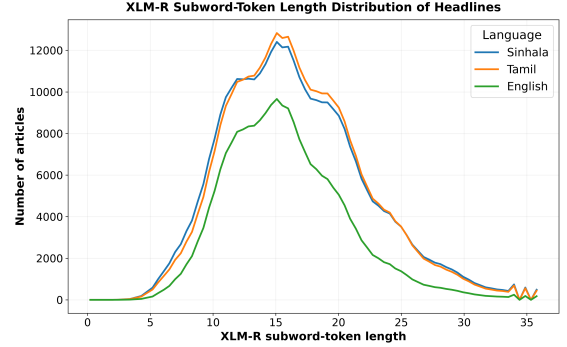


Figure 1: Token length distribution of headlines

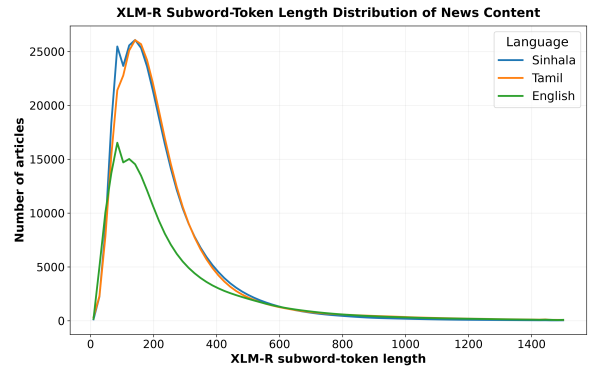


Figure 2: Token length distribution of contents

outlet-specific and topic-specific patterns across Sinhala, Tamil, and English, while the third evaluates generative ability by requiring models to produce concise headlines from full articles. Together, these tasks provide a realistic testbed for analysing both language understanding and generation in the Sri Lankan multilingual context.

The task definitions, models, and evaluation metrics were informed by the experimental design of the NSINA corpus (Hettiarachchi et al., 2024), while being adapted to the trilingual setting and task requirements of CEYNEWS.

For each task, we constructed fixed training, validation, and test splits and evaluated the models under three experimental settings:

Language	Sentences	Tokens	Unique Types	Sent./Article	Tok./Article	Tok./Sent.	Median Tok./Sent.
Tamil	715,429	10.09 M	514,069	7.15	100.93	14.11	13
Sinhala	595,661	12.07 M	218,946	5.96	120.70	20.26	18
English	825,785	18.64 M	214,032	8.26	186.36	22.57	22

Table 2: Linguistic statistics computed from 100,000 randomly sampled articles per language using article content only. Tokens refer to word-level tokens.

1. **Supervised fine-tuning:** Models are fine-tuned on the combined trilingual training data and evaluated separately on each language. We focus on multilingual transformer models
2. **Few-shot evaluation:** LLaMA 3.1-8B-Instruct is evaluated without task-specific fine-tuning to assess few-shot performance on the defined tasks.
3. **Cross-lingual transfer:** Finally, the best-performing model from the supervised fine-tuning is trained on one language and evaluated on another to assess transferability between Sinhala and Tamil.

The following subsections describe the tasks, the machine learning models trained for each task, and the corresponding results.

3.1 Task 1: News Source Identification

News source identification classifies articles by their originating outlet using textual content. It may capture outlet-specific linguistic patterns, reflecting differences in editorial policy, tone, and lexical choice.

3.1.1 Data

For news source identification, the input consists of the cleaned article content, and the target label represents the originating news outlet. Source labels were automatically derived from the provenance metadata recorded during data collection rather than manually annotated. Articles collected from the Ada Derana, Hiru News, and ITN News websites were assigned the corresponding source labels. To prevent the models from relying on explicit source indicators, URLs, source attributions, reporter bylines, and website-navigation remnants were removed from the article content during preprocessing.

We randomly sampled 39,680 articles from each available language–source combination, resulting in a total of 317,440 articles across eight language–source combinations. ITN News was

not included for English because it does not provide an English-language news section.

The data were divided into training, validation, and test sets using an 80/10/10 ratio with source stratified sampling. The data from all three languages were then combined, and the training set was randomly shuffled before model training.

3.1.2 Models & Experimental Setup

We evaluated four multilingual encoder models: XLM-R Base (Conneau et al., 2020), XLM-R Large (Conneau et al., 2020), InfoXLM Large (Chi et al., 2021), and mmBERT (Warner et al., 2025). All models were fine-tuned on the combined trilingual corpus for one epoch using a learning rate of 2×10^{-5} , a batch size of 16, and a maximum sequence length of 256. Based on trilingual fine-tuning performance, XLM-R Large was selected for cross-lingual transfer experiments, where it was trained on one language and evaluated on both in-language and cross-language test sets. We report Macro-F1 scores (Sokolova and Lapalme, 2009).

3.1.3 Results

Table 3 reports the Macro-F1 scores for news source identification. In the supervised fine-tuning setting, all multilingual encoders perform strongly, with XLM-R Large achieving the best results for Sinhala (0.967) and Tamil (0.812), and mmBERT performing best for English (0.949). These high scores across architectures suggest that news outlets exhibit strong outlet-specific structural and stylistic signatures – such as characteristic sentence length, punctuation conventions, formatting patterns, and recurring phrasing – that persist even after removing explicit source indicators during preprocessing, making the task highly learnable despite these controls.

Figure 3 shows confusion matrices for the best-performing model (XLM-R Large) across all three languages, providing insight into the lower Tamil performance noted above. For Sinhala, misclassifications are minimal and evenly rare across

all outlet pairs, consistent with its near-perfect Macro-F1. For Tamil, however, errors are concentrated asymmetrically: ITN News articles are frequently misclassified as Ada Derana (1,233 instances), substantially more than any other confusion pair, while Hiru News remains comparatively well-separated from both. This suggests that ITN and Ada Derana’s Tamil-language editions may share more similar journalistic style or content conventions than either shares with Hiru. For English, which includes only two sources, confusion is moderate and roughly symmetric between the two classes.

In contrast, cross-lingual transfer is substantially weaker: models perform well on the language they are trained on but degrade sharply when evaluated on other languages, indicating limited transferability for source-specific stylistic features. Notably, transfer between the two local languages is consistently stronger than transfer to English in both directions. This asymmetry indicates that Sinhala and Tamil source-identification features transfer more readily between each other than either does to English.

Setting	Model	Sinhala	Tamil	English
Fine-tuning	InfoXLM Large	0.964	0.789	0.926
	mmBERT	0.943	0.802	0.949
	XLM-R Base	0.964	0.799	0.931
	XLM-R Large	0.967	0.812	0.935
Cross-lingual	Sinhala Fine-tuned	0.968	0.345	0.273
	Tamil Fine-tuned	0.390	0.814	0.114

Table 3: Macro-F1 scores for news source identification

3.2 News Category Classification

News category classification assigns articles to predefined topical categories based on their content. It is a standard NLP task with applications in information organisation and retrieval (Bracewell et al., 2009).

3.2.1 Data

For this task, the article content was used as the model input, and the news category served as the target label. Category labels were derived from the metadata provided by the original news websites rather than manually annotated. Ada Derana articles were excluded from this task because category information was not available on its individual article pages.

Hiru News provided five category labels: Local, International, Sports, Business, and Entertainment.

ITN News used a broader and less consistent category scheme, including composite labels such as *Sports _ Local News* and *Breaking News _ International News*. We therefore applied a rule-based mapping to harmonise the source-specific labels into five common categories: Local, International, Sports, Business, and Entertainment. We then selected the four most frequent categories: (i) Local, (ii) International, (iii) Sports, and (iv) Business. Entertainment was excluded because it contained substantially fewer articles than the other categories.

For the Local, International, and Sports categories, we randomly selected 7,000 articles per language and category. For Business, all available Tamil and Sinhala articles were retained because of their limited availability. The resulting category-classification benchmark contains 81,309 articles in total.

The data were divided into training, validation, and test sets using an 80/10/10 ratio with category-stratified splitting. Data from all three languages were then combined, and the resulting multilingual training set was randomly shuffled before model training.

3.2.2 Models & Experimental Setup

We evaluated the same four multilingual encoder models used for source identification: XLM-R Base, XLM-R Large, mmBERT, and InfoXLM Large. All models were fine-tuned on the combined trilingual corpus for one epoch with a learning rate of 2×10^{-5} , a batch size of 16, and a maximum sequence length of 256. For cross-lingual transfer, we selected InfoXLM Large, the best-performing model in the trilingual fine-tuning setting.

3.2.3 Results

Table 4 reports the Macro-F1 scores for category classification. In the supervised fine-tuning setting, all models perform competitively across languages, with InfoXLM Large achieving the best scores for Sinhala (0.900) and Tamil (0.934), and mmBERT performing best for English (0.933).

Figure 4 shows confusion matrices for the best-performing model (InfoXLM Large) across all three languages. Sports remains almost perfectly separated in every language, consistent with its near-perfect recall. Local is the most error-prone true category across all languages, but its confusions are distributed differently in each: in Sinhala,

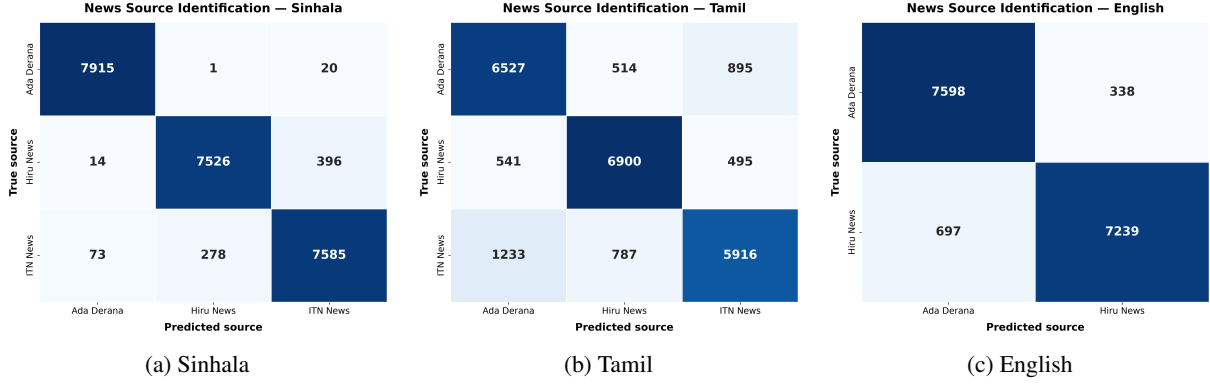


Figure 3: Confusion matrices for news source identification (XLM-R Large) across all three languages.

Local articles are most often misclassified as International, whereas in Tamil and English, Local errors are more evenly split between International and Business.

Cross-lingual transfer shows some degradation, but the drop is moderate compared to source identification, suggesting that category classification transfers better across languages. This is consistent with the per-class pattern above: Sports articles rely heavily on named entities and specialised terminology, including player names, team names, and tournament names that often remain recognisable even across scripts, whereas source identification depends on outlet-specific style, wording, and editorial conventions that are more language- and script-bound and therefore harder to transfer.

Setting	Model	Sinhala	Tamil	English
Fine-tuning	InfoXLM Large	0.900	0.934	0.930
	mmBERT	0.885	0.931	0.933
	XLM-R Base	0.894	0.928	0.928
	XLM-R Large	0.899	0.935	0.929
Cross-lingual	Sinhala	0.904	0.890	0.850
	Tamil	0.854	0.925	0.836

Table 4: Macro-F1 scores for news category classification

3.3 Headline Generation

Headline generation is a Natural Language Generation (NLG) task where the model generates a news headline from the full article content.

3.3.1 Data

For headline generation, the News Content field was used as the model input, and the corresponding Headline field was used as the target output. We extracted the headlines from the original news articles during data collection.

We used all available Tamil and English article-headline pairs, while the Sinhala dataset was down-sampled to approximately match the size of the Tamil dataset. The resulting benchmark contains 360,948 Sinhala articles (37.9%), 360,884 Tamil articles (37.9%), and 229,420 English articles (24.1%), giving a total of 951,252 article-headline pairs.

The data were divided into training, validation, and test sets using an 80/10/10 ratio. Data from all three languages were then combined, and the resulting multilingual training set was randomly shuffled before model training.

3.3.2 Models & Experimental Setup

We evaluated three multilingual sequence-to-sequence models: mBART-large-50 (Liu et al., 2020), mT5-base (Xue et al., 2021), and mT5-large (Xue et al., 2021). Models were fine-tuned on the combined trilingual corpus for one epoch with a learning rate of 1×10^{-5} , batch size of 8, maximum input length of 256, and maximum target length of 64. For cross-lingual transfer, we selected mT5-large. We report BLEU (Papineni et al., 2002) and Translation Edit Rate (TER) (Snoover et al., 2006).

3.3.3 Results

Table 4 reports BLEU and TER scores for headline generation. In the fine-tuned setting, mBART-50 performs best for Sinhala and Tamil, achieving the highest BLEU and lowest TER, while mT5-large performs best for English and shows stable performance across all three languages. Compared to the classification tasks above, absolute BLEU scores are low across all languages and models (0.13–0.23) even in the fine-tuned setting. This is expected given the nature of the task: unlike classification, where each example has a single cor-

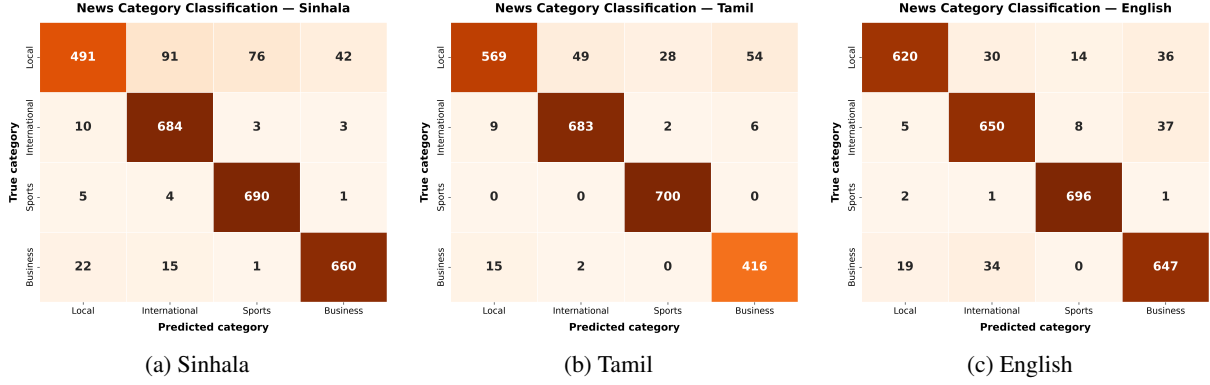


Figure 4: Confusion matrices for news category classification (InfoXLM Large) across all three languages.

rect label, headline generation is an open-ended summarisation task with many valid outputs for a given article, so n-gram overlap metrics such as BLEU and TER can penalise fluent, semantically appropriate headlines that do not closely match the exact reference wording. We report BLEU and TER as they are the standard metrics used in prior Sinhala headline generation work (Hettiarachchi et al., 2024). This limitation may be more pronounced for Sinhala and Tamil, whose morphologically rich, agglutinative structure allows a single word to take many valid inflected forms; exact lexical matching can therefore treat such variants as mismatches, which may contribute to the comparatively lower scores observed for these languages relative to English.

Cross-lingual transfer shows a substantial drop across all languages, with very low BLEU scores and high TER values. This indicates that headline generation is difficult to transfer across languages, likely because it requires not only content understanding but also language-specific generation and summarisation style or conventions. This is a stronger requirement than the pattern-matching underlying the classification tasks discussed earlier: whereas source identification and category classification rely on recognising structural or topical signals in text the model has already seen, headline generation requires producing fluent, grammatically well-formed output in a target language, which is substantially harder to achieve without in-language supervision.

3.4 Few-shot Evaluation

We evaluated a Llama model because its variants are openly available, relatively lightweight, and widely used in practical applications (Llama, 2023). Based on these considerations, we selected

Setting	Model	Sinhala		Tamil		English	
		BLEU	TER	BLEU	TER	BLEU	TER
Fine-tuned	mBART-50	0.167	0.946	0.234	0.811	0.151	0.926
	mT5-base	0.134	1.058	0.201	0.877	0.183	0.815
	mT5-large	0.150	0.998	0.219	0.838	0.205	0.779
Cross-lingual	Sinhala	0.019	1.830	0.032	2.364	0.025	2.240
	Tamil	0.021	2.123	0.037	2.416	0.027	2.826

Table 5: BLEU and TER scores for headline generation

Llama 3.1-8B-Instruct for few-shot in-context learning evaluation. We did not use SinLlama (Aravinda et al., 2025) because it is a Sinhala-specific pretrained model and no instruction-tuned variant was available for a comparable prompting-based evaluation. For each task and language, we evaluated Llama on a class-stratified subsample of 500 instances drawn from the same held-out test split used for the fine-tuned models. This sampling strategy was used to make the evaluation computationally tractable while maintaining representation across classes.

For classification tasks, the number of demonstrations was set to the number of classes, so that each class was represented once in the prompt. Accordingly, we used 4-shot prompts for category classification, 3-shot prompts for Sinhala and Tamil source identification, and 2-shot prompts for English source identification. For headline generation, which does not involve discrete class labels, we used a fixed 3-shot prompt. All prompts were constructed using the Llama instruction chat template.

The detailed few-shot evaluation results for news source identification, news category classification, and headline generation, are provided in Tables 6, 7, and 8, respectively, in Appendix A. The model performs relatively better on English, achieving a Macro-F1 of 0.737 for category classification and 0.545 for source identification, while

obtaining a BLEU score of 0.067 for headline generation. In contrast, performance on Sinhala and Tamil remains extremely low across the evaluated tasks.

For Sinhala and Tamil, the model often generates corrupted or unreadable character sequences, suggesting tokenisation and script-handling issues.

4 Research Directions

The findings from this study point to several directions for future research. Building on CEYNEWS, future work can explore the following directions:

- **Improved cross-lingual transfer:** The weak transfer observed in source identification and headline generation suggests the need for better language alignment methods. Future work could explore bi-encoder architectures, multilingual contrastive learning, and shared representation alignment techniques to reduce the gap between in-language and cross-lingual performance.
- **Comparative multilingual news analysis:** Since CEYNEWS contains comparable news articles across Sinhala, Tamil, and English, it can support social science research on how the same or related events are reported differently across languages, such as framing, emphasis, sentiment, terminology, and narrative structure across language communities.
- **Parallel and semi-parallel corpus construction:** Although CEYNEWS is large-scale and trilingual, it is comparable rather than parallel. Future work could use cross-lingual news alignment techniques to identify article pairs or clusters describing the same events, enabling more controlled multilingual and translation-oriented research.
- **Script-robust multilingual models:** The poor few-shot performance of Llama 3.1-8B-Instruct on Sinhala and Tamil, including corrupted or unreadable outputs, suggests limitations in tokenisation and script representation. This also opens an important direction for future research.

5 Conclusion

This paper introduced CEYNEWS, a large-scale trilingual Sri Lankan news corpus comprising over

1.02 million articles and 136 million tokens in Sinhala, Tamil, and English, collected from three major outlets between 2013 and 2026. We derived three benchmarks – news source identification, news category classification, and headline generation – and evaluated multilingual models under supervised fine-tuning, cross-lingual transfer, and few-shot settings. The results show that fine-tuned multilingual models perform strongly, but cross-lingual transfer remains task-dependent: category classification transfers relatively well, while source identification degrades sharply across languages. Few-shot evaluation with LLaMA 3.1-8B-Instruct further reveals serious script-level robustness issues for Sinhala and Tamil. By releasing CEYNEWS, we aim to support multilingual NLP, cross-lingual transfer research, and more equitable language technology development for Sri Lankan multilingual communities.

6 Limitations

The main contribution of this work is the CEYNEWS corpus and its standardised multilingual benchmark splits. The reported model results are intended as baseline results to demonstrate the usefulness of the corpus, rather than as state-of-the-art results. To ensure a controlled comparison, all models were fine-tuned for one epoch using fixed hyperparameters across languages. More extensive hyperparameter tuning and ablation studies are left for future work.

7 Ethical Considerations

We considered ethical and privacy aspects during the construction and release of the CEYNEWS corpus. Before crawling, we examined the publicly available robots.txt files of all three news sources. The collection followed the available crawling directives and did not intentionally access paths explicitly disallowed by these files.

To reduce the risk of author profiling, the publicly released dataset has been randomized and does not preserve the original chronological or author-based ordering of the articles.

Acknowledgment

Ajitha Sivakulasingam and Kengatharaiyer Sarveswaran acknowledge the support received from Google through the Research Scholar Program and from the University of Jaffna through its research grant for carrying out this research.

References

- HWK Aravinda, Rashad Sirajudeen, Samith Karunathilake, Nisansa de Silva, Rishemjit Kaur, and Surangika Ranathunga. 2025. Sinllama-a large language model for sinhala. In *2025 Moratuwa Engineering Research Conference (MERCon)*, pages 617–622. IEEE.
- Damián Blasi, Antonios Anastasopoulos, and Graham Neubig. 2021. [Systematic inequalities in language technology performance across the world’s languages](#). Preprint, arXiv:2110.06733.
- David B. Bracewell, Jiajun Yan, Fuji Ren, and Shingo Kuroiwa. 2009. [Category classification and topic discovery of japanese and english news articles](#). *Electronic Notes in Theoretical Computer Science*, 225:51–65. Proceedings of the Irish Conference on the Mathematical Foundations of Computer Science and Information Technology (MFCSIT 2006).
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. [Language models are few-shot learners](#). Preprint, arXiv:2005.14165.
- Anis Charfi, Wajdi Zaghouani, Syed Hassan Mehdi, and Esraa Mohamed. 2019. [A fine-grained annotated multi-dialectal Arabic corpus](#). In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 198–204, Varna, Bulgaria. INCOMA Ltd.
- Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. [InfoXLM: An information-theoretic framework for cross-lingual language model pre-training](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588, Online. Association for Computational Linguistics.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.
- Ona de Gibert, Graeme Nail, Nikolay Arefyev, Marta Bañón, Jelmer van der Linde, Shaoxiong Ji, Jaume Zaragoza-Bernabeu, Mikko Aulamo, Gema Ramírez-Sánchez, Andrey Kutuzov, Sampo Pyysalo, Stephan Oepen, and Jörg Tiedemann. 2024. [A new massive multilingual dataset for high-performance language technologies](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1116–1128, Torino, Italia. ELRA and ICCL.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Vinura Dhananjaya, Piyumal Demotte, Surangika Ranathunga, and Sanath Jayasena. 2022. [BERTifying Sinhala - a comprehensive analysis of pre-trained language models for Sinhala text classification](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7377–7385, Marseille, France. European Language Resources Association.
- Iakovos Evdaimon, Hadi Abdine, Christos Xypolopoulos, Stamatis Outsios, Michalis Vazirgiannis, and Giorgos Stamou. 2024. [GreekBART: The first pretrained Greek sequence-to-sequence model](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7949–7962, Torino, Italia. ELRA and ICCL.
- Mohammad Ghaseminejad Raeini. 2025. [The evolution of language models: From n-grams to llms, and beyond](#). *Natural Language Processing Journal*, 12:100168.
- Hansi Hettiarachchi, Damith Premasiri, Lasitha Randunu Chandrakantha Uyangodage, and Tharindu Ranasinghe. 2024. [NSina: A news corpus for Sinhala](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12307–12312, Torino, Italia. ELRA and ICCL.
- Hansi Hettiarachchi, Tharindu Ranasinghe, Alistair Plum, Paul Rayson, Ruslan Mitkov, Mohamed Medhat Gaber, Damith Premasiri, Fiona Anting Tan, and Lasitha Uyangodage. 2026. [Overview of the second workshop on language models for low-resource languages \(LoResLM 2026\)](#). In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages (LoResLM 2026)*, pages 651–661, Rabat, Morocco. Association for Computational Linguistics.
- Hansi Hettiarachchi, Tharindu Ranasinghe, Paul Rayson, Ruslan Mitkov, Mohamed Gaber, Damith Premasiri, Fiona Anting Tan, and Lasitha Randunu Chandrakantha Uyangodage. 2025. [Overview of the first workshop on language models for low-resource languages \(LoResLM 2025\)](#). In *Proceedings of the First Workshop on Language Models for*

- Low-Resource Languages*, pages 1–8, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: a massively multilingual multi-task benchmark for evaluating cross-lingual generalization. In *Proceedings of the 37th International Conference on Machine Learning*, ICML’20. JMLR.org.
- Varsha Jeyarajalingam, Menan Velayuthan, Sumirtha Karunakaran, Nivethiga Rasan, and Kengatharaiyer Sarveswaran. 2026. Evaluating linguistic knowledge of LLMs in Tamil: The ILAKKANAM benchmark. In *Proceedings of the Second Workshop on Challenges in Processing South Asian Languages (ChiPSAL 2026) @ LREC*, pages 176–189. European Language Resources Association (ELRA).
- Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. [The state and fate of linguistic diversity and inclusion in the NLP world](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.
- R Kailainathan. 1999. Linguistic heritage of sri lanka tamil.
- Bhadriraju Krishnamurti. 2003. *The dravidian languages*. Cambridge University Press.
- Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. [Multilingual denoising pre-training for neural machine translation](#). *Transactions of the Association for Computational Linguistics*, 8:726–742.
- Hugo Touvron Llama. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2407.21783*.
- Ahrane Mahaganapathy, Sumirtha Karunakaran, Kavitha Navakulan, and Kengatharaiyer Sarveswaran. 2026. [Bridging dialectal variation: A phonetic transcription tool for Tamil](#). In *Proceedings of the 13th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 234–241, Rabat, Morocco. Association for Computational Linguistics.
- Dushyanthi Mendis and Harshana Rambukwella. 2020. Sri lankan englishes. In *The Routledge handbook of world Englishes*, pages 178–193. Routledge.
- Kurt Micallef and Claudia Borg. 2025. [MELABenchv1: Benchmarking large language models against smaller fine-tuned models for low-resource Maltese NLP](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 20505–20527, Vienna, Austria. Association for Computational Linguistics.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Tharindu Ranasinghe, Isuri Anuradha, Damith Premasiri, Kanishka Silva, Hansi Hettiarachchi, Lasitha Uyangodage, and Marcos Zampieri. 2025a. [Sold: Sinhala offensive language dataset](#). *Language Resources and Evaluation*, 59(1):297–337.
- Tharindu Ranasinghe, Hansi Hettiarachchi, Constantin Orasan, and Ruslan Mitkov. 2025b. [MUSTS: Multilingual semantic textual similarity benchmark](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 331–353, Vienna, Austria. Association for Computational Linguistics.
- Tharindu Ranasinghe, Hansi Hettiarachchi, Nadeesha Chathurangi Naradde Vidana Pathirana, Damith Premasiri, Lasitha Uyangodage, Isuri Nanomi Arachchige, Alistair Plum, Paul Rayson, and Ruslan Mitkov. 2025c. [Sinhala encoder-only language models and evaluation](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8623–8636, Vienna, Austria. Association for Computational Linguistics.
- Surangika Ranathunga, Asanka Ranasinghea, Janaka Shamala, Ayodya Dandeniya, Rashmi Galapaththia, and Malithi Samaraweera. 2025. [A multi-way parallel named entity annotated corpus for english, tamil and sinhala](#). *Preprint*, arXiv:2412.02056.
- Harold F Schiffman. 1999. *A reference grammar of spoken Tamil*. Cambridge University Press.
- Nuwan I. Senaratna. 2026. [Sri lanka document datasets: A large-scale, multilingual resource for law, news, and policy](#). *Preprint*, arXiv:2510.04124.
- Matthew Shardlow, Fernando Alva-Manchego, Riza Batista-Navarro, Stefan Bott, Saul Calderon Ramirez, Rémi Cardon, Thomas François, Akio Hayakawa, Andrea Horbach, Anna Hülsing, Yusuke Ide, Joseph Marvin Imperial, Adam Nohejl, Kai North, Laura Occhipinti, Nelson Pérez Rojas, Nishat Raihan, Tharindu Ranasinghe, Martin Solis Salazar, and 2 others. 2024a. [An extensible massively multilingual lexical simplification pipeline dataset using the MultiLS framework](#). In *Proceedings of the 3rd Workshop on Tools and Resources for People with READING Difficulties (READI) @ LREC-COLING 2024*, pages 38–46, Torino, Italia. ELRA and ICCL.
- Matthew Shardlow, Fernando Alva-Manchego, Riza Batista-Navarro, Stefan Bott, Saul Calderon Ramirez, Rémi Cardon, Thomas François,

Akio Hayakawa, Andrea Horbach, Anna Hülsing, Yusuke Ide, Joseph Marvin Imperial, Adam Nohejl, Kai North, Laura Occhipinti, Nelson Pérez Rojas, Nishat Raihan, Tharindu Ranasinghe, Martin Solis Salazar, and 3 others. 2024b. [The BEA 2024 shared task on the multilingual lexical simplification pipeline](#). In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 571–589, Mexico City, Mexico. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linea Micciulla, and John Makhoul. 2006. [A study of translation edit rate with targeted human annotation](#). In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Marina Sokolova and Guy Lapalme. 2009. A systematic analysis of performance measures for classification tasks. *Information processing & management*, 45(4):427–437.

Adsajan Thillainathan, Nishanthini Kanthakumar, Nivethiga Rasan, and Kengatharaiyer Sarveswaran. 2026. TaLK-Corpus: A Regionally Diverse Evaluation Set for Sri Lankan Tamil Speech. In *Proceedings of Speech Language Models in Low-Resource Settings: Performance, Evaluation, and Bias Analysis (SPEAKABLE) at LREC2026*, pages 194–200. European Language Resources Association (ELRA).

Naveed Unjum, Stephanie Evert, Kengatharaiyer Sarveswaran, Ruvan Weerasinghe, and Nevidu Jayatilke. 2025. Llms for low-resource languages: A survey. *Available at SSRN 6122026*.

Gorka Urbizu, Iñaki San Vicente, Xabier Saralegi, Rodrigo Agerri, and Aitor Soroa. 2022. [BasqueGLUE: A natural language understanding benchmark for Basque](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1603–1612, Marseille, France. European Language Resources Association.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. [Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547, Vienna, Austria. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. [mT5: A massively multilingual pre-trained text-to-text transformer](#). In

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 483–498, Online. Association for Computational Linguistics.

A Few-shot Evaluation Results

This appendix provides the detailed results of the few-shot evaluation described in Section 3.4. The evaluation covers three tasks: headline generation, news category classification, and news source identification. Results are reported separately for Tamil, Sinhala, and English using the same evaluation settings described in the main paper. For headline generation, BLEU and Translation Edit Rate (TER) scores are reported, while Macro-F1 is reported for the two classification tasks.

Language	Macro-F1
Sinhala	0.0040
Tamil	0.0040
English	0.5448

Table 6: News source identification results in the few-shot setting.

Language	Macro-F1
Sinhala	0.0000
Tamil	0.0000
English	0.7366

Table 7: News category classification results in the few-shot setting.

Language	BLEU	TER
Sinhala	0.0004	7.2935
Tamil	0.0013	5.5479
English	0.0667	10.5605

Table 8: Headline generation results in the few-shot setting.

The TER scores should be interpreted together with the generated headline length and quality. The English model generally produces longer and more complete headlines, which can result in a higher number of edits when the generated headline differs in length from the reference. Although the generated headline is semantically related, its different wording and length lead to a higher TER. In contrast, the Sinhala and Tamil outputs are often very short or incomplete. Therefore, TER alone does not fully reflect the quality of the generated headlines.