CheckSent-BN: A Bengali Multi-Task Dataset for Claim Checkworthiness and Sentiment Classification from News Headlines

Pritam Pal and Dipankar Das
Jadavpur University, Kolkata, India
{pritampal522, dipankar.dipnil2005}@gmail.com

Abstract

This paper presents **CheckSent-BN** (Claim **Check**worthiness and **Sen**timent Classification in **B**engali **N**ews Headline), a novel multi-task dataset in Bengali comprising approximately 11.5K news headlines annotated for two critical natural language processing (NLP) tasks: claim checkworthiness detection and sentiment classification. To address the lack of high-quality annotated resources in Bengali, we employ a cost-effective annotation strategy that utilizes three large language models (GPT-4o-mini, GPT-4.1-mini, and Llama-4), followed by majority voting and manual verification to ensure label consistency. We provide benchmark results using multilingual and Bengali-focused transformer models under both single-task and multi-task learning (MTL) frameworks. Experimental results show that IndicBERTv2, BanglaBERT, and mDeBERTa model-based frameworks outperform other multilingual models, with IndicBERTv2 achieving the best overall performance in the MTL setting. CheckSent-BN establishes the first comprehensive benchmark for joint claim checkworthiness and sentiment classification in Bengali news headlines, offering a valuable resource for advancing misinformation detection and sentiment-aware analysis in low-resource languages. The CheckSent-BN dataset is available at: https://github.com/pritampal98/check-sent-bn

1 Introduction

In the contemporary digital landscape, news consumption patterns have undergone a significant transformation, with consumers increasingly accessing information from diverse sources through mobile and online platforms, instantly. (Samuels and Mcgonical, 2020). As per a report by Reuters, the shift towards digital news consumption has been particularly pronounced in India, where 71% of users prefer digital news over traditional media

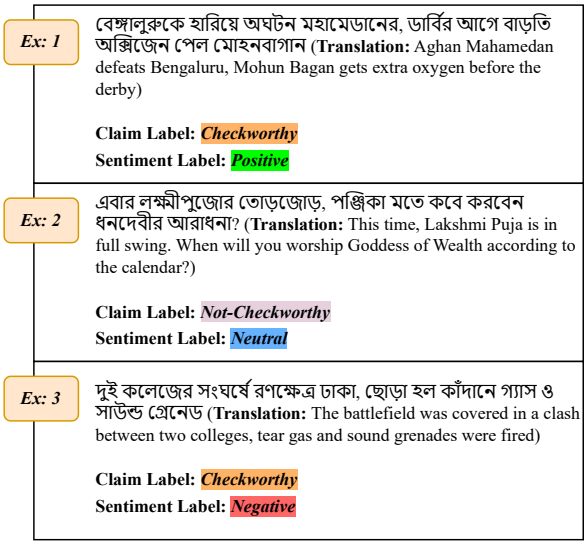

Figure 1: Example Bengali news headlines from the CheckSent-BN dataset with corresponding claim checkworthiness and sentiment labels (English translations provided)

in 2024[1]; this number increased to 76% in 2025[2].

Digital news organizations, on the other hand, frequently produce catchy and attention-grabbing headlines designed to maximize user engagement and encourage readers to click on their articles. These compelling headlines help boost user engagement rates and, consequently, revenue for the news organizations. However, this resulting information ecosystem, where revenue is the top-most priority, presents a significant challenge regarding the reliability and factual accuracy of news content.

Claim checkworthiness detection, the stepping stone towards fact-checking a given claim, gained significant progress in resource-rich languages such as English (Gencheva et al., 2017; Arslan et al., 2020; Nakov et al., 2018; Abumansour and Zubiaga, 2022; Majer and Šnajder, 2024) and Arabic (Jaradat et al., 2018; Abumansour

[1]https://bit.ly/reuters-digital-news-report-2024
[2]https://bit.ly/reuters-digital-news-report-2025

Proceedings of the Second Workshop on Bangla Language Processing (BLP-2025), pages 119–130

and Zubiaga, 2022). However, the detection of claim-checkworthiness in resource-constrained languages, such as Bengali, remains largely unexplored, particularly in the context of Bengali news headlines.

Bengali, with over 230 million native speakers globally, holds the distinction of being the sixth most spoken language worldwide (Alam et al., 2021). As the second most widely spoken language in India and the national language in Bangladesh, it confronts significant challenges in natural language processing (NLP) research due to its complex morphological structures, extensive use of compound words, and scarcity of high-quality annotated datasets, which have hindered the development of robust NLP systems. Despite some recent efforts in Bengali NLP, a critical gap remains in comprehensive datasets that address multiple classification tasks simultaneously.

The present article focuses on developing a multi-task Bengali news headline dataset with a primary focus on claim checkworthy detection, along with an additional task of sentiment classification. The key contributions in this paper can be summarized as follows:

- We present CheckSent-BN, the first comprehensive multi-task Bengali news headline dataset, comprising approximately 11.5K samples that focus on claim-checkworthiness identification and sentiment classification. An example from the dataset is shown in Figure 1.

- We proposed a cost-effective and faster way of data annotation in the resource-constrained Bengali language, utilizing three distinct large language models (LLMs): GPT-4o-mini (OpenAI et al., 2024), GPT-4.1-mini, and Llama-4 (Touvron et al., 2023). Further, we applied a majority voting scheme and manual revision to ensure annotation quality and consistency.

- We established baseline performance by developing multi-task learning (MTL) and single-task learning (STL) frameworks utilizing pre-trained multilingual transformer models, providing benchmarks for future research and exploration.

2 Related Work

This section provides an overview of recent research on claim checkworthiness detection and sentiment classification, with a specific focus on Bengali and low-resource languages.

2.1 Claim Detection

Recent advancements in artificial intelligence and NLP have enabled researchers to move from basic claim detection methods (Rosenthal and McKeown, 2012; Chakrabarty et al., 2019; Pathak et al., 2020; Gupta et al., 2021; Wührl and Klinger, 2021; Sundriyal et al., 2021; Gangi Reddy et al., 2022) to identifying the check-worthiness of claims (Jaradat et al., 2018; Wright and Augenstein, 2020). This progression has led to more advanced techniques, such as claim span identification (Sundriyal et al., 2022; Mittal et al., 2023), where specific phrases in a text that constitute a claim are pinpointed.

While significant strides have been made in high-resource languages like English, research on claim detection and claim check-worthiness identification is still limited in resource-constrained languages like Bengali. Early efforts by Dhar and Das (2021) introduced an expectation-maximization (EM) approach combined with principal component analysis (PCA) to identify claims in low-resource Indian languages, including Bengali. Their study revealed that dimensionality reduction techniques could improve classifier performance in resource-constrained contexts. Additionally, Rahman et al. (2025) created a claim detection dataset in Bengali, comprising 5,000 samples. They evaluated this dataset by incorporating traditional machine learning models with deep word embeddings, discovering that ensemble methods outperformed individual models, especially when using domain-specific Bangla embeddings.

On a broader scale, Dutta et al. (2023) released a multilingual dataset that includes English, Hindi, and Bengali, featuring factual claims extracted from Indian Twitter. Supporting efforts, such as those by Mittal et al. (2023), introduced multilingual claim span datasets based on Bengali social media texts, emphasizing span-level rather than sentence-level check-worthiness. Poddar et al. (2024) organized a shared task at the ICPR-2024 conference on multilingual claim span identification in Hindi and English tweets. Supporting this effort, Roy et al. (2025) extended the ICPR-2024 shared task dataset with more Hindi, English, Bengali, and CodeMix tweets annotated with claim spans. Recently, the CLEF-2025 CheckThat! Lab further integrated Bengali into global claim verification frameworks, which include subjectivity

detection and claim normalization (Alam et al., 2025).

2.2 Sentiment Classification

In recent years, researchers have focused on more efficient transformer-based approaches. For example, Islam et al. (2020) proposed a multilingual-BERT-based sentiment classification method. The authors combined multilingual BERT embeddings with LSTM, GRU, and CNN networks, demonstrating improved performance compared to traditional embeddings such as FastText and Word2Vec. Additionally, Bhowmick and Jana (2021) developed a transformer-based sentiment analysis technique by fine-tuning two transformer models: BERT and XLM-RoBERTa. A more recent study by Pal et al. (2025) introduced a multi-task learning framework for sentiment analysis and emotion recognition in Bengali text, utilizing three transformer models: mBERT, MuRIL, and IndicBERT.

Moreover, several resources for Bengali sentiment analysis have been developed by various researchers. Islam et al. (2021) proposed 'SentNoB', a comprehensive Bengali sentiment analysis dataset consisting of approximately 15.7K noisy Bengali texts, which were curated from prominent Bangladeshi news article comments and YouTube comments. Ahmed Masum et al. (2021) created 'BAN-ABSA', an aspect-based Bengali sentiment analysis dataset with nearly 9,000 samples. In addition to dataset development, the authors established baseline frameworks and evaluated them on their dataset using Bi-LSTM and CNN techniques. Islam and Alam (2024) developed a novel dataset called 'BanglaDSA', which comprises approximately 200K Bengali comments. Along with this dataset, the authors proposed a new approach that combines Skipgram with Bangla-BERT, outperforming all existing machine learning and deep learning methods. Rashid et al. (2024) introduced another Bengali sentiment analysis dataset with a sample size of 78K by collecting reviews from two popular e-commerce websites in Bangladesh. Furthermore, Islam and Masudul Alam (2023) conducted a study focused on Bengali reviews, developing a dataset of around 44K curated reviews from restaurant Facebook pages and groups. Kabir et al. (2023) created a Bengali book review dataset, which includes approximately 158K entries.

Furthermore, Hasan et al. (2023) organized a shared task on sentiment analysis in the Bengali language at the BLP workshop, where over 25 participants submitted systems ranging from traditional machine learning approaches to pretrained transformer models, as well as state-of-the-art LLM methodologies.

Although significant progress has been made in checkworthy claim detection and sentiment classification tasks, there remains a notable research gap in creating a multi-task Bengali dataset. Additionally, the use of LLMs for data annotation in resource-constrained languages, such as Bengali, has not been extensively explored. This research aims to address this gap by developing a larger Bengali claim check-worthy dataset containing ≈11,500 samples, along with sentiment labels. This new dataset surpasses the previously established Bengali claim check-worthy dataset by Rahman et al. (2025), which included only 5,000 samples.

3 Dataset Development

The development of our dataset was conducted in three distinct phases: (1) Data Collection, (2) Data Annotation utilizing multiple LLMs, and (3) Final Label Selection via Majority Voting.

3.1 Data Collection

Data were systematically collected from prominent online Bengali news portals, specifically Bartaman[3], Sangbad Pratidin[4], Ei Samay[5], and News18 Bangla[6]. News headlines were extracted from the main news page (home page) of each portal to ensure broad coverage across various domains and topics. The Python BeautifulSoup web-scraping library was used to collect these headlines daily. The data collection period lasted from August 10, 2024, to January 12, 2025. However, data collection for the Ei Samay news portal began later, on October 5, 2024. In contrast, the collection of news from the News18 Bangla portal was discontinued after a few days due to incomplete and low-quality headlines. All collected data were stored in an Excel spreadsheet for subsequent analysis.

Following the completion of data collection, duplicate entries were removed. Subsequently, all news headlines underwent a rigorous review by three computer science interns to identify and correct any inconsistencies, such as HTML tags or unrecognized characters. Finally, a total of 11,568

[3] https://bartamanpatrika.com/
[4] https://www.sangbadpratidin.in/
[5] https://eisamay.com/
[6] https://bengali.news18.com/

unique headlines were collected from the aforementioned news portals, including 4,610 headlines from Bartaman, 5,715 from Sangbad Pratidin, 858 from Ei Samay, and 385 from News 18 Bangla.

3.2 Data Annotation

The data annotation process was carried out using three distinct LLMs. While manual annotation is considered the gold standard for creating high-quality datasets, it often encounters significant challenges, particularly in resource-limited languages such as Bengali. These challenges include: (1) a scarcity of skilled annotators, (2) the time-consuming nature of the process, and (3) the prohibitively high overall cost.

To address these issues, we adopted a cost-effective and efficient approach to data annotation by leveraging LLMs. Given the state-of-the-art performance of LLMs across various NLP tasks, we utilized three specific models: GPT-4o-mini, Llama-4, and GPT-4.1-mini. Each model was provided with a concise prompt to annotate the claim checkworthiness and sentiment labels. Each LLM was accessed via its corresponding API, and the following prompt was provided to each LLM:

```
You are an efficient language model that can do
many tasks. Now act as a professional Bengali data
annotator. You have provided a news headline in the
Bengali language. Your task is to:

1. Identify news headline claim is Checkworthy
or Not.
Checkworthy (Label = 1): Headlines that contain
factual claims needing verification
Not Checkworthy (Label = 0): Headlines that are
opinions, questions, satire, etc.

2. Identify sentiment in news headline as
'POSITIVE', 'NEGATIVE', or 'NEUTRAL'.
POSITIVE: Expresses an optimistic, successful,
complimentary, or positive outlook.
NEGATIVE: Expresses loss, criticism, failure, or a
negative outlook.
NEUTRAL: Neutral or simply informational, no emotional
coloring.

Now annotate the above-mentioned annotations in
the following news headline.
NEWS HEADLINE: {txt}

Note that only provide the exact annotation
tags in list format (Eg, [Claim Checkworthy Label,
'Sentiment Label']), no extra explanation is needed.
For clear understanding, I am providing you with some
examples.

ANNOTATION EXAMPLES:
========================
** 12 unique annotation examples were provided (See
Appendix A)**
```

This approach ensured a scalable, efficient method for annotating news headlines with claim checkworthiness and sentiment labels, which were subsequently aggregated via majority voting to select the final annotated label.

Although strict instructions were given to the LLMs to provide only claim and sentiment labels, in some instances (approximately 150 headlines), Llama-4 provided explanatory results alongside the labels. These samples were manually identified and formatted adequately by the authors.

3.3 Final Annotated Label Selection from Three LLMs' Annotated Labels

Following the annotation by three distinct LLMs, the final label for each data point was determined through majority voting. For both claim checkworthiness and sentiment, the label with the most frequent outcome was selected as the final label.

The inter-annotator agreement among the three LLM annotators was evaluated using both Fleiss' Kappa (Fleiss, 1971) and Gwet's AC1 (Gwet, 2006) metrics. For assessing the claim checkworthiness, the Fleiss' Kappa score was 0.45, while Gwet's AC1 score was 0.83. In terms of sentiment annotation, the Fleiss' Kappa score was 0.63, and Gwet's AC1 score was 0.67. These scores generally indicate a good level of agreement among the different LLM annotators.

Instances where no majority label was found through automated voting were meticulously analyzed and annotated by the authors. Additionally, all LLM-annotated data, following majority voting, underwent a thorough review by three undergraduate computer science interns. Headlines identified by the interns as inconsistent were further reviewed and resolved by the authors wherever applicable. A flow diagram of the overall data annotation process is provided in Figure 2. The distribution of claim checkworthy and sentiment labels for training and testing splits is provided in Table 1.

	Label	#Train	#Test
Claim	Check-worthy	8143	2030
	Not Check-worthy	1111	284
Sentiment	Negative	4573	1123
	Neutral	3187	825
	Positive	1494	366

Table 1: Distribution of claim checkworthiness and sentiment labels in train and test splits of CheckSent-BN.

4 Methodology

This section presents a comprehensive methodology for classifying claim-checkworthiness and sentiment for the CheckSent-BN dataset. Given

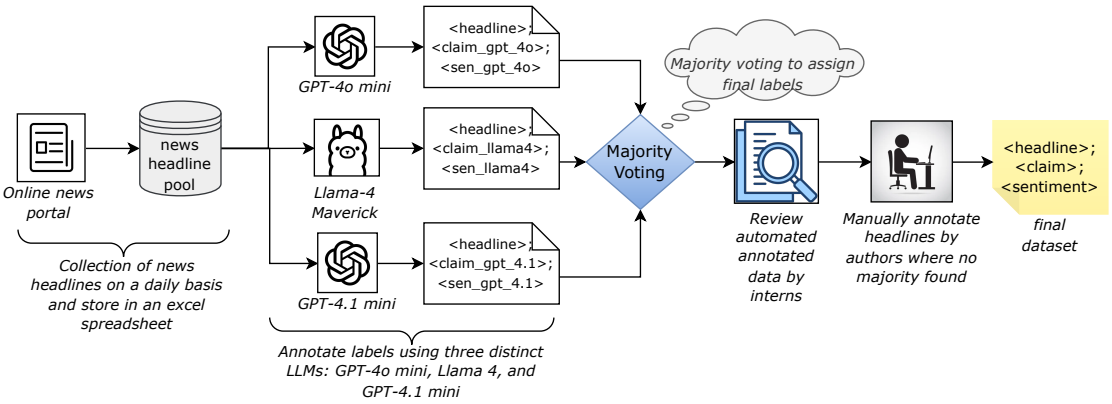

Figure 2: Overview of the annotation pipeline for CheckSent-BN, including data collection, annotation with three LLMs, majority voting, and manual verification.

a news headline S, our main objective is to develop an MTL framework that can classify claims as checkworthy or not and identify the sentiment as positive, negative, or neutral in S using a single neural network. We experimented with several transformer-based pre-trained models, ranging from the lightweight multilingual DistilBERT (mDistilBERT) to Indian language (including Bengali) focused models such as IndicBERT, BanglaBERT, etc. The overall system framework is demonstrated in Figure 3.

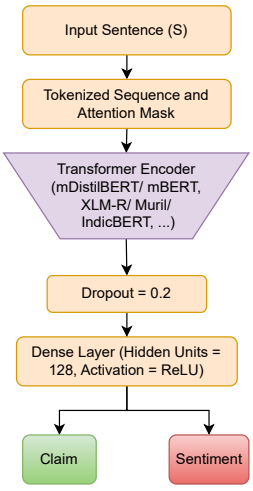

Figure 3: Flow diagram of the MTL framework. Pre-trained multilingual and Bengali-focused transformers are fine-tuned jointly for claim checkworthiness and sentiment classification.

Tokenization: Before proceeding to framework development and training, all the news headlines were tokenized into a sequence of tokens. The tokenization was performed using the corresponding pre-trained model's tokenizer, for example, for mBERT, the BERT tokenizers were used, and so on. The tokenizers were imported using the HuggingFace API. All the news headlines were tokenized to a fixed sequence length of 50 tokens, and the Input IDs (numeric representation of tokens) and attention masks were stored for further processing.

Framework Description: A diverse range of pre-trained multilingual transformer models was chosen to train and fine-tune the MTL framework, including lightweight mDistilBERT (Sanh et al., 2019), powerful mBERT (Devlin et al., 2018), XLM-RoBERTa (Conneau et al., 2019), mDeBERTa (He et al., 2021), and language-focused models such as BanglaBERT (Bhattacharjee et al., 2022) for Bengali, and MuRIL (Khanuja et al., 2021), IndicBERT (Kakwani et al., 2020), and IndicBERTv2 (Doddapaneni et al., 2023) for Indian languages. While mBERT and XLM-RoBERTa are trained on more than 100 languages, including Bengali, and excel in the majority of benchmark datasets, MuRIL, IndicBERT, and IndicBERTv2 are trained explicitly on Indian languages, including Bengali, allowing them to understand indian contexts accurately than other multilingual transformer models.

On the other hand, the BanglaBERT was explicitly trained on the Bengali language only, which allows it to understand the Bengali language more effectively than other models.

In the transformer models, the input IDs and attention masks, which were obtained from the corresponding transformer model's tokenizer function, were provided as inputs to the models. The pooler output from the transformer models, which is a learned linear transformation followed by a $tanh$ activation function applied to the last hidden state representation of the special starting token in the transformer models, was further passed through a

dropout layer with a dropout rate of 0.2.

Next, the output of the dropout layer was passed through a dense layer of 128 hidden units with the ReLU activation function.

$$\mathbf{Z_{dense}} = \text{ReLU}(\mathbf{Z_{dropout}})$$

Here, $\mathbf{Z_{dropout}}$ represents the output of the dropout layer, and $\mathbf{Z_{dense}}$ represents the output of the dense layer.

Classification: For multi-task classification, the output of the dense layer ($\mathbf{Z_{dense}}$) was passed through two task-specific output layers as depicted in Figure 3. The first layer was used for claim classification, employing two hidden units, and the second layer was for sentiment classification, which used three hidden units. Both the output layers used softmax as their activation function.

$$\mathcal{P}_* = softmax(\mathbf{Z_{dense}})$$
$$\hat{\mathcal{Y}}_* = \underset{j}{argmax}(P_*)$$

Here, $\mathcal{P}_*$ denotes the probability value for each class in a classification task, $\hat{\mathcal{Y}}_*$ indicates the predicted class value, and j represents the number of classes involved in the classification task.

Training: To train the framework, the training data was first divided into a 9:1 ratio, where 90% of the data was used for training and the remaining 10% was used as a validation set. The `SparseCategoricalCrossEntropy` loss function was used during training, and the loss was monitored for the validation set.

$$\mathcal{L}_{total} = \mathcal{L}_{claim} + \mathcal{L}_{sentiment}$$

Here, $\mathcal{L}_{claim}$ and $\mathcal{L}_{sentiment}$ represent the loss for claim checkworthiness and sentiment classification tasks, and $\mathcal{L}_{total}$ is the overall loss function. The primary objective during training was to minimize the total loss ($\mathcal{L}_{total}$) for the validation data.

The AdamW optimizer (Loshchilov and Hutter, 2019) was selected for optimization, utilizing a weight decay of 0.01 and an epsilon value of 1e-8. The learning rate was set to 2e-5, and the frameworks were trained for a maximum of ten epochs with a batch size of 16.

5 Experiment and Result

This section provides a brief overview of the experimental setup and the outcomes obtained from the experiments.

5.1 Experimental Setup

All experiments were conducted using TensorFlow and Keras, and the models were trained on an NVIDIA RTX 5000 GPU. The experiments were employed in two setups: an MTL configuration, where the tasks of claim checkworthiness identification and sentiment classification shared the same neural network with two classification heads, and a single-task learning (STL) configuration, where separate neural networks were developed for each task, with each network featuring a single classification head. A diagrammatic representation of the STL framework is provided in Appendix C.

To ensure a fair comparison, all MTL and STL frameworks were trained with the same hyperparameters as mentioned in Section 4. For evaluation, the precision, recall, and macro F1-score were calculated on the test dataset.

5.2 Result

Claim Checkworthiness Identification: The results of the claim checkworthiness identification are summarized in Table 2. It is observed from this table that the IndicBERTv2 model excels all other models in both the STL and MTL frameworks, achieving F1-scores of 82.91 and 83.86, respectively. Additionally, the majority of transformer models (mDistilBERT, mBERT, mDeBERTa, BanglaBERT, IndicBERT, and IndicBERTv2) demonstrate improved claim identification results with the MTL framework compared to their STL counterparts.

Notably, the IndicBERTv2 model exhibits a performance enhancement of 1.13% in the MTL framework compared to its STL framework. The other transformer model-based multi-task learning frameworks, including mDistilBERT, mBERT, mDeBERTa, BanglaBERT, and IndicBERT, show performance improvements of 2.07%, 5.15%, 1.05%, 2.48%, and 0.09%, respectively, compared to their corresponding single-task learning frameworks. However, a slight decline in performance is observed for the XLM-RoBERTa and MuRIL models in the MTL framework, with F1-score drops of 1.69% and 2.65%, respectively.

Sentiment Classification: The results for sentiment classification are presented in Table 3. Similarly, for claim checkworthiness identification in the context of sentiment classification, the IndicBERTv2 model outperforms other transformer models. Moreover, the IndicBERTv2-based MTL

	STL			MTL		
	Prec.	**Rec.**	**F1.**	**Prec.**	**Rec.**	**F1.**
mDistilBERT	78.64	69.37	72.73	73.69	74.90	74.27
mBERT	77.61	68.12	71.45	77.30	73.74	75.33
XLM-R	79.42	77.61	78.47	82.97	73.52	77.14
mDeBERTa	84.80	78.37	81.13	83.72	80.49	81.99
BanglaBERT	87.91	75.77	80.27	**86.21**	79.40	82.31
MuRIL	83.57	76.07	79.15	85.23	72.63	77.06
IndicBERT	79.55	65.29	69.30	79.81	65.31	69.36
InidcBERTv2	86.13	80.38	82.91	85.48	**82.43**	**83.86**

Table 2: Performance of transformer models on claim checkworthiness detection under STL and multi-task MTL frameworks. (Best Precision, Recall, and F1-score are in bold font)

framework demonstrates an improvement of 1.26% in terms of F1-score compared to the IndicBERTv2-based STL framework.

Regarding other transformer models, XLM-RoBERTa, mDeBERTa, BanglaBERT, MuRIL, and IndicBERT yielded better results within the MTL framework, with F1-score enhancements of 0.93%, 0.45%, 1.53%, 1.3%, and 0.47%, respectively, compared to their corresponding STL counterparts. In contrast, mDistilBERT and mBERT did not perform as effectively for sentiment classification in the multi-task learning scenario, resulting in similar or lower performance compared to the STL frameworks, with performance drops of 0.87% and 0.38%, respectively.

	STL			MTL		
	Prec.	**Rec.**	**F1.**	**Prec.**	**Rec.**	**F1.**
mDistilBERT	65.33	66.24	65.46	64.75	65.36	64.89
mBERT	71.43	66.15	67.37	69.42	66.07	67.11
XLM-R	74.07	74.73	74.35	74.84	75.30	75.05
mDeBERTa	72.02	78.37	78.64	79.91	78.27	79.00
BanglaBERT	82.63	77.62	78.98	80.00	80.11	80.05
MuRIL	75.73	77.40	75.79	76.23	77.51	76.79
IndicBERT	64.03	61.09	61.78	65.61	60.85	62.07
IndicBERTv2	80.39	**81.09**	80.82	**83.70**	80.02	**81.52**

Table 3: Performance of transformer models on sentiment classification under STL and MTL frameworks. (Best Precision, Recall, and F1-score are in bold font)

5.3 Observations

Upon performing all the experiments and analysing the results, a few observations are made:

First, it is observed that for both checkworthy claim identification and sentiment classification, the IndicBERTv2 demonstrates impressive results, irrespective of the STL and MTL frameworks. Additionally, the BanglaBERT and mDeBERTa models indicate a strong performance compared to other transformer models. This improvement suggests a better contextual understanding of the Bengali language compared to other transformer models.

Second, the joint learning of claim-checkworthiness detection and sentiment classification leads to more effective identification of claim-checkworthy sentences than using STL classifiers in most cases. This observation shows that the sentiment classification task effectively assists in identifying claim checkworthiness within an MTL environment.

Third, the performance of the IndicBERT model across all tasks and frameworks is relatively low. One possible reason for this is that IndicBERT was trained on the ALBERT model using 12 Indian languages, resulting in a smaller and more lightweight model compared to other transformer models, such as mBERT, MuRIL, and XLM-RoBERTa. As a result, it may struggle to identify nuanced contexts in text, leading to lower performance compared to other transformer-based frameworks.

6 Error Analysis

Due to a diverse range of experiments with different transformer-based models in both MTL and STL frameworks, the error analysis for each model is a challenging task. Therefore, to simplify the error analysis, we conducted the error analysis between the best-performing models (i.e., IndicBERTv2) for both MTL and STL frameworks. To conduct the error analysis, we calculated the confusion matrices for each task in each framework.

Claim Checkworthiness Detection: The confusion matrices for claim checkworthiness detection are provided in Figure 4. Although the IndicBERTv2-based MTL framework demonstrates strong overall performance, it slightly lacks in identifying claim-checkworthy sentences. Out of 2030 claim checkworthy instances, the MTL framework identifies 96.9% sentences as claim checkworthy, whereas the STL identifies claim checkworthy texts with 97.4% accuracy.

On the other hand, the STL framework correctly identifies 63.4% of the 284 non-claim checkworthy instances, while the MTL framework achieves 68%, demonstrating better performance.

Sentiment Classification: The confusion matrices for sentiment classification for IndicBERTv2-based STL and MTL frameworks are provided in Figure 5. The confusion matrices indicate that

ID	News Headline	True Label		Predicted STL		Predicted MTL	
		Claim	Sentiment	Claim	Sentiment	Claim	Sentiment
S_1	বছর শেষে ছুটির আমেজ, ভিড়ে ঠাসা দিঘা থেকে দার্জিলিং, দেখুন ছবি (**Translation:** End of year holiday atmosphere, crowded Digha to Darjeeling, see photos)	n-claim	neutral	claim	positive	n-claim	positive
S_2	'সিনেমাপাড়ার একটাই স্বর, জাস্টিস ফর RG Kar', পথে নামল টলিউড (**Translation:** 'Cinemapara has one voice, Justice for RG Kar', Tollywood takes to the road)	claim	neutral	no-claim	positive	claim	neutral
S_3	খসছে পদ্মের পাঁপাড়ি! উপনির্বাচনে ৬-এ শূন্য পেয়ে কত দাঁড়াল বিজেপির বিধায়ক সংখ্যা? (**Translation:** The lotus petals are falling! How many MLAs did BJP have after getting zero in 6 by-elections?)	claim	negative	claim	neutral	claim	negative
S_4	২ লক্ষ কোটি টাকার কুম্ভ ইকনমি! হিন্দুত্বকে সামনে রেখে ঢালাও ব্যবসা যোগীরাজ্যে, আশায় বুক বাঁধছে বণিকসভা (**Translation:** Kumbh Economy of 2 lakh crore rupees! Keeping Hindutva in the forefront, business will pour into Yogi Rajya, the Chamber of Commerce is full of hope)	claim	positive	claim	positive	claim	negative

Table 4: Example misclassifications from IndicBERTv2 models in STL and MTL settings. Each row displays the news headline, gold labels (true labels), and predicted labels (with an English translation provided).

Figure 4: Confusion matrices for claim checkworthiness detection using IndicBERTv2 in STL and MTL setups.

Figure 5: Confusion matrices for sentiment classification using IndicBERTv2 in STL and MTL setups.

the IndicBERTv2-based STL framework achieves 94.1% and 72.8% accuracy in identifying negative and neutral sentiments, respectively. In contrast, the IndicBERTv2-based MTL framework achieves accuracies of 94.7% and 77.8%, respectively, resulting in a decrease in the number of misclassified cases for these categories.

Conversely, the STL framework outperforms the MTL framework in identifying positive sentiments, achieving an accuracy of 76.8%, compared to the MTL framework's 67.5%. This suggests that STL is more effective at recognizing positive sentiments, although MTL exhibits better performance in the negative and neutral classes.

Along with the confusion matrix, a few examples of error cases are provided in Table 4. For instance, in example S_1, whereas the original claim was non-checkworthy, the IndicBERTv2-based STL framework incorrectly classified it as claim-checkworthy. Additionally, both STL and MTL frameworks mistakenly classify the sentiment as 'positive' where the original sentiment was 'neutral'. While the news headline in S_1 carries a slight positive emotional tone, it primarily describes a factual situation: "*year-end holiday atmosphere*". The phrase "*crowded Digha to Darjeeling*" is a neutral description of a high-traffic situation, not necessarily negative or positive. Therefore, the 'neutral' sentiment is justified for the headline. However, neither the STL nor the MTL framework captures these contextual nuances in the text and incorrectly categorizes it as 'positive'.

Considering another example, S_3, the original sentiment in the news headline is negative, indicating a poor performance of the BJP (a political party in India) in the bi-election results. In this case, the STL framework incorrectly predicts the text as "negative," whereas the MTL framework successfully identifies its sentiment label. This suggests that the joint learning of claim, sentiment, and news content enables the MTL framework to determine the sentiment label accurately.

7 Conclusion

This paper proposes 'CheckSent-BN', a dataset comprising 11,568 instances annotated with labels for two distinct tasks: claim checkworthiness identification and sentiment classification. We developed two baseline frameworks: the STL frame-

work, with one classification head for each task, and the MTL framework, which has two classification heads. We experimented with eight different multilingual transformer models, and the experimental results show that the IndicBERTv2, BanglaBERT, and mDeBERTa model-based frameworks demonstrate a strong performance over all classification tasks. Notably, the IndicBERTv2-based MTL framework achieved the best performance across all classification tasks.

Future directions include expanding the dataset's sample size, comparing LLM annotations with human annotations, and adding labels such as 'clickbait' or 'sarcasm' to enhance the dataset's scope.

Limitations

Our proposed work has several potential limitations. First, the annotation of claim checkworthy and sentiment labels was conducted using three LLMs. While we performed a superficial manual verification with three computer science interns, relying on LLMs for labeling may compromise the overall quality of the dataset. In future work, we aim to hire professional Bengali data annotators to ensure more accurate verification of these labels.

Second, we utilized the mini variants of the GPT models, specifically GPT-4o-mini and GPT-4.1-mini, for cost-effectiveness. Although these models can adequately annotate claim checkworthy and sentiment labels, the "mini" variants do not fully leverage the capabilities of the full GPT models.

Third, there is a significant imbalance in the claim checkworthy labels: 9,254 are labeled checkworthy, while only 2,314 are labeled non-checkworthy. This imbalance can lead to bias in our MTL and STL frameworks due to the predominance of annotated data. We plan to address this issue in future work by developing a more balanced dataset.

Fourth, the news headlines used in our study were curated from prominent news websites focused on the state of West Bengal, India. However, there are other Bengali-speaking regions in India, such as Tripura and parts of Assam, that have their own regional newspapers in Bengali, which are not included in our current work. Additionally, news headlines from Bangladeshi news portals were also excluded. In future work, we intend to incorporate Bengali news headlines from these other areas, including Tripura and Bangladesh, to broaden the dataset's scope.

Acknowledgment

This work was supported by the Defence Research and Development Organisation (DRDO), New Delhi, under the project "Claim Detection and Verification using Deep NLP: an Indian perspective".

References

Amani S. Abumansour and Arkaitz Zubiaga. 2022. Check-worthy claim detection across topics for automated fact-checking. *Preprint*, arXiv:2212.08514.

Mahfuz Ahmed Masum, Sheikh Junayed Ahmed, Ayesha Tasnim, and Md. Saiful Islam. 2021. Banabsa: An aspect-based sentiment analysis dataset for bengali and its baseline evaluation. In *Proceedings of International Joint Conference on Advances in Computational Intelligence*, pages 385–395, Singapore. Springer Singapore.

Firoj Alam, Arid Hasan, Tanvirul Alam, Akib Khan, Janntatul Tajrin, Naira Khan, and Shammur Absar Chowdhury. 2021. A review of bangla natural language processing tasks and the utility of transformer models. *Preprint*, arXiv:2107.03844.

Firoj Alam, Julia Maria Struß, Tanmoy Chakraborty, Stefan Dietze, Salim Hafid, Katerina Korre, Arianna Muti, Preslav Nakov, Federico Ruggeri, Sebastian Schellhammer, Vinay Setty, Megha Sundriyal, Konstantin Todorov, and Venktesh V. 2025. The clef-2025 checkthat! lab: Subjectivity, fact-checking, claim normalization, and retrieval. *Preprint*, arXiv:2503.14828.

Fatma Arslan, Naeemul Hassan, Chengkai Li, and Mark Tremayne. 2020. A benchmark dataset of check-worthy factual claims. *Preprint*, arXiv:2004.14425.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Ahmad, Kazi Samin Mubasshir, Md Saiful Islam, Anindya Iqbal, M. Sohel Rahman, and Rifat Shahriyar. 2022. BanglaBERT: Language model pretraining and benchmarks for low-resource language understanding evaluation in Bangla. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1318–1327, Seattle, United States. Association for Computational Linguistics.

Anirban Bhowmick and Abhik Jana. 2021. Sentiment analysis for Bengali using transformer based models. In *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, pages 481–486, National Institute of Technology Silchar, Silchar, India. NLP Association of India (NLPAI).

Tuhin Chakrabarty, Christopher Hidey, and Kathy McKeown. 2019. IMHO fine-tuning improves claim detection. 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

558–563, Minneapolis, Minnesota. 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. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Rudra Dhar and Dipankar Das. 2021. Leveraging expectation maximization for identifying claims in low resource Indian languages. In *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, pages 307–312, National Institute of Technology Silchar, Silchar, India. NLP Association of India (NLPAI).

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2023. Towards leaving no Indic language behind: Building monolingual corpora, benchmark and models for Indic languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12402–12426, Toronto, Canada. Association for Computational Linguistics.

Subhabrata Dutta, Rudra Dhar, Prantik Guha, Arpan Murmu, and Dipankar Das. 2023. A multilingual dataset for identification of factual claims in indian twitter. In *Proceedings of the 14th Annual Meeting of the Forum for Information Retrieval Evaluation*, FIRE '22, page 88–92, New York, NY, USA. Association for Computing Machinery.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Revanth Gangi Reddy, Sai Chetan Chinthakindi, Yi R. Fung, Kevin Small, and Heng Ji. 2022. A zero-shot claim detection framework using question answering. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6927–6933, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Pepa Gencheva, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, and Ivan Koychev. 2017. A context-aware approach for detecting worth-checking claims in political debates. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 267–276, Varna, Bulgaria. INCOMA Ltd.

Shreya Gupta, Parantak Singh, Megha Sundriyal, Md. Shad Akhtar, and Tanmoy Chakraborty. 2021. LESA: Linguistic encapsulation and semantic amalgamation based generalised claim detection from online content. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3178–3188, Online. Association for Computational Linguistics.

Kilem Li Gwet. 2006. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48.

Md. Arid Hasan, Firoj Alam, Anika Anjum, Shudipta Das, and Afiyat Anjum. 2023. BLP-2023 task 2: Sentiment analysis. In *Proceedings of the First Workshop on Bangla Language Processing (BLP-2023)*, pages 354–364, Singapore. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

Khondoker Ittehadul Islam, Md Saiful Islam, and Md Ruhul Amin. 2020. Sentiment analysis in bengali via transfer learning using multi-lingual bert. In *2020 23rd International Conference on Computer and Information Technology (ICCIT)*, pages 1–5.

Khondoker Ittehadul Islam, Sudipta Kar, Md Saiful Islam, and Mohammad Ruhul Amin. 2021. SentNoB: A dataset for analysing sentiment on noisy Bangla texts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3265–3271, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Md. Shymon Islam and Kazi Masudul Alam. 2024. Sentiment analysis of bangla language using a new comprehensive dataset bangdsa and the novel feature metric skipbangla-bert. *Natural Language Processing Journal*, 7:100069.

Md. Shymon Islam and Kazi Masudul Alam. 2023. Sentiment analysis on bangla food reviews using machine learning and explainable nlp. In *2023 26th International Conference on Computer and Information Technology (ICCIT)*, pages 1–6.

Israa Jaradat, Pepa Gencheva, Alberto Barrón-Cedeño, Lluís Màrquez, and Preslav Nakov. 2018. ClaimRank: Detecting check-worthy claims in Arabic and English. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 26–30, New Orleans, Louisiana. Association for Computational Linguistics.

Mohsinul Kabir, Obayed Bin Mahfuz, Syed Rifat Raiyan, Hasan Mahmud, and Md Kamrul Hasan. 2023. Banglabook: A large-scale bangla dataset for sentiment analysis from book reviews. *Preprint*, arXiv:2305.06595.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., 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 EMNLP*.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha Talukdar. 2021. Muril: Multilingual representations for indian languages. *Preprint*, arXiv:2103.10730.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *Preprint*, arXiv:1711.05101.

Laura Majer and Jan Šnajder. 2024. Claim check-worthiness detection: How well do LLMs grasp annotation guidelines? In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pages 245–263, Miami, Florida, USA. Association for Computational Linguistics.

Shubham Mittal, Megha Sundriyal, and Preslav Nakov. 2023. Lost in translation, found in spans: Identifying claims in multilingual social media. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3887–3902, Singapore. Association for Computational Linguistics.

Preslav Nakov, Alberto Barrón-Cedeño, Tamer Elsayed, Reem Suwaileh, Lluís Màrquez, Wajdi Zaghouani, Pepa Atanasova, Spas Kyuchukov, and Giovanni Da San Martino. 2018. Overview of the clef-2018 checkthat! lab on automatic identification and verification of political claims. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 372–387, Cham. Springer International Publishing.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 9 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Pritam Pal, Dipankar Das, and Anup Kumar Kolya. 2025. Bilingual sentiment and emotion analysis: A multi-task learning framework for bengali and english. In *Proceedings of the 16th Annual Meeting of the Forum for Information Retrieval Evaluation*, FIRE '24, page 61–66, New York, NY, USA. Association for Computing Machinery.

Archita Pathak, Mohammad Abuzar Shaikh, and Rohini Srihari. 2020. Self-supervised claim identification for automated fact checking. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, pages 213–227, Indian Institute of Technology Patna, Patna, India. NLP Association of India (NLPAI).

Soham Poddar, Biswajit Paul, Moumita Basu, and Saptarshi Ghosh. 2024. ICPR 2024 Competition on Multilingual Claim-Span Identification (ICPR-CSI 2024).

Md. Rashadur Rahman, Rezaul Karim, Mohammad Shamsul Arefin, Pranab Kumar Dhar, Gahangir Hossain, and Tetsuya Shimamura. 2025. Facilitating automated fact-checking: a machine learning based weighted ensemble technique for claim detection. *Discover Applied Sciences*, 7(1):73.

Mohammad Rifat Ahmmad Rashid, Kazi Ferdous Hasan, Rakibul Hasan, Aritra Das, Mithila Sultana, and Mahamudul Hasan. 2024. A comprehensive dataset for sentiment and emotion classification from bangladesh e-commerce reviews. *Data in Brief*, 53:110052.

Sara Rosenthal and Kathleen McKeown. 2012. Detecting opinionated claims in online discussions. In *2012 IEEE Sixth International Conference on Semantic Computing*, pages 30–37.

Rudra Roy, Pritam Pal, Dipankar Das, Saptarshi Ghosh, and Biswajit Paul. 2025. Enhancing textual understanding: Automated claim span identification in english, hindi, bengali, and codemix. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI era*, pages 1030–1035, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Antony Samuels and John Mcgonical. 2020. News sentiment analysis. *Preprint*, arXiv:2007.02238.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Megha Sundriyal, Atharva Kulkarni, Vaibhav Pulastya, Md. Shad Akhtar, and Tanmoy Chakraborty. 2022. Empowering the fact-checkers! automatic identification of claim spans on Twitter. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7701–7715, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Megha Sundriyal, Parantak Singh, Md. Shad Akhtar, Shubhashis Sengupta, and Tanmoy Chakraborty. 2021. Desyr: Definition and syntactic representation based claim detection on the web. *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Dustin Wright and Isabelle Augenstein. 2020. Claim check-worthiness detection as positive unlabelled learning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 476–488, Online. Association for Computational Linguistics.

Amelie Wührl and Roman Klinger. 2021. Claim detection in biomedical Twitter posts. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 131–142, Online. Association for Computational Linguistics.

A Examples Used During Data Annotation by LLM

ANNOTATION EXAMPLES:
==========

"২০২৪-এ ভারতের জিডিপি ৭.৫ শতাংশ হারে বাড়বে বলে আশা আরবিআই-এর" --> [1, 'POSITIVE'];

"কলকাতায় আজ থেকে শুরু হচ্ছে আন্তর্জাতিক বইমেলা" --> [1, 'NEUTRAL'];

"নয়া দিল্লিতে তীব্র দূষণে স্কুল বন্ধ ঘোষণা" --> [1, 'NEGATIVE'];

"ভারত-পাকিস্তান ম্যাচে রোহিত শর্মার দুর্দান্ত সেঞ্চুরি" --> [1, 'POSITIVE'];

"আজকের রাজনীতি নীতিহীন ও দুর্নীতিপূর্ণ" --> [0, 'NEGATIVE'];

"কলকাতার ট্র্যাফিক এখন আগের চেয়ে অনেক নিয়ন্ত্রিত" --> [0, 'POSITIVE'];

"ভারতের সিনেমা বিশ্ব দরবারে সম্মান পাচ্ছে" --> [0, 'POSITIVE'];

"ভারতীয় সেনাবাহিনীর নতুন হেলিকপ্টার যুক্ত হলো বাহিনীতে" --> [1, 'POSITIVE'];

"বিজেপি-তৃণমূল সংঘর্ষে আহত ১০, ভাঙচুর ও অগ্নিসংযোগ" --> [1, 'NEGATIVE'];

"চেন্নাইয়ে ডেঙ্গু আক্রান্তের সংখ্যা বেড়েছে ৪০ শতাংশ" --> [1, 'NEGATIVE'];

"রানির স্টাইলে এবার মুগ্ধ নেটিজেনরা" --> [0, 'POSITIVE'];

"ধর্ষণের অভিযোগে পুলিশ কর্মী গ্রেপ্তার" --> [1, 'NEGATIVE'];

Figure A.1: Illustrative prompts used for claim checkworthiness detection and sentiment annotation in Bengali news headlines. The set comprises 12 examples spanning diverse domains, including politics, sports, entertainment, the economy, and social issues. These examples were provided to LLMs during the annotation phase to guide consistent labeling across both tasks.

B Statistical Distribution of Data

Label	Max	Min	Mean	Median	Mode	St. Dev.
Non-Checkworthy	37	6	10.192	10	10	2.802
Checkworthy	38	6	10.188	10	10	2.477
Negative	35	6	10.264	10	10	2.338
Neutral	37	6	10.032	10	10	2.558
Positive	38	6	10.298	10	10	2.921

Table B.1: Statistical distribution of the number of words across claim checkworthiness detection (Non-checkworthy, Checkworthy) and sentiment classification (Negative, Neutral, Positive) labels.

C STL Framework (Flow Diagram)

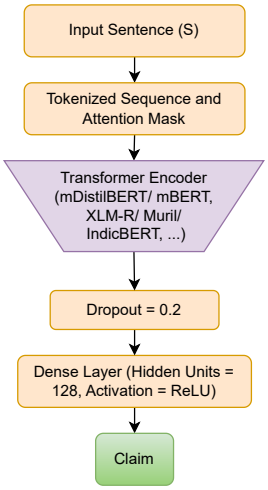

Figure C.1: Diagrammatic representation of the STL framework for identifying claim checkworthiness. The framework is identical to the MTL framework. However, instead of using two classification heads in the MTL framework, the STL framework employs a single classification head for each task.