Advancing Subjectivity Detection in Bengali News Articles Using Transformer Models with POS-Aware Features

Md Minhazul Kabir, Kawsar Ahmed
Mohammad Ashfak Habib and **Mohammed Moshiul Hoque**
Department of Computer Science and Engineering
Chittagong University of Engineering & Technology, Chattogram-4349, Bangladesh
{u1904040, u1804017}@student.cuet.ac.bd
{ashfak, moshiul_240}@cuet.ac.bd

Abstract

Distinguishing fact from opinion in text is a nuanced but essential task, particularly in news articles where subjectivity can influence interpretation and reception. Identifying whether content is subjective or objective is critical for sentiment analysis, media bias detection, and content moderation. However, progress in this area has been limited for low-resource languages such as Bengali due to a lack of benchmark datasets and tools. To address these constraints, this work presents **BeNSD** (Bengali News Subjectivity Detection), a novel dataset of 8,655 Bengali news article texts, along with an enhanced transformer-based architecture (*POS-Aware-MuRIL*) that integrates parts-of-speech (POS) features with MuRIL embeddings at the input level to provide richer contextual representation for subjectivity detection. A range of baseline models is evaluated, and the proposed architecture achieves a macro F1-score of 93.35% in subjectivity detection for the Bengali language. The code of the work is available on GitHub[1].

1 Introduction

The expansion of digital platforms and online journalism has resulted in a substantial increase in text-based content. News articles, blogs, and social media posts now serve as primary sources of information for a global audience. These media have a significant impact on public discourse and opinion formation. The proliferation of digital information has also introduced challenges, including the dissemination of biased, subjective, and manipulative content that complicates the consumption of accurate information and shapes public sentiment. Transitions within a single sentence from factual reporting to personal opinion can significantly alter the reader's interpretation and perception. Subjectivity detection in news articles refers to the process of distinguishing between factual statements and those that express opinions, emotions, or personal judgments. Subjectivity detection helps identify opinion-based statements, supporting balanced reporting and thereby enhancing the credibility of news sources.

Effective subjectivity detection supports news credibility, mitigates bias, and promotes the distribution of impartial information (Satapathy et al., 2022). Recent advancements in subjectivity detection have primarily been observed in high-resource languages, including English and other European languages. In contrast, subjectivity detection in Bengali remains understudied due to limited linguistic resources. As the volume of Bengali content on digital platforms grows, the need for effective subjectivity detection becomes increasingly important. However, progress is hindered by a scarcity of public datasets and the complexity of Bengalis syntax, morphology, and annotation ambiguity. This work contributes in the following ways to address current constraints in subjectivity detection in Bengali:

- We introduce **BeNSD**, a new annotated dataset comprising 8,655 Bengali news sentences, each labelled as either subjective (SUBJ) or objective (OBJ)
- We propose a transformer-based model that leverages MuRIL language model with POS embeddings. By integrating syntactic features with deep contextual representations, the model enhances its ability to recognize the subtle linguistic markers that distinguish subjective from objective expressions in Bengali texts.

2 Related Work

Subjectivity detection in news articles has seen marked progress in high-resource languages such

[1] https://github.com/R1FA7/Subjectivity-Detection-in-Bengali-News-Articles

Proceedings of the Second Workshop on Bangla Language Processing (BLP-2025), pages 131–141

as English. Paran et al. (2024) developed a model for subjectivity detection using datasets of 1,776 English and 2,675 Arabic news article sentences. Their approach achieved the highest F1 Scores of 72.6% on Arabic and 50.36% on English using Llama-3-8b, with the dataset size potentially contributing to the results. Antici et al. (2023) introduced annotation guidelines and a corpus of 1,049 annotated sentences for subjectivity detection in English news articles. Both monolingual and multilingual classification setups were examined. In the monolingual setting, m-BERT achieved macro-F1 scores of 75% for English and 74% for Italian. The multilingual use of m-BERT yielded 5% (for English) and 3% (for Italian) improvements in macro F1 scores. Pachov et al. (2023) applied a dataset of 1,019 English sentences, with a majority voting ensemble achieving the highest macro F1 score of 0.77. Frick (2023) studied subjectivity classification using LLM-augmented data with datasets of 1,292 English and 1,291 German sentences, employing BERT, GPT, and their combination. Their approach yielded an F1 value of 0.73, but performance was inconsistent when ChatGPT was used in few-shot settings. Dey et al. (2023) used a multilingual dataset in six languagesEnglish, Arabic, Dutch, German, Italian, and Turkishand applied transformer architectures, including BERT, M-BERT, and XLM-RoBERTa. The XLM-RoBERTa large model achieved the highest F1 score of 0.82 on a dataset comprising 7,828 texts. Additionally, a BERT-based multitask learning framework combining sentiment and subjectivity detection was introduced by Satapathy et al. (2022), with the addition of a Neural Tensor Network resulting in a 24% absolute improvement in accuracy, reaching 95.1%. On the IMDB dataset of 10,000 English movie reviews, A recent work Sagnika et al. (2021) proposed an opinion mining technique for subjectivity identification using an attention-based CNN-LSTM model, achieving an accuracy of 0.971.

In contrast to the progress made in high-resource languages, subjectivity detection in low-resource languages remains in a rudimentary stage. Suwaileh et al. (2025) proposed a dataset for subjectivity detection in Arabic news sentences, comprising 3,600 manually annotated sentences named ThatiAR. Their framework leveraged various transformers and LLMs, but using the 3-shot and 5-shot settings of GPT-4, they achieved the highest weighted F1 score of 0.80. Chaturvedi et al. (2017) proposed a framework using an extreme learning machine with Bayesian networks and fuzzy recurrent neural networks (FRNNs) for subjectivity detection and achieved an accuracy of 89%. A recent study (Dwivedi et al., 2024) explored subjectivity analysis in nine low-resource Indian languages using GPT-4 and BARD through in-context learning and prompt engineering. They showed that language-specific prompts significantly improve performance in multilingual settings. Around 7,000 domain-specific sentences were collected, and the data was balanced for subjective and objective classes by Dwivedi and Ghosh, 2022. They designed a lexical-rule-based Finite State Transducer (FST) for five Indian languages: Bengali, Hindi, Odia, Khorti, and Kannauji. Their system achieved an average accuracy of 84% across five languages.

Differences with existing research: Although notable progress has been made in subjectivity detection across high-resource languages, a significant gap remains in Bengali. In our exploration, no benchmark dataset is available for subjectivity detection in Bengali news articles, and no prior systems have been proposed for this task. Furthermore, previous methods have often overlooked the role of syntactic structures, such as POS, in shaping subjectivity. Instead, they have relied primarily on semantic features. To address these gaps, this work differs from existing studies in two crucial ways: (i) it introduces the first large-scale, manually annotated dataset, **BeNSD** for subjectivity detection in Bengali news articles, and (ii) it proposes a *POS-Aware* transformer-based model, combining syntactic and contextual cues, to improve subjectivity detection in Bengali.

3 Development of Dataset: BeNSD

This work develops **BeNSD**, a dataset for detecting subjectivity in news articles, as benchmark datasets are currently unavailable in Bengali. In news media, it is often hard to distinguish between subjective and objective text as the differences can be subtle. A *subjective text* expresses personal opinions, emotions, speculations, rhetorical questions, sarcasm, exaggerations, or unsupported conclusions. In contrast, the *objective text* presents information neutrally, reports events as they happen, includes third-party statements, and uses data-supported conclusions (Antici et al., 2021).

3.1 Data Collection and Preprocessing

Bengali news articles were collected through a combination of manual browsing (42 articles) and web scraping (251 articles) from prominent Bengali news portals. To ensure data diversity, sources were selected to represent a wide range of topics, including politics, health, sports, entertainment, and social issues. Initially, 8,800 raw news texts were gathered, with the majority (8,144) collected from Prothom Alo and the least amount (37) from Jugantor. The accumulated data was collected between January 21, 2025, and April 14, 2025. Figure 1 shows the source-wise distribution of collected data.

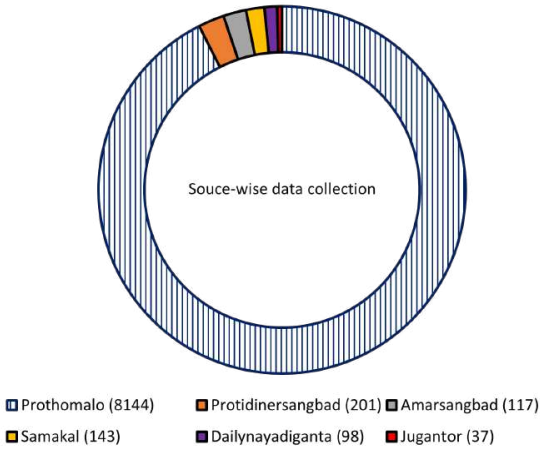

Figure 1: Source-wise accumulated data distribution. The values in the legends indicate the amount of data collected from each source.

To reduce manual annotation effort and redundancy, cleaning and filtering steps were applied: non-Bengali characters, extra punctuation, and special symbols were removed; duplicates were discarded; and articles with fewer than three meaningful words were filtered out. After preprocessing, 115 texts were removed, leaving 8,685 valid texts for manual annotation.

3.2 Data Annotation

After preprocessing, 8,685 news texts were given to annotators to label as *subjective (SUBJ)* or *objective (OBJ)*. Three undergraduate computer science students independently labelled each article, and their work was reviewed by an expert with over 20 years of experience in NLP research. Annotators followed clear class definitions to ensure consistent labelling, and all texts were annotated independently, without regard to prior context. Initially, they determined whether each sentence was subjective or objective by applying the provided guidelines. Before starting the main task, they practiced on a small set of example sentences to ensure they understood the difference between the two classes. This training helped clarify how they interpreted the definitions. Annotators also wrote brief justifications for their choices, which helped resolve disagreements during expert review. Since each text was labelled by three annotators, the final class label was decided through majority voting. After annotation, the expert removed 30 texts because their tone was ambiguous. The final dataset includes 8,655 annotated texts.

Raw Text	Processed Text	Remarks
বর্ণিত ম্যাক্স-মিন পদ্ধতি (Max-min principle), তথা সর্বাধিক দলের ন্যূনতম জরুরি সংস্কারে একমত হওয়া একটি বাস্তবসম্মত পথ তৈরি করতে পারে	বর্ণিত ম্যাক্স-মিন পদ্ধতি, তথা সর্বাধিক দলের ন্যূনতম জরুরি সংস্কারে একমত হওয়া একটি বাস্তবসম্মত পথ তৈরি করতে পারে	Removed non-Bengali characters
একটি ছেলে — যে এখানেই বড় হয়েছে... সবকিছু জিতেছে!	একটি ছেলে যে এখানেই বড় হয়েছে সবকিছু জিতেছে	Remove special symbol & punctutation
অনেকেই এসেছেন	Discarded	Text fewer than three words

Table 1: Preprocessing examples.

To evaluate annotation quality, we measured inter-annotator agreement using Cohens kappa coefficient (Cohen, 1960). The resulting kappa score was 0.92, indicating almost perfect agreement and demonstrating the reliability of the annotated dataset. The finalized dataset was then converted into a standardized format (e.g., Excel) for further processing.

3.3 Dataset Statistics

The **BeNSD** dataset is randomly split into training (80%), validation (10%), and test (10%) sets, with label stratification to maintain class balance for model development and evaluation. Table 2 provides a summary of the developed dataset. It is observed that OBJ samples have more words than SUBJ samples across all sets. On average, each text contains 13 to 14 words. The validation and test sets exhibit significantly higher lexical diversity (approximately 0.5) than the training set (approximately 0.24), indicating that they use a broader vocabulary despite being smaller.

We conducted several statistical analyses to better understand the dataset's characteristics. Figure 2 illustrates the distribution of SUBJ and OBJ text

Set	Class	Samples	W_T	W_U	W_{Avg}	W_U/W_T
Train	OBJ	4007	56,321	13,704	14.06	0.2433
	SUBJ	2917	38,629	9,568	13.24	0.2477
Subtotal		6924	94,950	18,859	-	-
Val	OBJ	501	6,895	3,539	13.76	0.5133
	SUBJ	364	4,900	2,445	13.46	0.4990
Subtotal		865	11,795	5,123	-	-
Test	OBJ	501	7,176	3,680	14.32	0.5128
	SUBJ	365	4,819	2,370	13.20	0.4918
Subtotal		866	11,995	5,173	-	-
Total		8,655	118,740	29,155	-	-

Table 2: Dataset statistics across training, validation, and test splits, where W_T, W_U, W_{Avg}, and W_U/W_T denote the total word count, unique word count, average words per sample, and lexical diversity ratio, respectively.

lengths in the dataset. The diagram shows that most OBJ sentences fall within the 7- to 15-word range, indicating a concentration of short news pieces. SUBJ sentences follow a similar pattern.

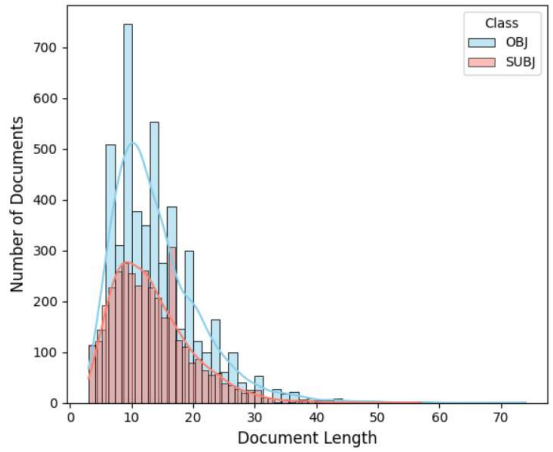

Figure 2: Distribution of sentence length (in words) for *SUBJ* and *OBJ* classes

To further explore the vocabulary and writing patterns, word clouds were generated separately for the subjective and objective classes. Figure 3 shows word clouds of the top 100 words from the dataset. Figure 3 revealed that the subjective class

Figure 3: Word clouds of the top 100 words from the dataset.

frequently includes words such as রাজনৈতিক (political), ভাল (good), and জরুরি (urgent), etc., which reflect opinions, emotions, or evaluations. In contrast, বাংলাদেশ (Bangladesh), কথা (talk), and সালে (in the year), etc., focus more on reporting events and presenting information in objective samples.

4 Methodology

This section describes our POS-aware transformer approach to subjectivity detection, integrating syntactic cues with contextual representations. We also outline the machine learning, deep learning, and transformer baselines evaluated for comparison.

4.1 Baselines

To evaluate the effectiveness of the proposed approach, we also implemented a range of ML, DL, and transformer-based baselines. All models are trained and assessed on the developed **BeNSD** dataset. Preprocessing, tokenization, and feature extraction techniques are applied uniformly across models where applicable.

- **ML Baselines:** This study examines several traditional ML models, including Logistic Regression (LR), Support Vector Machine (SVM), Naïve Bayes (NB), Random Forest (RF), XGBoost, and their ensembles. The text data is first converted to lowercase, and stopwords are removed. Next, the TF-IDF and CountVectorizer methods are used to convert the text into numerical features. These features are then used to train each model, with hyperparameters adjusted based on the validation results. Key hyperparameters include 100 estimators for RF, *logloss* evaluation metric for XGBoost, and a maximum of 1000 iterations for Logistic Regression.

- **DL Baselines:** To improve the quality of text representation, we explore DL models with pre-trained word embeddings. We experiment with four architectures: CNN, LSTM, BiLSTM, and their hybrid combinations. All models start with an embedding layer initialized with pre-trained Bengali embeddings from Word2Vec (Goldberg and Levy, 2014), GloVe (Pennington et al., 2014), or FastText (Bojanowski et al., 2016). We tune key hyperparameters, such as learning rate, batch size, and sequence length, using validation data, and apply dropout and early stopping to reduce overfitting. All models are

trained for up to 20 epochs, with early stopping saving the best model based on the validation loss. The LSTM model with 300-dimensional Word2Vec embeddings, a learning rate of 0.001, a batch size of 64, and the Adam optimizer performs best, achieving optimal results around 7 epochs. Table 3 illustrates the tuned hyperparameters for DL models. Hyperparameters are tuned using a mix of grid and random search across predefined ranges. We fine-tune key hyperparameters, such as sequence length, LSTM units, batch size, optimizer, and learning rate, based on validation performance. To reduce overfitting, we employ dropout and L2 regularization techniques.

Hyperparameter	Search Space	L	B	C + B
Vocabulary Size	[5000, 10000, 15000]	10000	10000	10000
Sequence Length	[64, 128, 256]	128	128	128
Embed. Dim	[100, 300]	300	300	300
LSTM Units	[64, 128, 256]	128	64	64
Dense Units	[32, 64, 128]	64	32	32
Batch Size	[32, 64, 128]	64	64	64
Learning Rate	[0.001, 0.01, 0.0001]	0.001	0.001	0.001
Optimizer	[adam, rmsprop]	adam	adam	adam
Conv Filters	[64, 128, 256]	–	–	64
Conv Kernel Size	[3, 5, 7]	–	–	3

Table 3: Hyperparameter summary of DL models. L, C, and B denote LSTM, CNN, and BiLSTM methods, respectively.

- **Transformer Baselines:** This work examines seven pre-trained transformer baselines and evaluates their performance on the dataset developed for subjectivity detection in Bengali news articles. The models are: Bangla-BERT-1 (Sarker, 2020), Bangla-BERT-2 (Bhattacharjee et al., 2021), Multilingual Bidirectional Encoder Representations from Transformers (m-BERT) (Devlin et al., 2018), distilled version of BERT (distilBERT) (Sanh et al., 2019) , cross-lingual version of robustly Optimized BERT (XLM-Roberta) (Conneau et al., 2020), Multilingual Representations for Indian Languages (MuRIL) (Khanuja et al., 2021) and indicBERT (Kakwani et al., 2020).

  All models are available in the Hugging Face Transformers library[2]. The m-BERT model has 12 layers, 12 attention heads, and 110 million parameters. We also evaluate distilbert-base-multilingual-cased, which has 6 layers and 768 hidden dimensions. MuRIL is trained on 17 Indian languages, including Bengali, using both monolingual and transliterated data. It provides improved contextual understanding for Indic-language tasks and is part of our evaluation. Bangla-BERT-1 is trained on the Bengali Common Crawl corpus using the base BERT architecture. We also include Bangla BERT-2, which is trained on a larger Bengali corpus and optimized for NLP tasks in Bengali. We chose XLM-RoBERTa for its strong multilingual performance. It is trained on 100 languages and has 12 transformer layers with 125 million parameters. IndicBERT is a lightweight model for Indic languages, including Bengali. We fine-tune all models on our dataset with various hyperparameter settings. We tune key hyperparameters, such as batch size, learning rate, weight decay, and the learning rate scheduler (linear with warmup), among others. Models are trained for up to 15 epochs with early stopping and the Adam optimizer. A learning rate of 1e-5 with weight decay of 0.01 is used. Evaluation and model checkpointing are performed every 300 steps based on the F1 score. Mixed-precision training (fp16) is enabled to accelerate training.

[2]https://huggingface.co

4.2 Proposed Architecture

Combining syntactic and lexical features with transformer models improves downstream task performance (Shi et al., 2022). Motivated by these findings, this work presents a POS-aware classification method that merges POS features with pre-trained contextual embeddings. The proposed model leverages Bengali POS tags and semantic information to enhance the identification of linguistic subjective patterns, providing a richer context for classification. Figure 4 illustrates the architecture of the proposed model.

4.2.1 Contextual and POS Embedding Extraction

For each input sentence, $S = [w_1, w_2, w_3, \ldots, w_m]$, the tokenizer produces a subword token sequence, $\text{TOK}(S) = [x_1, x_2, x_3, \ldots, x_n]$. Each token x_i is converted into an embedding vector, $\mathbf{E}^x = (\mathbf{e}_1^x, \mathbf{e}_2^x, \mathbf{e}_3^x, \ldots, \mathbf{e}_n^x)$, where $\mathbf{e}_i^x \in \mathbb{R}^{768}$ using the embedding layer. It combines token embedding $\mathbf{E}_{\text{token}}(x_i)$, positional embedding

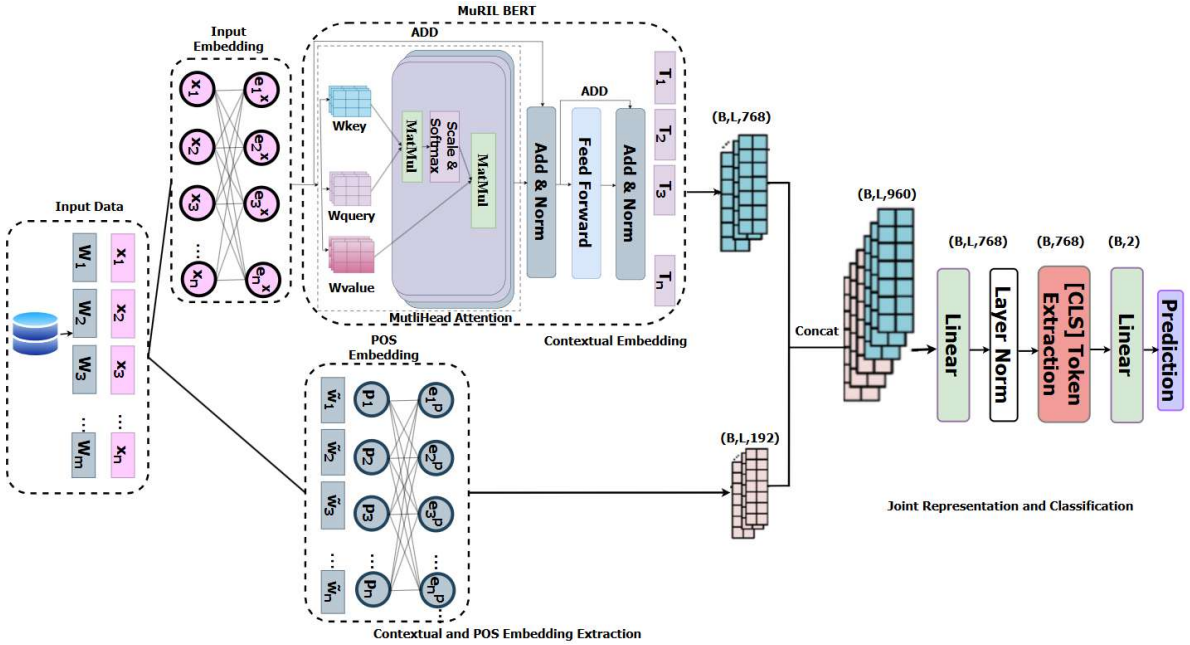

Figure 4: Architecture of the proposed model, where B: batch size, L: sequence length, x_i: subword tokens, $\tilde{w}_i$: original words mapped from tokens, and p_i: POS tag.

$\mathbf{E}_{\text{pos}}(i)$, and segment embedding $\mathbf{E}_{\text{segment}}(s)$. To obtain contextual embeddings, these input embeddings are fed into MuRIL's transformer layers as shown in Eq. 1.

$$\begin{aligned} T_i &= \text{MuRIL}(\mathbf{e}_i^x) \\ &= \text{MuRIL}\big(\mathbf{E}_{\text{token}}(x_i) + \mathbf{E}_{\text{pos}}(i) + \mathbf{E}_{\text{segment}}(s)\big) \end{aligned} \tag{1}$$

It produces contextual embedding, $\mathbf{T} = [\mathbf{T}_1, \mathbf{T}_2, \ldots, \mathbf{T}_n]$, $\mathbf{T}_i \in \mathbb{R}^{768}$ that captures semantic information.

For POS embedding, since tokenization may split a word into multiple subword tokens, we assign the POS tag of the original word to all its tokens. This ensures that the model receives the correct syntactic information aligned with its input tokens. For each token in the tokenized sequence, we track the original word it came from, denoted as $\tilde{w}_i$, and its POS tag, denoted as p_i. The BNLP library [3] is used for extracting POS tags. We map each POS tag p_i to an integer ID, denoted as $\text{id}(p_i)$, over a set of 14 predefined coarse-grained categories. These categories include core syntactic classes such as nouns, verbs, adjectives, adverbs, pronouns, prepositions, conjunctions, interjections, determiners, and punctuation, as well as special tokens like PAD, UNK, CLS, and SEP.

[3]https://github.com/sagorbrur/bnlp

We then create trainable POS embeddings of 64 dimensions as Eq. 2.

$$\mathbf{p}_i = \text{Embed}_{pos}[\text{id}(p_i)], \quad \mathbf{p}_i \in \mathbb{R}^{64} \tag{2}$$

To enrich the representation and improve the models ability to capture syntactic nuances, we project the POS embeddings into a higher-dimensional space, as shown in Eq. 3.

$$\tilde{\mathbf{p}}_i = \mathbf{W}_{proj}\mathbf{p}_i + \mathbf{b}_{proj}, \quad \tilde{\mathbf{p}}_i \in \mathbb{R}^{192} \tag{3}$$

where $\mathbf{W}_{proj} \in \mathbb{R}^{192\times64}$ is a trainable projection matrix and $\mathbf{b}_{proj} \in \mathbb{R}^{192}$ is a trainable bias vector. This projection enhances the performance of POS embeddings, making syntactic information more meaningful for the model.

4.2.2 Joint Representation and Classification

After projecting 64-dimensional POS embeddings to a 192-dimensional representation, for each token, we concatenate the contextual embedding $\mathbf{T}_i \in \mathbb{R}^{768}$ from MuRIL with the projected POS embedding $\tilde{\mathbf{p}}_i \in \mathbb{R}^{192}$ to form a fused representation as shown in Eq. 4.

$$\mathbf{z}_i = \begin{bmatrix} \mathbf{T}_i \\ \tilde{\mathbf{p}}_i \end{bmatrix}, \quad \mathbf{z}_i \in \mathbb{R}^{960} \tag{4}$$

This combined 960-dimensional representation, $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n]$, captures both the syntactic

and contextual aspects of each token. Each fused vector $\mathbf{z}_i$ is then passed through a feed-forward fusion layer with GELU activation function, which is chosen for its smooth non-linearity and effectiveness with small input values. After that, to stabilize the training and accelerate the convergence, we use layer normalization, which produces the transformed representation, $\mathbf{u}_i \in \mathbb{R}^{768}$ as defined in Eq. 5.

$$\mathbf{u}_i = \text{LayerNorm}\Big(\text{GELU}\big(\mathbf{W}_{\text{fusion}}\mathbf{z}_i + \mathbf{b}_{\text{fusion}}\big)\Big) \tag{5}$$

We train with a learning rate of 1×10^{-5}, weight decay of 0.01, and batch size of 16. A linear scheduler with a 0.1 warm-up ratio and early stopping with a patience of 10 epochs is applied. Finally, to perform classification, we extract the [CLS] token, which serves as the sentence-level representation. It is passed through a final dense layer to predict subjectivity. The whole prediction function can be expressed as Eq. 6.

$$\hat{y} = \arg\max\big(\text{softmax}(\mathbf{W}_{\text{cls}}\,\mathbf{u}_{[\text{CLS}]} + \mathbf{b}_{\text{cls}})\big) \tag{6}$$

5 Experiments

All experiments are conducted on the Kaggle platform using Python 3. Notebooks are executed in a GPU-enabled Kaggle environment with NVIDIA Tesla P100 GPUs and 16GB of RAM. For data manipulation, we use pandas (2.2.3) and numpy (1.26.4). Traditional machine learning models are implemented using scikit-learn (1.2.2). In contrast, deep learning models are built with Keras (3.8.0) and TensorFlow (2.18.0). For transformer-based models, we utilize PyTorch (version 2.6.0) and the Hugging Face Transformers library.

5.1 Results and Discussions

While the macro F1-score (F1) is used to assess model performance, other standard metrics such as accuracy (A), precision (P), and recall (R) are also reported. Table 4 shows how different baseline models performed on subjectivity detection in Bengali on **BeNSD** dataset.

Performance of ML and DL Models: Traditional ML approaches are less effective for Bengali subjectivity detection. SVM with TF-IDF achieves the highest mF1 among ML models (73.06%). In contrast, DL models improve performance: the best, LSTM with Word2Vec, achieves an F1 of 83.49%an approximate 10-point gain. This highlights the importance of distributed representations and sequential modeling in detecting subjectivity in Bengali text.

Superior performance of Transformer models: Transformer-based models consistently outperform both traditional ML and DL approaches. It shows performance ranging from 80.13% (IndicBERT) to 92.09% (MuRIL). Notably, multilingual models like mBERT (87.20%) and XLM-R (90.33%), as well as Bengali-specific models such as Bangla-BERT variants (88.05-91.64%), demonstrate superior performance. These results suggest that both multilingual knowledge and language-specific contextual information play a crucial role in enhancing the detection of subjectivity in Bengali texts..

Impact of POS Integration Varies by Model Architecture: The integration of POS embedding shows varying degrees of improvement across different transformer models. Lower-performing models, such as IndicBERT, gain substantially from POS integration (+2.40 F1 scores), while some mid-tier models, like mDistilBERT, show modest improvements (+0.64 F1 scores). However, the relationship is not uniform across all high-performing models. Bangla-BERT-1 shows minimal gain (+0.01 F1), but other strong performers achieve more significant benefits.

POS-Aware Technique Enhances Model's Performance: The proposed POS-Aware-MuRIL achieves the highest performance with a F1 score of 93.35%. This represents a notable 1.26-point improvement over the base MuRIL. The improvement results from both the integration of POS embedding with the transformer's contextual embedding and an effective architecture. This architecture employs GELU activation and layer normalization to process combined features. These techniques enable the model to capture both contextual and syntactic information, thereby helping identify features indicative of subjectivity in Bengali news articles. Moreover, consistent improvements in recall across all POS-enhanced models (ranging from +0.47 to +1.90) suggest that syntactic features help identify subtle subjective patterns that purely contextual models might miss. Appendix A presents an ablation study to analyze the impact of POS embedding on performance.

Impact of Fusion Strategy: We evaluate five fusion mechanisms for integrating POS embeddings with MuRIL BERT representations. Table 5 illustrates the results of fusion techniques. Among

ML Models				
Classifier	**A (%)**	**P(%)**	**R(%)**	**F1(%)**
XGBoost (CountVec)	71.59	70.17	56.71	62.73
RF (TF-IDF)	72.63	67.68	67.12	67.40
RF (CountVec)	73.21	67.92	69.04	68.48
NB (TF-IDF)	76.10	78.42	59.73	67.81
LR (TF-IDF)	76.21	76.07	63.56	69.25
SVM (CountVec)	75.17	70.27	71.23	70.75
NB (CountVec)	75.52	71.19	70.41	70.80
LR (CountVec)	76.21	73.18	68.77	70.90
Ensemble (CountVec)	76.91	74.92	67.95	71.26
SVM (TF-IDF)	77.60	74.08	72.05	**73.06**
DL Models				
Classifier	**A(%)**	**P(%)**	**R(%)**	**F1(%)**
BiLSTM (GloVe)	79.45	70.19	79.18	78.50
CNN+BiLSTM (GloVe)	81.87	75.62	84.11	79.64
CNN+BiLSTM (FastText)	81.76	74.01	87.40	80.15
LSTM (GloVe)	82.79	76.34	85.75	80.77
LSTM (FastText)	84.64	81.52	82.19	81.86
BiLSTM (FastText)	84.99	82.19	82.20	82.19
BiLSTM (Word2Vec)	84.76	79.80	85.48	82.54
CNN+BiLSTM (Word2Vec)	84.87	80.31	84.93	82.56
LSTM (Word2Vec)	85.33	79.46	87.95	**83.49**
Transformers				
Classifier	**A(%)**	**P(%)**	**R(%)**	**F1(%)**
IndicBERT	80.48	79.99	80.38	80.13
+POS	82.79	82.35	82.90	82.53
Δ	+2.31	+2.36	+2.52	+2.40
mDistilBERT	84.53	84.22	83.95	84.07
+POS	85.22	85.10	84.44	84.71
Δ	+0.69	+0.88	+0.49	+0.64
mBERT	87.53	87.60	87.04	87.20
+POS	87.99	87.69	87.68	87.69
Δ	+0.46	+0.09	+0.64	+0.49
Bangla-BERT-1	88.22	87.82	88.12	88.05
+POS	88.22	87.85	88.59	88.06
Δ	+0.00	+0.05	+0.47	+0.01
XLM-Roberta	90.65	90.75	90.02	90.33
+POS	91.57	91.49	91.19	91.33
Δ	+0.92	+0.74	+1.17	+1.00
Bangla-BERT-2	91.80	92.02	92.03	91.64
+POS	92.38	92.03	92.52	92.23
Δ	+0.58	+0.36	+0.49	+0.59
MuRIL	92.38	92.79	91.63	92.09
+POS-Aware (Proposed)	93.42	93.10	93.53	**93.35**
Δ	+1.04	+0.31	+1.90	+1.26

Table 4: Performance of various models on the subjectivity detection task

them, concatenation achieves the highest macro F1 score of 93.35%, followed by gated fusion at 92.49%. The proposed concatenation-based fusion outperforms the gated fusion by 0.86% and the additive fusion by 0.91%. All fusion strategies demonstrate strong performance; however, the simple concatenation-based fusion shows the most robust overall performance when integrating both embeddings. Element-wise strategy, such as additive and multiplicative fusion, outperforms attention-based fusion. The strong performance of the concatenation-based fusion in our POS-Aware MuRIL model shows that keeping features separate helps the model use linguistic information more effectively.

Fusion Type	**Ac (%)**	**Pr (%)**	**Re (%)**	**F1 (%)**
Attention	92.26	92.82	91.42	91.95
Multiplicative	92.61	92.48	92.35	92.41
Additive	92.61	92.35	92.53	92.44
Gated	92.73	92.83	92.23	92.49
Concatenation	93.42	93.10	93.53	**93.35**

Table 5: Performance comparison of different fusion strategies

5.2 Comparison with Existing Approaches

To the best of our knowledge, no publicly available dataset exists for subjectivity detection in Bengali news articles. To facilitate comparison in this context, we implement and adapt several existing techniques from similar domains (Antici et al., 2023; Sagnika et al., 2021; Paran et al., 2024; Dey et al., 2023) to the **BeNSD** dataset, ensuring consistency across these approaches. Table 6 presents a comparative analysis of F1-scores. Notably, the proposed *POS-Aware MuRIL* model achieves the highest F1 score of 93.35%, improving by 9.28% over LLaMA (Paran et al., 2024) and 2.02% over the XLM-R (Dey et al., 2023).

Approach	**mF1 (%)**
CNN-LSTM+Attention (Sagnika et al., 2021)	77.65
LLaMA-3-8B (Paran et al., 2024)	84.07
mBERT (Antici et al., 2023)	88.65
XLM-R (Dey et al., 2023)	91.33
POS-Aware MuRIL (Proposed)	**93.35**

Table 6: Comparison of existing approaches with the proposed method for Bengali subjectivity detection.

5.3 Error Analysis

The results confirmed that the proposed model detects subjectivity in Bengali news articles more effectively than the baselines. To better understand how the model performs, we carried out a detailed error analysis using both quantitative and qualitative methods.

Quantitative Error Analysis: Figure 5 shows the confusion matrix for the proposed *POS-Aware MuRIL* model on test sets. Of the 866 samples, 810 were correctly classified and 56 were misclassified. Among 365 subjective samples, 31 were misclassified, and among 501 objective samples, 25 were misclassified. This suggests the model misclassified both SUBJ and OBJ instances at nearly the same rate. The errors are not strongly biased toward one class but reflect challenges in capturing

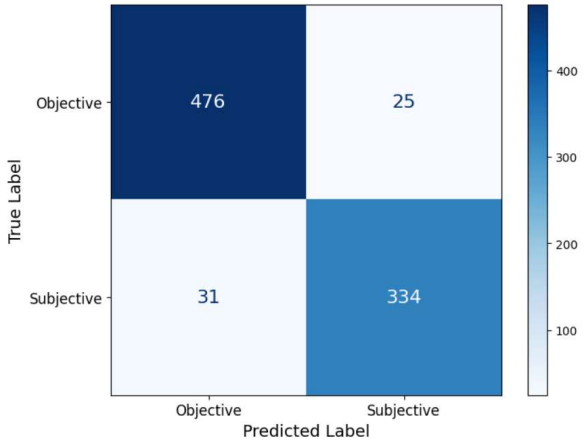

Figure 5: Confusion matrix for the proposed model on test data

nuanced distinctions. A primary reason is lexical ambiguity, in which certain words or phrases appear in both subject and object contexts depending on usage.

Figure 6 shows the 2D t-SNE plot, which indicates a clear separation between classes, with errors primarily located at the edges of the clusters, as highlighted by the rectangle. This suggests the

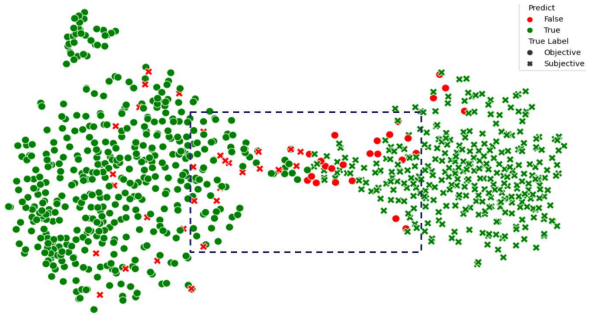

Figure 6: 2D t-SNE plot of *POS-Aware-MuRIL* embedding on test data.

model mainly struggles with ambiguous cases that fall between the two classes. These insights could help improve classification by adding more varied linguistic cues during training and refining feature representation.

Qualitative Error Analysis: Table 7 shows some sample texts that are correctly and incorrectly classified by the *POS-Aware MuRIL* model. The first and third samples were correctly classified, whereas the model failed to classify the second and fourth samples.

The analysis of incorrect predictions reveals that it is challenging to identify texts with hidden subjectivity. These cases are tough to categorize because they do not clearly show subjective meaning. Sarcasm, a critical tone, and rhetorical questions, for example, often lead to misclassification. Additionally, the model incorrectly labels factual descriptions as subjective when they contain descriptive adjectives. Without context, it is challenging to classify these samples solely based on the text. Subjectivity typically relies on broader cues that sentence-level models often overlook. Using more context and a wider range of training data could help improve the models performance.

Text	Ac	Pr
বাংলাদেশের স্বার্থেই রোহিঙ্গা সংকটের সমাধান বা বর্তমানের চেয়ে ভালো বিকল্প বের করা উচিত (*For Bangladeshs own interest, the Rohingya crisis should be resolved or a better alternative than the current one should be found.*)	SUBJ	SUBJ
যেন বলা হলো—সন্ধ্যার পর ছাত্রীরা হল থেকে বের হয় বলেই এমন ঘটনা ঘটে (*As if to say, such incidents happen because female students leave the hall after dark.*)	SUBJ	OBJ
অল্টম্যানের পক্ষ থেকে তাদের চীনা প্রতিদ্বন্দ্বী ডিপসিকের প্রশংসা করে বলা হয়, এটা ভালো একটি মডেল (*On behalf of Altman, their Chinese rival DeepSeek was praised, saying its a good model.*)	OBJ	OBJ
ঐতিহ্যবাহী এই উৎসব দেখতে উৎসুক জনতার কোনো কমতি ছিল না (*There was no shortage of eager crowds to watch this traditional festival.*)	OBJ	SUBJ

Table 7: Some correctly and incorrectly classified samples by *POS-Aware MuRIL* model

6 Conclusion

This work presented **BeNSD**, a manually annotated dataset for subjectivity detection in Bengali, containing 8,655 news articles from multiple online sources. Alongside, we introduce a transformer-based model (**POS-Aware-MuRIL**) that fuses MuRIL's contextual embeddings with POS embeddings to boost classification accuracy. Evaluation shows the proposed model surpasses machine learning, deep learning, and transformer baselines, achieving the top macro F1 score (93.35%). Building on these results, we plan to expand the dataset, refine syntactic feature extraction through improved POS tagging, and investigate ensemble and multitask learning to further enhance system robustness and generalization.

Limitations

Although the proposed approach shows promising results, some limitations remain. First, because the POS tagger is not fine-tuned for this dataset, some syntactic information may be misidentified, leading to inaccurate feature extraction and misclassifications in the text. Additionally, frequent words that appear in both subjective and objective texts create class ambiguity, making it difficult for the model to distinguish between categories and reducing its ability to generalize. Moreover, context-specific language and sarcasm can cause the model to misinterpret meaning, further limiting classification accuracy. Addressing these issues by fine-tuning linguistic tools and expanding the dataset is crucial to improving performance. Furthermore, integrating large language models (LLMs) in future work could enhance the results.

Acknowledgements

This work is supported by the Directorate of Research and Extension (DRE) and NLP Lab, Chittagong University of Engineering & Technology (CUET), Chittagong 4349, Bangladesh.

Ethics Statement

The *BeNSD* dataset was created by a team that included a native Bengali-speaking NLP expert and three undergraduate students with NLP backgrounds. The team employed clear annotation guidelines that outlined data sources, collection steps, and task formats to ensure data consistency and accuracy. They also manually checked and corrected the data to ensure its accuracy. An NLP expert with more than 20 years of experience reviewed the final annotations to confirm their reliability. Throughout the process, the team adhered to ethical guidelines to ensure the data was fair and trustworthy.

References

Francesco Antici, Luca Bolognini, Matteo Antonio Inajetovic, Bogdan Ivasiuk, Andrea Galassi, and Federico Ruggeri. 2021. Subjectivita: An italian corpus for subjectivity detection in newspapers. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 12th International Conference of the CLEF Association, CLEF 2021, Virtual Event, September 21–24, 2021, Proceedings 12*, pages 40–52. Springer.

Francesco Antici, Andrea Galassi, Federico Ruggeri, Katerina Korre, Arianna Muti, Alessandra Bardi, Alice Fedotova, and Alberto Barrón-Cedeño. 2023. A corpus for sentence-level subjectivity detection on english news articles. *arXiv preprint arXiv:2305.18034*.

Abhik Bhattacharjee, Tahmid Hasan, Kazi Samin, M Sohel Rahman, Anindya Iqbal, and Rifat Shahriyar. 2021. Banglabert: Combating embedding barrier for low-resource language understanding. *arXiv preprint arXiv:2101.00204*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information.

Iti Chaturvedi, Edoardo Ragusa, Paolo Gastaldo, Rodolfo Zunino, and Erik Cambria. 2017. Bayesian network based extreme learning machine for subjectivity detection. *Journal of the Franklin Institute*, 355.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

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*.

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

K. Dey, P. Tarannum, M. A. Hasan, and S. R. H. Noori. 2023. Nn at checkthat!-2023: Subjectivity in news articles classification with transformer based models. In *Working Notes of CLEF 2023 Conference and Labs of the Evaluation Forum*, volume 3497 of *CEUR Workshop Proceedings*.

Satyam Dwivedi and Sanjukta Ghosh. 2022. Subjectivity identification through lexical rules. *SN Computer Science*, 3(1):32.

Satyam Dwivedi, Sanjukta Ghosh, and Shivam Dwivedi. 2024. Navigating linguistic diversity: In-context learning and prompt engineering for subjectivity analysis in low-resource languages. *SN Comput. Sci.*, 5(4).

Raphael Antonius Frick. 2023. Fraunhofer sit at checkthat!-2023: Can llms be used for data augmentation & few-shot classification? detecting subjectivity in text using chatgpt. In *CLEF (Working Notes)*, pages 329–336.

Yoav Goldberg and Omer Levy. 2014. word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method.

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, Sumanth Doddapaneni, Sunayana Sitaram, Monojit Choudhury, and Kalika Bali. 2021. Muril: Multilingual representations for indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 607–622.

Georgi Pachov, Dimitar Dimitrov, Ivan Koychev, and Preslav Nakov. 2023. Gpachov at checkthat! 2023: A diverse multi-approach ensemble for subjectivity detection in news articles. *Preprint*, arXiv:2309.06844.

Ashraful Paran, Md Hossain, Symom Shohan, Jawad Hossain, Shawly Ahsan, and Moshiul Hoque. 2024. Semanticcuetsync at checkthat! 2024: Finding subjectivity in news articles using llama notebook for the checkthat! lab at clef 2024.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Santwana Sagnika, B. Mishra, and Saroj Meher. 2021. An attention-based cnn-lstm model for subjectivity detection in opinion-mining. *Neural Computing and Applications*, 33:1–14.

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

Sagor Sarker. 2020. Banglabert: Bengali mask language model for bengali language understading.

Ranjan Satapathy, Shweta Pardeshi, and Erik Cambria. 2022. Polarity and subjectivity detection with multitask learning and bert embedding. *Preprint*, arXiv:2201.05363.

Yu Shi, Xi Zhang, and Ning Yu. 2022. Pl-transformer: a pos-aware and layer ensemble transformer for text classification. *Neural Computing and Applications*, 35.

R. Suwaileh, M. Hasanain, F. Hubail, W. Zaghouani, and F. Alam. 2025. Thatiar: Subjectivity detection in arabic news sentences. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, pages 2587–2602.

A Ablation Study

To identify the best configuration for tasks, we conduct an ablation analysis of key factors, including the impact of POS embedding dimensions for integrating POS embeddings with MuRIL BERT. As POS embeddings play a pivotal role in performance, we vary the POS embedding dimension from 16 to 512 to determine the optimal size for capturing syntactic information. Figure 7 shows that dimensions between 32 and 128 yield stable performance, with 64 achieving the best F1 score of 0.9335. Smaller dimensions fail to capture a sufficient number of syntactic patterns. On the other hand, larger dimensions add unnecessary parameters without improving performance. This indicates that moderate embedding sizes offer a better trade-off between expressiveness and efficiency.

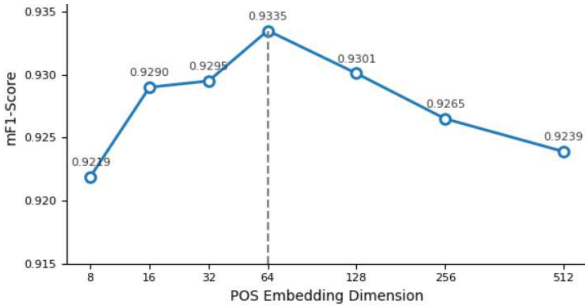

Figure 7: Impact of POS embedding dimension on model performance