Ushoshi2023 at BLP-2023 Task 2: A Comparison of Traditional to Advanced Linguistic Models to Analyze Sentiment in Bangla Texts

Sharun Akter Khushbu
Daffodil International University
sharun.cse@diu.edu.bd

Nasheen Nur
Florida Institute of Technology
nurn@fit.edu

Mohiuddin Ahmed
University of North Carolina at Charlotte
mahmed27@uncc.edu

Nashtarin Nur
United International University
nashtarin.nur@gmail.com

Abstract

This article describes our analytical approach designed for BLP Workshop-2023 Task-2: in Sentiment Analysis. During actual task submission, we used DistilBERT. However, we later applied rigorous hyperparameter tuning and preprocessing, improving the result to 68% accuracy and a 68% F1 micro score with vanilla LSTM. Traditional machine learning models were applied to compare the result where 75% accuracy was achieved with traditional SVM. Our contributions are a) data augmentation using the oversampling method to remove data imbalance and b) attention masking for data encoding with masked language modeling to capture representations of language semantics effectively, by further demonstrating it with explainable AI. Originally, our system scored 0.26 micro-F1 in the competition and ranked 30th among the participants for a basic DistilBERT model, which we later improved to 0.68 and 0.65 with LSTM and XLM-RoBERTa-base models, respectively.

1 Introduction

Sentiment analysis and opinion-mining techniques determine a text's sentiment or emotional polarity and then analyze it (Medhat et al., 2014). Throughout diverse fields, such as marketing, customer feedback analysis, and social media monitoring, sentiment analysis has gained significant attention in recent years. While sentiment analysis has been extensively studied in languages like English, there is a growing interest in applying this technique to other languages, including Bangla. Analyzing sentiment in Bangla text presents unique challenges due to its complex grammar, script, and nuances. This article aims to explore sentiment analysis in the Bangla language with an example dataset provided for the BLP workshop competition for task 2 using sequential data analysis models, such as LSTM and large language models, along with traditional models. This multi-class classification task determines whether the sentiment expressed in the text is positive, negative, or neutral.

Even though LSTM provides the highest performance among the deep learning models, XLM-RoBERTa-base (Singh et al., 2022) uses Masked Language Modeling (MLM) to handle multilingual and cross-lingual tasks, making it a powerful tool for understanding and generating text in multiple languages. MLM is a pre-training objective used in models like XLM-RoBERTa-base. Using MLM, a fraction of input tokens are replaced with unique [MASK] tokens, and the model is trained to predict the original tokens from the context provided by the surrounding tokens. MLM is a self-supervised learning task where a model learns to understand the statistical properties of the language by making predictions. We provide the competition results on the GitHub[1] which was implemented with DistillBERT. The final implementation with the higher accuracy and comparative analysis on different models is available in the GitHub[2].

Our rigorous experiments on a dataset and with various models have resulted in the following observations in addition to designing the system.

- Observation 1: Classifiers with no boosting, oversampling, or undersampling gave lower recall with a lower false positive rate (FPR). Without techniques like boosting, oversampling, or undersampling, a classifier tends to be biased toward the majority class. For example, after applying these techniques and masking, we get 66% accuracy for the XLM-RoBERTa-base, which was previously 41.45% on the XLM-RoBERTa-base.The classifier is conservative when clas-

[1] https://github.com/blp-workshop/blp_task2#leaderboard

[2] https://github.com/sharunakter/BLPWorkshop_2023_SentimentAnalysisInBangla

Proceedings of the First Workshop on Bangla Language Processing (BLP-2023), pages 293–299
December 7, 2023 ©2023 Association for Computational Linguistics

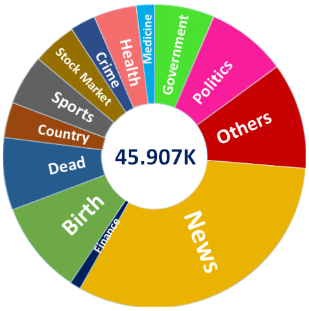


Figure 1: Data Distribution for Different Categories

sifying instances as the minority class. It generates too many false positive predictions (i.e., predicting the minority class when it is the majority class), which keeps the FPR low. Moreover, oversampling with boosting combats the data skew for all the models.

- Observation 2: XAI on XLM-RoBERTa-base's output shows how the MLM approach captures the nuanced sentiment expressed in Bangla text, even in the presence of code-mixing, sarcasm, or subtle linguistic cues. By understanding sentiment polarity and the context in which sentiments are expressed, it is possible to gain a deeper understanding. The randomly masked tokens were replaced with the special [MASK] token, creating partially masked sequences. The partially masked sequences were fed into the pre-trained XLM-RoBERTa-base model, which has been fine-tuned for sentiment analysis and language understanding tasks.

2 Background

2.1 Dataset Description

The dataset contains tweets or news-related public comments (Hasan et al., 2023a) to identify multi-class classification. Bangla data on various topics, such as political issues, incidents, COVID-19 facts, and country news from various online sources, are manually collected. The distribution of three classification labels, "negative," "positive," and "neutral," for training, dev, and test datasets are "19612", "17090", and "9205" respectively with a total datapoints of 45907.

2.2 Related Work

Interpreting implicit and underspecified phrases in instructional texts is vital to elicit plausible clarification and understanding (Roth et al., 2022; Islam et al., 2021). Researchers are increasingly focusing on sentiment analysis for low-resource languages like Bangla using traditional supervised machine learning such as multinomial Naïve Bayes (Sharif et al., 2019), SVM, Random forest and decision tree, and deep learning approaches such as deep recurrent neural network (Hassan et al., 2016), Glove word embedding with convolutional neural network (Mahmud et al., 2022), transfer learning using multilingual BERT (M-BERT) (Islam et al., 2020), transformer-based approach (Bhowmick and Jana, 2021; Hasan et al., 2023b). The lack of sufficient labeled data and domain and gender agnostic data limit the performance of those approaches (Islam et al., 2023a). Considering the scarcity of annotated data and the problem of predicting the lexical complexity of single-word and multi-word expressions, (Taya et al., 2021) used an ensemble model over a set of transformer-based model with hand-crafted features to increase the model generalization and robustness. To improve the quality of the sentiment analysis task of low-resource languages such as Bangla, the authors (Rahman and Kumar Dey, 2018; Sultana et al., 2022) proposed aspect-based sentiment analysis using BOW and supervised machine learning techniques and provided two datasets for aspect-based sentiment analysis. Many researchers claimed that transfer learning with adaptive pre-training effectively improves sentiment prediction tasks in low-resource languages by selecting appropriate source languages (Wang et al., 2023). Candidate source language selection through forward and backward strategies will increase compute requirements. To discover the effectiveness of semantic and syntactic parsing and the effects of subjective aspects on sentiment analysis, the authors at (Morio et al., 2022) performed a graph-based and seq2seq-based analysis with the help of a pre-trained language model and discovered that both research approaches perform well in extracting structured sentiment.

Considering the challenges for the Bangla dataset, the sentiGold (Islam et al., 2023a) developed a comprehensive dataset for sentiment analysis and provides a word embedding method, BanglaBERT, which performs well on formal Bangla text. However, the performance degrades for controversial text because of the need to be trained on informal data.

3 System Overview: Experiment and Setup

This section describes our data preprocessing steps for traditional machine learning models, vanilla deep learning models, and transformers. Next, we discuss the training and hyperparameter tuning of each model group.

3.1 Preprocessing and Data Augmentation

Bangla sentiment annotation is a challenging task because of its diversified syntaxes. Our task is to detect sentiment with three polarities: positive, negative, and neutral. We filtered out duplicate text if structural and semantic similarity were high (Islam et al., 2023b). Several syntaxes have been removed from the text, including punctuation marks, links, emoji, hashtag signs, and usernames (Mukta et al., 2021). We removed all non-Bangla characters and stop words and implemented Porterstemmer (Budiasih et al., 2009) to identify the root words and suffixes. Following preprocessing, boosting is applied with oversampling. There is a lack of balance in the class distribution of the Bangla dataset provided. Therefore, to balance the class distribution, we used oversampling techniques (Tahir et al., 2023) on the dataset. We merged the train and dev-test set to train the model. We applied the upsampling technique to the combined dataset with a ratio of 1.0 for the negative class.

3.2 Training and Hyperparameter Tuning

We used an 80-20 training-validation split for training all the classifiers: complex deep learning models, pre-trained transformers, and traditional machine learning algorithms.

Deep Learning and MLM: We experimented with following vanilla deep learning models: LSTM (Bhowmik et al., 2022), LSTM CNN (Chowdhury et al., 2022) and pretrained transformer models such as multilingual-BERT (M-BERT) (Tarannum et al., 2022), XLM-RoBERTa-base (Singh et al., 2022), DistilBERT (Suri, 2022; Fröbe et al., 2023), BanglaBERT (Sarker, 2020).

After the first round of analysis, we continue with both multilingual models BERT and XLM-RoBERTa-base and train our datasets with rigorous hyperparameter tuning and masked language modeling. The number of parameters and network size are responsible for the computation time and performance of the learning.

The number of labels determines the size of the last fully-connected dense layer. To predict the likelihood of the label, softmax activation with sparse categorical cross-entropy is applied on top of the model. The total parameter size for XLM-RoBERTa-base was 278045955, which took approximately 2 hours to complete the training on 8 GB RAM. We use a transformer toolkit for transfer learning in Bangla language (Hasan et al., 2019). The hyperparameters for hidden and feed-forward sizes are 768 and 3072, with 12 heads and 12 transformer blocks, regularized by a dropout of 10%, and the vocabulary size is 250002. XLM-RoBERTa-base model and other transformer models were fine-tuned with a batch size ranging from [16, 32], learning rate (Adam) range [3e-5, 2e-5], and number of epochs is 3. Tokenizer tools in the Huggingface (Zhang et al., 2019) repository were used to tokenize and preprocess the dataset.

For LSTM training, the parameters are maximum features = 500, embedding_dimension = 128, input length = 300, vocabulary size = 5000, and learning rate 0.01 with a decay value 1e-6. For 3 class labels, the batch size is 64, and the epoch number is 50. Additionally, there is one dense layer for sequential learning, 2 units of 1D MaxPooling layers, and a dropout of 0.2. Relu and Softmax were used for embedding. We used Adam optimization and sparse categorical cross-entropy as loss function. Table 1 reports the output for evaluation metrics and individual class labels on the test dataset for all deep learning models.

Traditional Machine Learning Models: We experimented with traditional approaches such as (I) Linear Regression (LR), (ii) Decision Tree (DT), (iii) Random Forest (RF), (iv) Multinomial Naïve Bayes (MNB), (v) K-Nearest Neighbour (KNN), (vi) Support Vector Machine (SVM) (Sazzed, 2021) and (vii) Stochastic Gradient Descent (SGD). We first transformed the preprocessed data into TF-IDF vectors with weighted n-gram (unigram, bigram, and trigram) to use contextual information. Table 2 reports the output for the traditional machine learning models.

4 Evaluations and Discussion on Results

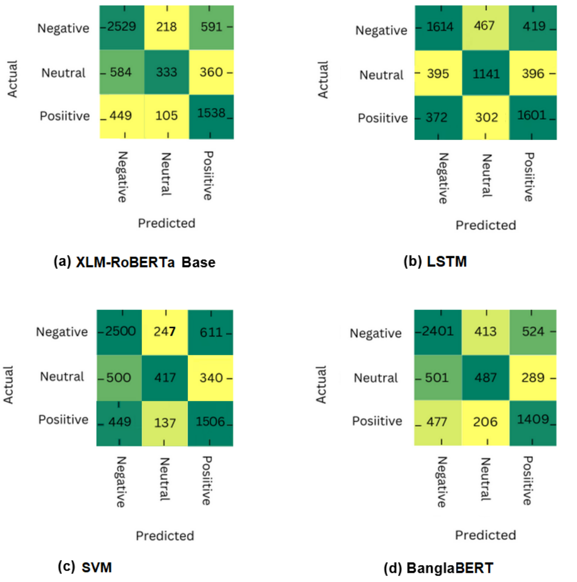
In the original competition, we generated the results using a basic DistillBERT model without any preprocessing and fine-tuning. DistillBERT can process maximum 10k data - even batch-

Table 1: Evaluation of Top Deep Learning Models based on Individual Class Labels

Class Label	Model	Accuracy	Precision	Recall	F1	Micro F1	Macro F1
Negative	LSTM	0.68	0.70	0.64	0.67	0.68	0.62
Neutral			0.70	0.78	0.74		
Positive			0.63	0.63	0.63		
Negative	XLM-RoBERTa-base	0.66	0.71	0.76	0.73	0.65	0.58
Neutral			0.51	0.26	0.34		
Positive			0.62	0.74	0.67		
Negative	BanglaBERT	0.64	0.71	0.72	0.71	0.64	0.59
Neutral			0.44	0.38	0.41		
Positive			0.63	0.67	0.65		
Negative	Multilingual BERT	0.64	0.68	0.77	0.72	0.64	0.57
Neutral			0.46	0.29	0.36		
Positive			0.65	0.66	0.66		
Negative	DistilBERT	0.55	0.54	0.54	0.54	0.55	0.51
Neutral			0.60	0.64	0.61		
Positive			0.20	0.33	0.24		

Table 2: Evaluation Metrics: Traditional ML Models

Traditional Models	Accuracy	Precision	Recall	F1-Score
LR	71.91	72.54	71.91	71.52
DT	64.81	64.31	64.81	64.18
RF	72.66	73.55	72.66	72.00
MNB	71.22	72.51	71.22	70.83
KNN	53.69	54.79	53.69	53.64
SVM	75.02	75.26	75.02	74.85
SGD	60.40	65.94	60.40	58.69



(a) XLM-RoBERTa Base

(b) LSTM

(c) SVM

(d) BanglaBERT

Figure 2: Confusion Matrix for Deep Learning Models

wise processing and averaging the output scores couldn't give a good result. We improved with a rigorous comparative analysis with vanilla deep learning, transformer-based LLMs, and traditional machine learning models that can handle large datasets. SVM achieved the highest accuracy and F1-score of 75.02% and 74.85% (Table 2). Unlike transformer-based models, LSTM and traditional models require extensive preprocessing, data cleaning, and oversampling. Moreover, upsampling and boosting improves all of the models. For example, before oversampling and boosting, XLM-RoBERTa-base reported 41% accuracy, where it improved to 66% after applying them (Table 1).

XLM-RoBERTa-base better predicts actual positive labels (Figure 2). However, it reports higher false negative (FN) values for negative classes and more false positive (FP) values for positive classes. In contrast, BanglaBERT reports fewer FP and FN values for each class but fails to predict TP with about 103 data points deviation. Therefore, XLM-RoBERTa-base and BanglaBERT performed well on the test dataset. FP and FN for each class in LSTM made minimal impact on accuracy because their values show a slightly equal distribution. Though LSTM generates better accuracy than the transformer model, transformers produce more correct instances for negative and other classes. In Figure 4, the learning curve backs up the finding of the unstable nature of the LSTM model, showing how it is underfitting. We saw the similar pattern for traditional ML models such as SVM. Therefore, models like LSTM and SVM may not generalize to another dataset with new

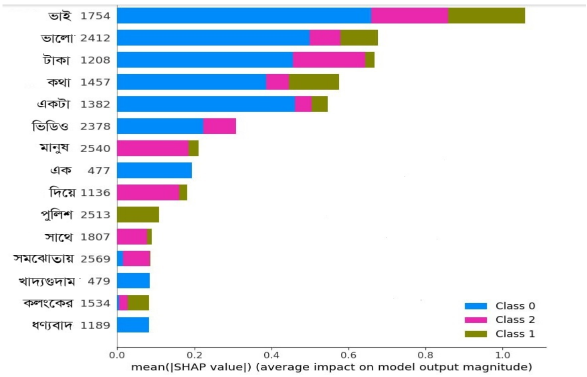


Figure 3: SHAP on XLM-RoBERTa-base output- blue (Positive), green (Negative) and pink (Neutral)

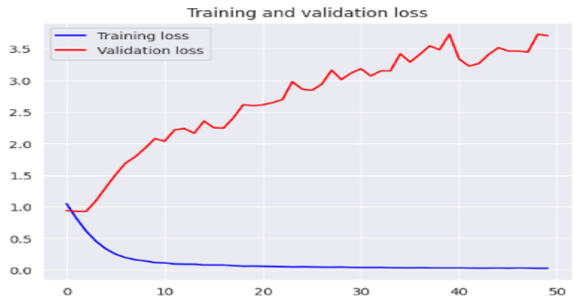


Figure 4: Learning Curve for Underfit LSTM

test instances. Since the class distribution is imbalanced in the dataset, we also calculated other metrics such as F1-score, precision, and recall, which basically signifies if the model is doing a better job for lowered-numbered classes. For example, none of the models did a great job with the "neutral" class showing a lower f1-score, precision, and recall, which also syncs with the confusion matrix.

We used SHAP (Lundberg and Lee, 2017) - a state-of-the-art explainable AI (XAI) tool, to interpret the classification results of the XLM-RoBERTa-base transformer model's output, in our case "accuracy" (Figure 3). This SHAP plot combines the significance of the features with their impacts. The Y-axis lists the features from top to bottom or most important to least important order. The labels on the Y-axis represent the most influential word features for XLM-RoBERTa-base and their associated indexing in the word vector. The x-axis shows the Shapely values from 0 to 1. Blue, green, and pink spectrum are representations of Shapley values for "positive," "negative," and "neutral" classes. Not only the length of the spectrum but also the color has significance. For example, the "পুলিশ" feature correlates less than 20% with the model output accuracy. However, this word influences a post's identification as only negative (green color). Another good example is the "ভাই" feature, the most influential feature in the predictions with XLM-RoBERTa-base. The Shapely value for blue (positive) is 70%, whereas for pink (neutral) and green (negative) is 20%. That means having a "ভাই" word in a post mostly co-related to a positive post, which is also intuitively correct since it is a respectful salutation. The Shapely values of the features are more positively correlated with the positive class (labeled with blue) since blue spectrums are larger than the others. The neutral class (labeled with pink) has the lowest correlation with the model output. This result also aligns with the confusion matrix (Figure 3), where prediction accuracy for positive classes is higher with XLM-RoBERTa. Therefore, the positive class operated on a higher accuracy scale with a higher correlation of approximately 70% with the most influential feature (feature 1754). The plot also shows that the impact of the "negative class" is very low- it does not frequently appear as the positive or neutral classes.

5 Limitations and Conclusion

In summary, we compared multiple ML approaches to discuss the multi-class sentiment analysis. We analyzed and compared results based on preprocessing techniques, rigorous output analysis, and XAI. Our analysis shows that the XLM-RoBERTa-base generates a stable model even with lower accuracy regarding confusion matrix, evaluation metrics, and XAI than LSTM and traditional models. The first challenge we faced is that vector assembler on huge data made the dimensions of the feature very large and computationally expensive, difficult to address with low computing resources. Secondly, the highly imbalanced dataset has only 20% "Neutral" labels, which skewed the prediction against this class and caused some models to underfit. Developing MLM-based masked models with oversampled datasets improved the quality of the classification tasks for XLM-RoBERTa. It understands contextual relationships between words better and effectively predicts missing or masked words within a sentence. Our future work will focus on mitigating the challenge of Bangla sentiment analysis for lacking high-quality datasets, generalizable tools, comprehensive sentiment lexicons, and standardized evaluation metrics.

Ethics Statement

All the authors are trained in the ethical conduct of research. Ethical usage of data, analysis, writing, and transparency of implementation have been maintained by sharing the implementation.

References

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).

Nitish Ranjan Bhowmik, Mohammad Arifuzzaman, and M Rubaiyat Hossain Mondal. 2022. Sentiment analysis on bangla text using extended lexicon dictionary and deep learning algorithms. *Array*, 13:100123.

Ni Nyoman Budiasih, TAB Wirayuda, and RN Dayawati. 2009. Analisis dan implementasi stemming teks berbahasa indonesia dengan menggunakan porter stemmer. *Tugas Akhir Teknik Informatika, Fakultas Teknik Informatika, Universitas Telkom: Bandung*.

Pallab Chowdhury, Ettilla Mohiuddin Eumi, Ovi Sarkar, and Md Faysal Ahamed. 2022. Bangla news classification using glove vectorization, lstm, and cnn. In *Proceedings of the International Conference on Big Data, IoT, and Machine Learning: BIM 2021*, pages 723–731. Springer.

Maik Fröbe, Benno Stein, Tim Gollub, Matthias Hagen, and Martin Potthast. 2023. Semeval-2023 task 5: Clickbait spoiling. In *Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2275–2286.

Md. Arid Hasan, Firoj Alam, Anika Anjum, Shudipta Das, and Afiyat Anjum. 2023a. Blp-2023 task 2: Sentiment analysis. In *Proceedings of the 1st International Workshop on Bangla Language Processing (BLP-2023)*, Singapore. Association for Computational Linguistics.

Md Arid Hasan, Firoj Alam, Shammur Absar Chowdhury, and Naira Khan. 2019. Neural machine translation for the bangla-english language pair. In *2019 22nd International Conference on Computer and Information Technology (ICCIT)*, pages 1–6. IEEE.

Md. Arid Hasan, Shudipta Das, Afiyat Anjum, Firoj Alam, Anika Anjum, Avijit Sarker, and Sheak Rashed Haider Noori. 2023b. Zero- and few-shot prompting with llms: A comparative study with fine-tuned models for bangla sentiment analysis.

Asif Hassan, Mohammad Rashedul Amin, Abul Kalam Al Azad, and Nabeel Mohammed. 2016. Sentiment analysis on bangla and romanized bangla text using deep recurrent models. In *2016 International Workshop on Computational Intelligence (IWCI)*, pages 51–56.

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. Ekramul Islam, Labib Chowdhury, Faisal Ahamed Khan, Shazzad Hossain, Md Sourave Hossain, Mohammad Mamun Or Rashid, Nabeel Mohammed, and Mohammad Ruhul Amin. 2023a. Sentigold: A large bangla gold standard multi-domain sentiment analysis dataset and its evaluation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, page 42074218, New York, NY, USA. Association for Computing Machinery.

Md Ekramul Islam, Labib Chowdhury, Faisal Ahamed Khan, Shazzad Hossain, Md Sourave Hossain, Mohammad Mamun Or Rashid, Nabeel Mohammed, and Mohammad Ruhul Amin. 2023b. Sentigold: A large bangla gold standard multi-domain sentiment analysis dataset and its evaluation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4207–4218.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc.

Md. Shihab Mahmud, Md. Touhidul Islam, Afrin Jaman Bonny, Rokeya Khatun Shorna, Jasia Hossain Omi, and Md. Sadekur Rahman. 2022. Deep learning based sentiment analysis from bangla text using glove word embedding along with convolutional neural network. In *2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–6.

Walaa Medhat, Ahmed Hassan, and Hoda Korashy. 2014. Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, 5(4):1093–1113.

Gaku Morio, Hiroaki Ozaki, Atsuki Yamaguchi, and Yasuhiro Sogawa. 2022. Hitachi at SemEval-2022 task 10: Comparing graph- and Seq2Seq-based models highlights difficulty in structured sentiment analysis. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1349–1359, Seattle, United States. Association for Computational Linguistics.

Md Saddam Hossain Mukta, Md Adnanul Islam, Faisal Ahamed Khan, Afjal Hossain, Shuvanon Razik, Shazzad Hossain, and Jalal Mahmud. 2021. A comprehensive guideline for bengali sentiment annotation. *Transactions on Asian and Low-Resource Language Information Processing*, 21(2):1–19.

Md. Atikur Rahman and Emon Kumar Dey. 2018. Datasets for aspect-based sentiment analysis in bangla and its baseline evaluation. *Data*, 3(2).

Michael Roth, Talita Anthonio, and Anna Sauer. 2022. SemEval-2022 task 7: Identifying plausible clarifications of implicit and underspecified phrases in instructional texts. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1039–1049, Seattle, United States. Association for Computational Linguistics.

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

Salim Sazzed. 2021. Abusive content detection in transliterated bengali-english social media corpus. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 125–130.

Omar Sharif, Mohammed Moshiul Hoque, and Eftekhar Hossain. 2019. Sentiment analysis of bengali texts on online restaurant reviews using multinomial naïve bayes. In *2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT)*, pages 1–6.

Sumit Singh, Pawankumar Jawale, and Uma Tiwary. 2022. silpa_nlp at semeval-2022 tasks 11: Transformer based ner models for hindi and bangla languages. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1536–1542.

Nasrin Sultana, Rehena Sultana, Risul Islam Rasel, and Mohammed Moshiul Hoque. 2022. Aspect-based sentiment analysis of bangla comments on entertainment domain. In *2022 25th International Conference on Computer and Information Technology (ICCIT)*, pages 953–958.

Manan Suri. 2022. Pickle at semeval-2022 task 4: Boosting pre-trained language models with task specific metadata and cost sensitive learning. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 464–472.

Maryam Tahir, Ahmad Naeem, Hassaan Malik, Jawad Tanveer, Rizwan Ali Naqvi, and Seung-Won Lee. 2023. Dscc_net: Multi-classification deep learning models for diagnosing of skin cancer using dermoscopic images. *Cancers*, 15(7):2179.

Prerona Tarannum, Firoj Alam, Md Arid Hasan, and Sheak Rashed Haider Noori. 2022. Z-index at checkthat! lab 2022: Check-worthiness identification on tweet text. *arXiv preprint arXiv:2207.07308*.

Yuki Taya, Lis Kanashiro Pereira, Fei Cheng, and Ichiro Kobayashi. 2021. OCHADAI-KYOTO at SemEval-2021 task 1: Enhancing model generalization and robustness for lexical complexity prediction. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 17–23, Online. Association for Computational Linguistics.

Ming Wang, Heike Adel, Lukas Lange, Jannik Strotgen, and Hinrich Schütze. 2023. Nlnde at semeval-2023 task 12: Adaptive pretraining and source language selection for low-resource multilingual sentiment analysis. *ArXiv*, abs/2305.00090.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.