M1437 at BLP-2023 Task 2: Harnessing Bangla Text for Sentiment Analysis: A Transformer-based Approach

Majidur Rahman and **Özlem Uzuner**
George Mason University, Virginia, USA
{mrahma37, ouzuner}@gmu.edu

Abstract

Analyzing public sentiment on social media is helpful in understanding the public's emotions about any given topic. While numerous studies have been conducted in this field, there has been limited research on Bangla social media data. Team M1437 from George Mason University participated in the Sentiment Analysis shared task of the Bangla Language Processing (BLP) Workshop at EMNLP-2023. The team fine-tuned various BERT-based Transformer architectures to solve the task. This article shows that $BanglaBERT_{large}$, a language model pre-trained on Bangla text, outperformed other BERT-based models. This model achieved an F1 score of 73.15% and top position in the development phase, was further tuned with external training data, and achieved an F1 score of 70.36% in the evaluation phase, securing the fourteenth place on the leaderboard. The F1 score on the test set, when $BanglaBERT_{large}$ was trained without external training data, was 71.54%.

1 Introduction

Social networking platforms have emerged as avenues where people share their thoughts and feelings on diverse subjects such as entertainment, politics, and education (Chen et al., 2022). Natural Language Processing (NLP) can effectively evaluate the sentiment of a text (Medhat et al., 2014) and explore the information discussed in social networking platforms. However, most research in this field has focused on English as the primary language; many other languages (e.g., Bangla) have remained largely unexplored (Sazzed, 2020; Islam et al., 2020).

Despite being the seventh most commonly spoken language worldwide, as well as the sixth in terms of native speakers (Babbel, 2023), Bangla is regarded as a low-resource language (Alam et al., 2021). The inaugural Bangla Language Processing (BLP) Workshop (Hasan et al., 2023a) sought to address sentiment analysis of Bangla social media posts. Within the scope of this workshop's sentiment analysis shared task, two datasets were utilized: the **MUltiplatform BAngla SEntiment (MUBASE)** (Hasan et al., 2023b) dataset, which features tweets and Facebook posts paired with their corresponding sentiment polarity, and the **Sentiment on Noisy Bangla texts (SentNoB)** (Islam et al., 2021) dataset, which consists of user comments on news articles and social media videos in various domains, such as education, politics, etc.

This paper presents our solution to sentiment analysis in Bangla on the workshop datasets. Our experiments with various Bidirectional Encoder Representations from Transformers (BERT)-based models (Devlin et al., 2019) indicated that $BanglaBERT$ (Bhattacharjee et al., 2022), a BERT language model that is pretrained on more than 27 GB of Bangla data is effective for classifying Bangla text sentiment. This system achieved an F1 score of 73.15% during the development phase. To further improve performance, we supplemented the training set with the CogniSenti dataset (Hasan et al., 2020) containing Facebook posts and tweets authored by Bangla speakers. This updated system achieved the best F1 score of 70.36% on the test set, securing the fourteenth place on the evaluation leaderboard. Without training data from CogniSenti Dataset, the F1 score was 71.54%. Our code is publicly available on GitHub[1].

2 Related Work

Extensive research has been carried out regarding sentiment analysis in languages with abundant resources, such as English. Traditional sentiment analysis approaches on resource-abundant languages relied heavily on syntactic parsing (Nasukawa and Yi, 2003). The advent of Transformer-based architectures (Vaswani et al., 2017), such as

[1] https://github.com/majidurrahman1437/blp-shared-task2

Proceedings of the First Workshop on Bangla Language Processing (BLP-2023), pages 279–285
December 7, 2023 ©2023 Association for Computational Linguistics

BERT (Devlin et al., 2019), greatly improved the state-of-the-art (Socher et al., 2013) on sentiment classification (Munikar et al., 2019).

Low-resource languages have traditionally lagged behind these advancements. In recent years, however, the NLP community has turned its attention to low-resource languages like Bangla. Sentiment analysis for low-resource languages became one of the tasks to receive attention. The availability of high-quality datasets, such as aspect-based sentiment analysis (ABSA) of Bangla text (Rahman et al., 2018) dataset, has supported sentiment analysis in Bangla. Example approaches to sentiment analysis on Bangla primarily utilized long short-term memory (LSTM) units (Hochreiter and Schmidhuber, 1997; Tripto and Ali, 2018; Rezaul Karim et al., 2020). The SentNoB dataset (Islam et al., 2021) was introduced in 2021, which consists of noisy Bangla texts. Islam et al. showed that combining lexical features resulted in better performance than neural models for SentNoB. Hasan et al. developed the CogniSenti dataset (Hasan et al., 2020), which leverages Transformer models like XLM-RoBERTa (Conneau et al., 2020) to predict sentiment polarity in Bangla text, with promising results.

In a recent comparative study of various Bangla sentiment classification datasets using different Transformer-based architectures, XLM-RoBERTa outperformed all models (Alam et al., 2021). These results demonstrate the growing potential of Transformer-based architectures to improve language processing even in low-resource languages such as Bangla. BanglaBERT (Bhattacharjee et al., 2022) is a language model based on BERT, pre-trained on a large dataset of 27.5 GB of Bangla text. It has yielded state-of-the-art results in Bangla sentiment classification. While there are some promising research directions for Large Language Models (LLM) to perform Bangla sentiment analysis (Hasan et al., 2023b), existing pre-trained language models, such as BanglaBERT, can outperform them. Although there has been a sentiment analysis shared task for Indian languages, including Bangla, in the past (Patra et al., 2015), there has been a lack of initiatives to organize such a task for the Bangla language specifically. The Bangla sentiment analysis shared task at the first BLP workshop (Hasan et al., 2023a) aims to highlight the research efforts of Bangla researchers from around the world.

3 Methods

3.1 Data

The dataset used in this shared task consists of samples from the MUltiplatform BAngla SEntiment (MUBASE) (Hasan et al., 2023b) and SentNoB (Islam et al., 2021) datasets. The former contains Bangla language posts from social media platforms like Twitter and Facebook, which have undergone manual annotation for sentiment analysis. The latter comprises comments from multiple social media domains; each has also been manually annotated for sentiment.

The dataset comprises three sentiment classes: Negative, Neutral, and Positive. The proportion of Negative, Neutral, and Positive examples is kept uniform across the training and validation splits, whereas in the test split, the ratio is almost similar to the train and validation split. The distribution of labels across data splits is illustrated in Figure 1.

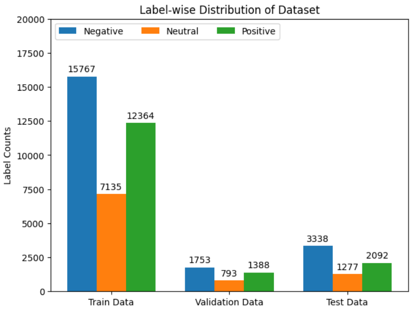

Figure 1: Label-wise Distribution of the Dataset

External Training Data: As can be observed from the class distribution of the training data, the "Neutral" class is under-represented compared to the other two sentiment classes. In order to boost the representation of the "Neutral" class and to reduce the class imbalance, we integrated external training data from the CogniSenti dataset (Hasan et al., 2020) to the original training set. The CogniSenti dataset consists of 6570 instances, encompassing three sentiment categories (Negative, Neutral, Positive) extracted from Tweets and Facebook posts written by native Bangla speakers. It features examples from diverse domains, including politics, current affairs, etc. By merging the complete dataset with the provided training set, we create a new training set of more than 41k samples. The distribution of the dataset across various categories is presented in Table 1.

Dataset	Class	Number of Samples
CogniSenti	Negative	1333
	Neutral	3749
	Positive	1488
	Total	6570
Merged (BLP Train Set + CogniSenti)	Negative	17100
	Neutral	10884
	Positive	13852
	Total	41836

Table 1: Data Distribution of External Training Data (CogniSenti Dataset), Along With Merged Training Data Per Class

3.2 BanglaBERT

BanglaBERT language model utilizes ELECTRA (Clark et al., 2020) as its foundation due to ELECTRA's superior computational efficiency compared to BERT. BanglaBERT is pre-trained on 27.5 GB of Bangla text from various sources such as news, encyclopedias, and blogs. The $BanglaBERT_{base}$ model includes 12 Transformer Encoder layers with 768 hidden units, 12 attention heads, and 110M parameters, while the $BanglaBERT_{large}$ model boasts 24 Transformer Encoder layers with 1024 hidden units, 16 attention heads, and 335M parameters (Bhattacharjee et al., 2022).

3.3 Evaluation

The official evaluation metric for the Sentiment Analysis shared task is the micro-F1 score (Pedregosa et al., 2011).

3.4 Experimental Setup

We utilized BanglaBERT with the aid of HuggingFace transformers library (Wolf et al., 2019). Our model is trained on NVIDIA DGX-A100 GPU nodes, with a maximum input sequence length of 512. We conducted hyperparameter tuning on the learning rate, seed, training batch size, and number of training epochs to achieve optimal performance. The model undergoes ten epochs of training, with a training batch size of 32 and a seed value of 18. We set the learning rate 3e-5 utilizing the Adam (Kingma and Ba, 2014) optimizer and a linear warmup with a warmup ratio of 0.001. We develop our models on the provided development set and validate utilizing the development-test (dev-test) and test sets during the development and evaluation phases, respectively.

4 Results and Discussions

During the development phase, our system attains the top position in the leaderboard, which is evaluated using the dev-test split. During the evaluation phase, our model ranks as the fourteenth-best model evaluated using the test split, as illustrated in Table 2.

Model	F1 Score (%)
Development Phase	
M1437	**73.15**
MoFa_Aambela	73.03
yangst	72.88
Hari_vm	72.48
amlan107	72.24
Evaluation Phase	
MoFa_Aambela	**73.10**
yangst	72.67
amlan107	71.79
Hari_vm	71.72
PreronaTarannum	71.64
ShadmanRohan	71.55
M1437 (latest submission)	70.36
M1437 (best submission)	71.54

Table 2: Performance Comparison on the Dev-Test Set and Test Set of Our System Submissions

4.1 Performance with External Data

Upon merging the CogniSenti dataset with the BLP sentiment analysis shared task train set, we analyze our latest submission, which utilizes $BanglaBERT_{large}$. Unfortunately, we discovered that incorporating external data did not improve the performance of our model. Following an in-depth investigation into our model's inaccuracies, we uncovered that 331 instances were classified as "Positive" when they should have been labeled as "Negative". Upon further analysis of these predictions, including phrases such as 'চ্যালেঞ্জের মুখে নার্সারি ব্যবসায়ীরা" (The nursery traders are facing challenges), "কেন ঝুঁকি থাকলেও এখনো মশক কর্মী নিয়োগ হচ্ছে না?" (Why the mosquito workers are still not recruited despite having risks?), we observed that our model struggled to detect the "Negative" sentiment in these samples accurately. On the contrary, the model that was trained without incorporating CogniSenti data accurately identified 91 of the 331 "Negative" samples.

Our further analysis discovered that 50% of the incorrect predictions were originally labeled as "Negative" but fell under the "Neutral" category. Likewise, 33.26% of mispredictions were previously labeled as “Positive” but were classified as “Neutral”. Examples of “Negative” sentiments that were misclassified as “Neutral” were found in the CogniSenti data-trained model, such as “এদিকে শহরের মানুষের বিদ্যুৎ অপচয় ও বিদ্যুৎ নির্ভরতা বাড়তেই আছে !” (Meanwhile, the city’s electricity consumption and electricity dependence continues to increase!), “আপনার কাছ থেকে এমন বক্তব্য আশা করি না” (I do not expect such a statement from you). However, these examples were predicted correctly by the model trained without CogniSenti data. The merged dataset had a higher proportion of “Neutral” to “Negative” samples, resulting in a more effective prediction of “Neutral” sentiment examples but leading to a higher number of mispredictions for the “Positive” and “Negative” sentiment examples compared to the model trained without CogniSenti data. This is supported by the fact that the model trained without CogniSenti data made 501 mispredictions for the “Negative” sentiment category, while the model trained with CogniSenti data made 663 mispredictions for the same category.

4.2 Performance Without External Data

Model	**F1 Score (%)**
Random Baseline	33.56
Majority Baseline	49.77
n-gram Baseline	55.14
Logistic Regression	55.05
Decision Tree Classifier	48.68
multi-lingual BERT-cased	64.20
XLM-RoBERTa_large	68.21
MuRIL_base	68.39
IndicBERT	70.82
$BanglaBERT_{base}$	71.49
$BanglaBERT_{large}$	**71.54**

Table 3: Performance Comparison on the Test Set Across Various BERT Models

Comparative Study Across Baselines: Prior to the commencement of the shared task, the organizers released three baseline scores for the dev-test set and the test set. The initial score, referred to as the random baseline, randomly predicts a label from the three likely class labels. The second score, known as the majority baseline, employs the “DummyClassifier” from the sklearn library (Pedregosa et al., 2011) and predicts the most frequent class label for each instance. Lastly, the third baseline, named the n-gram baseline, employs the TF-IDF vectorization (Salton and Buckley, 1988) technique to generate feature vectors and the Support Vector Machine classifier (Noble, 2006) to provide predictions on the test set. Moreover, we have conducted a comparison of our model’s performance, $BanglaBERT_{large}$, trained only on the BLP sentiment analysis train set by utilizing the test set specified in Table 3, with that of conventional machine learning classifiers, namely Logistic Regression (Wright, 1995) and Decision Tree Classifier (Swain and Hauska, 1977). To extract features, we utilized a similar TF-IDF vectorization technique and independently applied Logistic Regression and Decision Tree Classifier to generate predictions on the test set. Our assessment demonstrates that the baselines and traditional machine learning classifiers were not able to develop a robust model due to their inability to grasp the intricacies of the input text and context.

Comparative Study Across BERT models: We further assess the performance of our top-performing model as specified previously. Our findings reveal that the BERT-based models exceed the performance of other models chosen for comparison. One of these models is the multi-lingual BERT-cased (mBERT) (Devlin et al., 2019), which is trained in 104 languages, including Bangla. However, it’s worth noting that multi-lingual models typically yield better results for high-resource languages and may not perform as well on lower-resource languages like Bangla (Wu and Dredze, 2020). Multilingual language models such as MuRIL (Khanuja et al., 2021) and IndicBERT (Doddapaneni et al., 2023) have undergone pretraining on a range of Indian languages, including Bangla, through the use of monolingual, translated, and transliterated text. These models have demonstrated superior performance in comparison to mBERT, a similar multilingual language model. However, it is worth highlighting that although these models are multilingual, this is also the primary reason for their inability to surpass our model’s performance. Research has shown that XLM-RoBERTa (XLM-R) (Conneau et al., 2020), despite having more model parameters (550M), is unable to outperform $BanglaBERT_{large}$ due to its limited pretraining knowledge of Bangla text (8.7 GB). In contrast, $BanglaBERT_{large}$ has access to a vast amount of pretraining knowledge (27.5 GB) specific to the Bangla language. This highlights the importance of having a substantial amount of language-specific pretraining knowledge, which aids in generating robust context-

No. of Input Tokens	No. of Train Samples	No. of Test Samples	Prediction Correctness (%)
1 to 20	28108	5490	72.91
21 to 40	5612	895	67.82
41 to 60	994	174	65.52
61 to 80	241	69	53.62
81 to 100	101	29	48.28
101 to 150	109	28	46.43
151 to 200	40	12	33.33
201 or higher	170	10	60.00

Table 4: Performance Comparison of Test Set According to Input Token Length

aware embedding vectors and ultimately improves model performance.

Error Analysis: Based on our findings, it appears that the model trained without CogniSenti data exhibits higher true positive rates for the "Negative" and "Positive" classes at 84.99% and 75.76%, respectively, compared to only 29.44% for the "Neutral" class. Our model is more proficient at learning examples from the "Negative" and "Positive" classes while struggling with the "Neutral" class due to the data imbalance in our training set. In fact, 69.59% of the mispredictions regarding the "Neutral" class actually belong to the "Negative" class, which can be attributed to the larger number of "Negative" examples in our training set. To ensure unbiased outcomes, a well-balanced dataset with comparable sample sizes in each class is essential for optimal performance.

Examining FP and FN: We thoroughly analyze the mispredictions made by the model trained without CogniSenti data, specifically when it predicts a "Positive" sentiment instead of a "Negative" one, or vice versa, according to the gold label. We examine texts such as "উনার রেস্ট দরকার । । ।" (They need rest...), "একাদশ জাতীয় সংসদের ৯ম অধিবেশন শুরু ৬ সেপ্টেম্বর" (The 9th session of the eleventh National Parliament begins on September 6), "আস্তাগফিরুল্লাহ!" (God forgive us!), and "গ্রাহকের কাঁধে আর থাকছে না বাড়তি বিলের বোঁঝা" (The customer is no longer burdened with additional bills). While these texts are labeled as "Negative" in the gold label, the model may not have enough background knowledge to accurately label them as "Negative" instead of "Positive". Similar cases have been observed in texts such as "দেশে আরও ৩ জন করোনাভাইরাসে আক্রান্ত" (3 new people infected with coronavirus in the country), "ছিল না ফরম ফিলাপের টাকা, অভাব ছিল নিত্যসঙ্গী" (Neither there was money for filling up the form nor a daily companion) where the model predicts a text to be "Negative" whereas the actual label is "Positive". It has been observed that a significant number of erroneous predictions can be attributed to political and national affairs, which are over-represented in the dataset. It is imperative to acknowledge the potential biases that can result from such imbalances and to devise strategies for mitigating their impact to ensure the accuracy and reliability of the predictions. This issue highlights the importance of careful data curation and analysis in the context of predictive modeling, particularly when dealing with sensitive or complex domains.

Performance Comparison by Text Length: To assess the performance of the model trained without CogniSenti data on texts of varying lengths, we closely monitor its predictions on the test set. Our evaluation reveals that our model accurately predicts approximately most of the 5.5k samples with up to 20 tokens. However, as the input text length increases, the F1 score declines. Notably, the model's F1 score is highest (72.91%) for texts with up to 20 tokens, dropping to 33.33% for texts with 151 to 200 tokens. This suggests that the model learns to predict shorter texts more precisely, possibly due to more training examples with 20 tokens or less as per table 4. In order to facilitate the learning process for longer inputs, it may be advantageous to consider augmenting the training data with lengthier texts.

5 Conclusion

Team M1437 had the privilege of participating in the Bangla Sentiment Analysis challenge during the inaugural BLP workshop at EMNLP-2023. For this task, we prefer the $BanglaBERT_{large}$ as our language model due to its exceptional pre-trained proficiency in the Bangla language. During development, our system ranked first on the leaderboard. Although we achieved a comparable F1 score during the evaluation phase, we remain committed to exploring a range of Large Language Models (LLMs) to improve the true positive rates for longer input sequences.

Limitations

In an effort to enhance our model's ability to generalize across all labels, we integrated the CogniSenti dataset into the training set. Unfortunately, the model's performance did not meet our expectations in this particular scenario. However, this can be due to the specific dataset chosen and leaves open the question of whether other datasets would yield similar results. We, therefore, remain committed to examining other relevant datasets that can not only supplement the training data but also enhance the model's performance across all sentiment classes.

Ethics Statement

The dataset used in this research complies with a non-commercial share-alike international license by Creative Commons [2], which is taken under careful consideration. The research does not use this dataset for any commercial purpose.

References

Firoj Alam, Md Arid Hasan, Tanvir 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. *arXiv preprint arXiv:2107.03844*.

Babbel. 2023. The 10 most spoken languages in the world.

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.

Junhan Chen, Yumin Yan, and John Leach. 2022. Are emotion-expressing messages more shared on social media? a meta-analytic review. *Review of Communication Research*, 10.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *ICLR*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

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.

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, 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.

Md. Arid Hasan, Jannatul Tajrin, Shammur Absar Chowdhury, and Firoj Alam. 2020. Sentiment classification in bangla textual content: A comparative study. In *23rd International Conference on Computer and Information Technology (ICCIT)*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

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. IEEE.

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.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja

[2] https://creativecommons.org/licenses/by-nc-sa/4.0/

Nagipogu, Shachi Dave, et al. 2021. Muril: Multilingual representations for indian languages. *arXiv preprint arXiv:2103.10730*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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

Manish Munikar, Sushil Shakya, and Aakash Shrestha. 2019. Fine-grained sentiment classification using bert. In *2019 Artificial Intelligence for Transforming Business and Society (AITB)*, volume 1, pages 1–5.

Tetsuya Nasukawa and Jeonghee Yi. 2003. Sentiment analysis: Capturing favorability using natural language processing. In *Proceedings of the 2nd international conference on Knowledge capture*, pages 70–77.

William S Noble. 2006. What is a support vector machine? *Nature biotechnology*, 24(12):1565–1567.

Braja Gopal Patra, Dipankar Das, Amitava Das, and Rajendra Prasath. 2015. Shared task on sentiment analysis in indian languages (sail) tweets-an overview. In *Proc. of MIKE*, pages 650–655. Springer.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Md Rahman, Emon Kumar Dey, et al. 2018. Datasets for aspect-based sentiment analysis in bangla and its baseline evaluation. *Data*, 3(2):15.

Md Rezaul Karim, Bharathi Raja Chakravarthi, Mihael Arcan, John P McCrae, and Michael Cochez. 2020. Classification benchmarks for under-resourced Bengali language based on multichannel convolutional-lstm network. *arXiv*, pages arXiv–2004.

Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523.

Salim Sazzed. 2020. Cross-lingual sentiment classification in low-resource Bengali language. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 50–60, Online. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Philip H Swain and Hans Hauska. 1977. The decision tree classifier: Design and potential. *IEEE Transactions on Geoscience Electronics*, 15(3):142–147.

Nafis Irtiza Tripto and Mohammed Eunus Ali. 2018. Detecting multilabel sentiment and emotions from bangla youtube comments. In *Proc. of ICBSLP*, pages 1–6. IEEE.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Raymond E Wright. 1995. Logistic regression.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online. Association for Computational Linguistics.