BERT is all you need

Sifat Redwan
Department of Computer Science
University of Minnesota Duluth

Abstract

Sentiment analysis (SA) refers to the process of identifying sentiments or emotions within textual data. It is a crucial task in natural language processing (NLP) as it enables applications such as opinion mining, customer feedback analysis, and social media monitoring, which are vital for decision-making across various domains. Although extensive research has been conducted on SA, the majority of studies focus on English text, leaving other languages, such as Bangla, underexplored. Despite being widely spoken, Bangla remains a low-resource language, primarily due to the limited availability of labeled datasets and state-of-the-art language models. To address these challenges, this work leverages BanglaBERT on the benchmark sentiment analysis dataset SentNoB. Additionally, experiments were conducted using traditional machine learning (ML) models with linguistic features, as well as pretrained language models, to compare their performance. The findings indicate while pretrained language models perform better than linguistic-feature based methods they still, they still face challenges in capturing the unique and complex characteristics of language, particularly for low-resource languages like Bangla. I have made this project publicly available here https://github.com/SifatRedwan1/Sentiment_Analysis.git.

1 Introduction

Social media platforms are greatly utilized these days for expressing the opinions or views about any product, topic, event, or any breaking news from anywhere at any time. Sentiment analysis (SA) is a field which involves the analysis of such opinions. The purpose of sentiment analysis is to automatically determine the expressive direction of user reviews (Luo et al., 2016). Here, the polarity of textual data implies finding out whether the sentiments of given textual data are positive, negative, or neutral (Chauhan et al., 2021). SA has been successfully employed in financial market prediction, health issues, customer analytics, commercial valuation assessment, brand marketing, politics, crime prediction, and emergency management. To this day most large E-commerce platforms and political campaigns rely heavily on SA for their decision-making and strategy. An illustrative example is the publication by (Ravi and Ravi, 2015), who conducted a study summarizing over one hundred papers published between 2002 and 2015, focusing on the applications of sentiment analysis, the different approaches and open issues in the field. For these reasons SA has an critical impact all over the world. However, SA faces numerous challenges due to ambiguity in language, domain-specific language, multilingual language, emotion intensity and nuance, data quality and imbalance etc. Additionally, most of the research works done in SA is on English texts as English labeled texts are largely available. However, as people now can express their opinions in their languages in social media or media platforms, performing SA in scarced or low-resource languages poses new challenges. To address this issue I focused on performing SA in my first language Bangla.

Bangla is the seventh most spoken language worldwide and the second Indo-Aryan language after Hindi with 278M speakers which approximately is 3.5% of the world's population [1]. Despite being the seventh most spoken language in the world , Bangla is considered a resource-scarce language. Bangla as a language lacks labeled data collection and relies heavily on self-supervised pretrained language models (PLMs) which are mostly trained in English texts. However, as multilingual PLMs (Conneau and Lample, 2019; Conneau, 2019) are trained to perform on wide range

[1] https://en.wikipedia.org/wiki/List_of_languages_by_total_number_of_speakers

of languages they also require large amount of data for fine-tuning. They also tend to show degraded performance on low-resources language (Wu and Dredze, 2020). Motivated by this issue, I aim to fine-tune a robust model called BanglaBERT to achieve state-of-the-art performance on SentNoB, a widely recognized Bangla benchmark dataset for sentiment analysis in noisy texts. In summary the contribution of this project can be summarized as:

- Fine-tuning the BanglaBERT PLM on a benchmark Bangla Sentiment Analysis dataset.
- Experimenting on different techniques such as linguistic feature based ML methods, PLMs and analyzing their performances for comparison and future research directions.

2 Related Works

There are several approaches to sentiment analysis, which can broadly be categorized into three main types: lexicon/rule-based, machine learning-based, and deep learning-based methods. The lexicon/rule-based approach uses the polarity of words, while the ML method sees texts as a classification problem and can be further divided into unsupervised, semi-supervised, and supervised learning (Aqlan et al., 2019).

2.1 Lexicon or rule-based approaches

Lexicon-based or rule-based approaches to sentiment analysis rely on predefined dictionaries of words associated with specific sentiment scores. These methods calculate the sentiment of a given text by analyzing the sentiment scores of the words it contains. For example, VADER (Hutto and Gilbert, 2014), a rule-based sentiment analysis model optimized for social media, combining validated lexical features with grammatical rules, outperforming human raters and state-of-the-art ML models of that time.

In my baseline paper (Islam et al., 2021a), the authors have proposed that the old school handcrafted lexical features based methods provide superior performance than neural network and pretrained language models.

2.2 Machine Learning-based approaches

ML based sentiment analysis can be classified into unsupervised, semi-supervised, and supervised learning based approaches. Machine learning-based sentiment analysis leverages supervised learning techniques to classify text into sentiment categories by training models on human labeled datasets. Text is preprocessed through tokenization, stemming/lemmatization, and stopword removal before being transformed into numerical features using methods like Bag of Words (BoW), Term Frequency-Inverse Document Frequency (TF-IDF), or dense word embeddings (e.g., Word2Vec, GloVe). Common algorithms include Naive Bayes, Logistic Regression, Support Vector Machines (SVMs), and tree-based methods like Random Forests or Gradient Boosting (e.g., XGBoost). These approaches excel at generalization and domain-specific tasks, achieving higher accuracy than rule-based models, but are constrained by the need for high-quality labeled data and reliance on effective feature engineering. While supervised machine learning approaches were initially dominant, the introduction of transformers (Vaswani, 2017) has shifted the focus towards pretrained language models, which are now widely used for sentiment analysis tasks.

2.3 Deep Learning-based approaches:

Recent years have shown a trend of deep learning models applied in the field of NLP. Deep neural networks (DNNs) are made up of artificial neural networks having multiple hidden layers between the input layer and the output layer. Sentiment analysis tasks can be performed efficiently by implementing different deep learning models, which have been extended recently. These models include CNN (convolutional neural networks), RNN (recursive neural network), DNN (deep neural networks), RNN (recurrent neural networks) and DBN (deep belief networks) (Yadav and Vishwakarma, 2020). While deep neural network (DNN)-based models outperform traditional machine learning models, their effectiveness relies on training with extensive text corpora. For sentiment analysis, a large, labeled dataset is essential for accurate classification, but the process of data collection and annotation is often resource-intensive and time-consuming.

3 Dataset

In this paper I used an annotated sentiment analysis dataset named SentoNoB, made of informally written Bangla texts proposed in this paper (Islam et al., 2021a). This dataset comprises of public comments on news and videos collected from so-

cial media covering 13 different domains, including politics, education, and agriculture. These comments are labeled with one of the polarity labels, namely positive, negative, and neutral. One significant aspect of this dataset is that every comments are noisy and informally written in terms of grammatical correctness and context understanding. The dataset was mainly collected from the public comments of articles on the newspaper Prothom Alo [2], one of the most circulated newspaper in Bangladesh. Then, public comments were also collected from youtube in similar topics.

Out of ~31K collected comments, the comments that are written in only Bangla alphabets were kept. To reduce repetitiveness and noise, duplicates and instances shorter than three or longer than 50 words tokens were excluded. Additionally, to increase the uniqueness of the vocabulary, different words are incorporated as much as possible. Therefore, we prioritize the instances for annotation that will increase the percentage of the unique word in the dataset. While a diverse vocabulary can have a great impact on the model it eventually helps to create a robust system that can generalize well. Eventually, out of 31k comments ~15k comments were kept in the whole dataset. The dataset is publicly available. Here is the link to the dataset: SentNoB.

Class	Instances	#Sent/instance	#Words/instance
Negative	5,709 (36.3%)	1.64	16.33
Positive	6,410 (40.8%)	1.73	15.88
Neutral	3,609 (22.9%)	1.45	12.94
Total	15,728	1.63	15.37

Table 1: Brief statistics of the SentoNoB dataset per class label

3.1 Data Annotation

Three different annotators to label each instance with one of the five polarity labels Strong Negative, Moderate Negative, Neutral, Moderate Positive, and Strong Positive. For this task, we ten undergraduate students were employed and was provided with detailed annotation guidelines. Majority voting was used to assign the final class label, where we keep the neutral class unchanged but combine the two intensities of the polar classes and assign either Positive or Negative label. An inter-annotator agreement (Fleiss, 1971) of 0.53 was achieved which indicates a moderate agreement across the dataset. This is the highest such score among Bangla datasets.

[2] https://www.prothomalo.com/

Sentiment	Example in Bangla and English
Positive	[B] ভাই আপনার কথাই যাদু রয়েছে / [E] Brother, your words are magic.
Neutral	[B] ভাইয়া এই রেস্টুরেন্টা কোথায় / [E] Bro, where is this restaurant?
Negative	[B] এগুলো বড়লোক জানোয়ারের বাচ্চাদের কাজ / [E] These are the work of the children of the giant beast

Table 2: A Brief example of Bangla and its translation to English from the training dataset. Here, **B** represents Bangla and **E** represents English.

3.2 Dataset Statistics and Analysis

The dataset in total have 15, 728 instances. The average length of the instances is 1.63ś1.03 sentences and average sentence length is 15.37 ś 9.93 words. 40.8% of the data are labeled as Positive, 36.3% Negative, and 22.9% Neutral. Figure 1 provides the topic distribution of the dataset.

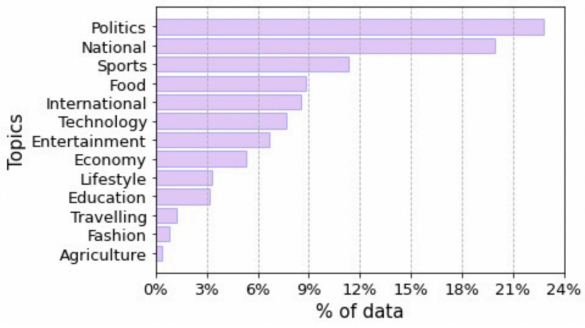

Figure 1: Topic distribution of dataset

4 Methodology

In this section, I describe the models used to perform classification on the SentNoB dataset. Initially, I trained linear SVM (Cortes and Vapnik, 1995) and Random Forest (Breiman, 2001) using hand-crafted linguistic features. Following this, I experimented with PLMs, given their recent success in downstream NLP tasks such as SA.

4.1 Linguistic Features

I experimented with different n-gram word (1-3) methods to extract features as well as their different combinations as these lexical methods have shown strong performance in classification tasks. Then I vectorized each instance using TF-Idf and Bag-of-words (BoW) to see the effect of those two embedding methods on the performance.

4.2 Pre-trained Language Model

In recent years large pre-trained language models like BERT (Devlin, 2018) have shown promising performance in NLP tasks. BERT is a transformer architecture (Vaswani, 2017) based model. Unlike recent language representation models, BERT is designed to learn deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, like text classification, named entity recognition, question answering, text summarization, machine translation etc.

4.3 BanglaBERT

BanglaBERT, introduced by (Bhattacharjee et al., 2021), is a BERT based Natural Language Understanding (NLU) model pretrained in Bangla, which achieves a outstanding performance in Bangla benchmark datasets.

BanglaBERT was pretrained using the ELECTRA (Clark, 2020), pretrained with the Replaced Token Detection (RTD) objective, where a generator and a discriminator model are trained jointly. The motivation behind using ELECTRA was computational efficiency. ELECTRA achieves comparable downstream performance to RoBERTa (Liu et al., 2019) or XLNet (Yang, 2019) with only a quarter of their training time.

BanglaBERT is available in multiple versions. BanglaBERT (small) and BanglaBERT (Large) has 13M and 335M parameters respectively.

4.3.1 Pretraining Data

Pretraining data for BanglaBERT was selected from 110 Bangla websites by their Amazon Alexa rankings and the volume and quality of extractable texts by inspecting each website. The contents included encyclopedias, news, blogs, e-books, stories, social media/forums, etc. The amount of data totaled around 35 GB. An important thing to note here is that, BanglaBERT is mostly trained on formal Bangla texts.

5 Experimental Setup

The experimental framework was implemented using Pytorch (Paszke et al., 2019), Scikit-learn (Pedregosa et al., 2011), and transformers (Wolf, 2019). As the baseline model I compared my results with the results from (Islam et al., 2021b) this paper. To remove noise from the dataset I used a normalizer (Hasan et al., 2020). The intend of this normalizer is to prepare the text by normalizing quotes, handling URLs, punctuation, emojis, and removing extra whitespace. This normalization pipeline has is directly useful to the performance of the model as without the normalizer the accuracy drops significantly. Due to the class imbalance, we perform per-topic stratified split to create training (80%), development (10%), and test (10%) sets.

For the purpose of training, validating and evaluating the model the dataset was split into 80-10-10 percentages. All the models were trained using UKKO and AKKA, machines provided by the CS department.

6 Evaluation metrics

I evaluated my model with macro averaged F-1 score. As the class distribution of dataset is imbalanced I used macro F-1 score which gives equal weight to each classes regardless of class imbalances. The evaluation metrics are described below:

- Macro F1-score: Macro F1-score is a metric used to evaluate a model's performance across multiple classes by calculating the F1-score for each class independently and then averaging them.

$$\textbf{Macro F}_1 = \frac{1}{C}\sum_{i=1}^{C} F_{1,i}$$

where:

- C is the total number of classes,
- $F_{1,i}$ is the F1-score for class i.

7 Results

Experimental results are provided in comparison with basline models in Table 3. Figure 2 represents the training and validation loss of the BanglaBERT large model

Model	Method	F-1
LinearSVM	Unigram(U)	63.19
LinearSVM	Bigram(B)	59.68
LinearSVM	Trigram(T)	55.56
LinearSVM	Char 2-gram (C2)	59.12
LinearSVM	Char 3-gram (C3)	62.6
LinearSVM	Char 4-gram (C4)	63.17
LinearSVM	Char 5-gram (C5)	63.57
Linear SVM	U + B + T + C2 + C3 + C4 + C5	**64.61**
mBERT	-	52.79
BERT base-uncased	-	60.5
BanglaBERT Small	-	**66.6**
BanglaBERT Large	-	**71.6**

Table 3: Comparison of BanglaBERT with baseline models from (Islam et al., 2021b) paper. Models above dotted line represents baseline models

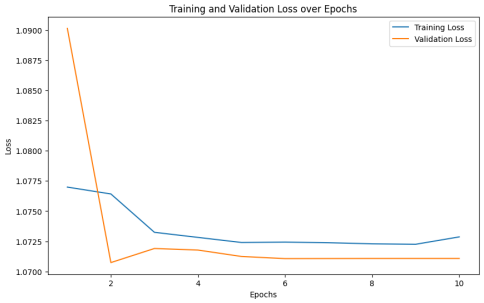

Figure 2: Training and Validation loss for BanglaBERT large model over epochs

8 Discussion

Baseline models were chosen from the paper (Islam et al., 2021b) where the SentNoB dataset was introduced. By thorough experimentation of different ML and neural networks the authors found that linguistic feature-based methods perform better then the neural networks and mBERT.

From Table 3 it can be seen that among word gram methods the unigram method performs better compared to bigram and trigram. Character n-gram (3-5) does not perform significantly better than unigram method. This can possibly imply that this task is highly dependent on word-level information rather than subword level.

However, multilingual BERT (mBERT) model performs lower than linguistic feature based methods. This can be possible due to two reasons: a) mBERTs training data is compiled of formal Bangla text from Wikipedia, whereas our dataset contains informal and noisy Bangla texts, and b) fine-tuning only the output layer makes mBERT under-trained for the task. Fine-tuning BanglaBERT on this dataset lifts the F-1 score by 3.08% and 10.82% respectively for the small and large version of the model.

Although BanglaBERT provides significant improvement it struggles to provide correct prediction of the sentiment polarity certain times. This can also account for the noisy and complex dialect of the SentNoB dataset. A large volume of the pretraining data for BanglaBERT comes from news, blogs, e-books, stories which contain formal and official Bangla texts which can be quite different than public comments in social media or news articles.

Sample from training dataset	Official forms of the sample
ভাল্লাগছেরে ভাই খুব্বব্বব	ভালো লাগছে রে ভাই খুবই
Bhagache's brother Very good	You are welcome brother very
বাংলাদেশ কুকুর দামআচে পুলিশ দাম নাই	বাংলাদেশে কুকুরের দাম আছে, পুলিশের দাম নেই
Bangladesh dog Price: Police price no	Dog price in Bengal Come on, Puleshar price That's it

Table 4: Random samples from the training set, including their official Bangla forms and English translations provided by Google Translate.

Table 4 presents examples of noisy, randomly selected samples from the training dataset. Due to the informal and unstructured nature of social media texts, individuals express opinions in diverse and unrestricted ways. This lack of standardization introduces significant variability among texts, making it challenging to effectively model and capture the complex representations inherent in social media comments.

Additionally, the quality of annotation can also have an effect in the performance of the model.

9 Conclusion

This paper aims to enhance the performance of sentiment analysis (SA) on the SentNoB dataset by comparing traditional lexicon-based machine learning (ML) methods with pretrained language models. Additionally, it examines the limitations and challenges faced by current models in accurately interpreting unique social media comments.
.

References

Ameen Abdullah Qaid Aqlan, B Manjula, and R Lakshman Naik. 2019. A study of sentiment analysis: Concepts, techniques, and challenges. In *Proceedings of International Conference on Computational Intelligence and Data Engineering*, pages 147–162. Springer Singapore, Singapore.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Uddin Ahmad, Kazi Samin, Md Saiful Islam, Anindya Iqbal, M Sohel Rahman, and Rifat Shahriyar. 2021. Banglabert: Language model pretraining and benchmarks for low-resource language understanding evaluation in bangla. *arXiv preprint arXiv:2101.00204*.

Leo Breiman. 2001. *Machine Learning*, 45(1):532.

Priyavrat Chauhan, Nonita Sharma, and Geeta Sikka. 2021. The emergence of social media data and sentiment analysis in election prediction. *Journal of Ambient Intelligence and Humanized Computing*, 12:2601–2627.

K Clark. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

A Conneau. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning*, 20(3):273297.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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

Tahmid Hasan, Abhik Bhattacharjee, Kazi Samin, Masum Hasan, Madhusudan Basak, M. Sohel Rahman, and Rifat Shahriyar. 2020. Not low-resource anymore: Aligner ensembling, batch filtering, and new datasets for Bengali-English machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2612–2623, Online. Association for Computational Linguistics.

C. Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1):216225.

Khondoker Ittehadul Islam, Sudipta Kar, Md Saiful Islam, and Mohammad Ruhul Amin. 2021a. 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.

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach.

Fang Luo, Cheng Li, and Zehui Cao. 2016. Affective-feature-based sentiment analysis using svm classifier. In *2016 IEEE 20th international conference on computer supported cooperative work in design (CSCWD)*, pages 276–281. IEEE.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Kumar Ravi and Vadlamani Ravi. 2015. A survey on opinion mining and sentiment analysis: tasks, approaches and applications. *Knowledge-based systems*, 89:14–46.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

T Wolf. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual bert? *arXiv preprint arXiv:2005.09093*.

Ashima Yadav and Dinesh Kumar Vishwakarma. 2020. Sentiment analysis using deep learning architectures: a review. *Artificial Intelligence Review*, 53(6):4335–4385.

Zhilin Yang. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.

A Appendix

Pretraining Data Sources

The following are some websites that were used as pretraining data. These lists are not exhaustive.

Encyclopedias

- bn.banglapedia.org
- bn.wikipedia.org
- songramernotebook.com

News

- anandabazar.com
- bangla.dhakatribune.com
- bbc.com
- banglatribune.com
- channelionline.com
- ctgtimes.com
- daily-bangladesh.com
- dainikdinkal.net
- dainikshiksha.com
- dw.com
- ittefaq.com.bd
- jagonews24.com
- kalerkantho.com
- manobkantha.com.bd
- prothomalo.com
- samakal.com
- tbsnews.net

Blogs

- amrabondhu.com
- bigganblog.org
- cadetcollegeblog.com
- muktangon.blog

Social Media/Forums

- banglacricket.com
- helpfulhub.com
- pchelplinebd.com
- techtunes.io