**Abstract**

The analysis of consumer sentiment, as expressed through reviews, can provide a wealth of insight regarding the quality of a product. While the study of sentiment analysis has been widely explored in many popular languages, relatively less attention has been given to the Bangla language, mostly due to a lack of relevant data and cross-domain adaptability. To address this limitation, we present BANGLA BOOK, a large-scale dataset of Bangla book reviews consisting of 158,065 samples classified into three broad categories: positive, negative, and neutral. We provide a detailed statistical analysis of the dataset and employ a range of machine learning models to establish baselines including SVM, LSTM, and Bangla-BERT. Our findings demonstrate a substantial performance advantage of pre-trained models over models that rely on manually crafted features, emphasizing the necessity for additional training resources in this domain. Additionally, we conduct an in-depth error analysis by examining sentiment unigrams, which may provide insight into common classification errors in under-resourced languages like Bangla. Our codes and data are publicly available at [https://github.com/mohsinulkabir14/BanglaBook](https://github.com/mohsinulkabir14/BanglaBook).

**1 Introduction**

The resources publicly available for scholarly investigation in the realm of Sentiment Analysis (SA) for the Bangla language are scarce and limited in quantity (Khatun and Rabeya, 2022; Sazzed, 2021; Rahman et al., 2019) despite its literary gravitas as the 6th most spoken language in the world with approximately 200 million speakers. In the existing literature on Bangla Text SA, as shown in Table 5, the largest dataset consists of 20,468 samples (Islam et al., 2022) while the smallest has a mere 1,050 samples (Tabassum and Khan, 2019). Besides these, Islam et al. (2020) created a dataset consisting of 17,852 samples and Islam et al. (2021) utilized a dataset of 15,728 samples. All other datasets apart from these either have <15,000 samples or are publicly unavailable. Another limitation of the existing research works in Bangla Text SA is the deficiency of datasets having product-specific review samples. Most of the available Bangla SA datasets are focused on user-generated textual content from cyberspace. The insights derived from these may not accurately represent sentiment in the context of product reviews, thus hindering their usefulness for businesses. The tonal and linguistic analysis of reviews from product-specific datasets can aid businesses to gain valuable insights into customer attitudes, preferences, and experiences which can then be leveraged to improve products and services, design targeted marketing campaigns, and make more informed business decisions. In this paper, we introduce a large-scale dataset, BANGLA BOOK, consisting of 158,065 samples of book reviews collected from online bookshops written in the Bangla language. This is the largest dataset for Bangla sentiment analysis to the best of our knowledge. We perform an analysis of the dataset’s statistical characteristics, employ various ML techniques to establish a performance benchmark for validating the dataset, and also conduct a thorough evaluation of the classification errors.

**2 Dataset Construction**

In order to create this dataset, we collect a total of 204,659 book reviews from two online bookshops (Rokomari² and Wafilife³) using a web scraper developed with several Python libraries, including BeautifulSoup, Selenium, Pandas, Openpyxl, and others.
and Webdriver, to collect and process the raw data.

For the data collection and preparation process of the BANGLABOOK dataset, we first compile a list of URLs for authors from online bookstores. From there, we procure URLs for the books. We meticulously scrape information such as book titles, author names, book categories, review texts, reviewer names, review dates, and ratings by utilizing these book URLs.

<table>
<thead>
<tr>
<th>Source</th>
<th>Language</th>
<th>Annotated</th>
<th>Unannotated</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rokomari</td>
<td>Bangla† + English</td>
<td>56,485</td>
<td>15,408</td>
<td>71,893</td>
</tr>
<tr>
<td></td>
<td>Bangla + Bangla† + English</td>
<td>13,144</td>
<td>4,114</td>
<td>17,258</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>154,301</td>
<td>44,328</td>
<td>198,629</td>
</tr>
<tr>
<td>Wafiifu</td>
<td>Bangla† + English</td>
<td>4,699</td>
<td>59</td>
<td>4,758</td>
</tr>
<tr>
<td></td>
<td>Bangla + Bangla† + English</td>
<td>896</td>
<td>3</td>
<td>899</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>5,395</td>
<td>65</td>
<td>6,050</td>
</tr>
<tr>
<td>Bangla</td>
<td>89,371</td>
<td>24,865</td>
<td>114,237</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Bangla† + English</td>
<td>56,855</td>
<td>15,411</td>
<td>72,266</td>
</tr>
<tr>
<td></td>
<td>Bangla + Bangla† + English</td>
<td>14,040</td>
<td>4,117</td>
<td>18,157</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>160,266</td>
<td>44,393</td>
<td>204,659</td>
</tr>
<tr>
<td>Untranslated Data (Removed)</td>
<td>2,201</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Final Dataset Size</td>
<td>158,065</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Summary statistics of our dataset. Bangla† denotes Romanized Bangla text.

<table>
<thead>
<tr>
<th>General Properties</th>
<th># of Reviews</th>
<th># of Books</th>
<th># of Reviewers</th>
<th># of Categories</th>
<th>Total Review Words</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>158,065</td>
<td>30,253</td>
<td>44,616</td>
<td>1,573</td>
<td>44,429,201</td>
</tr>
</tbody>
</table>

Table 2: General overview of BANGLABOOK.

2.1 Labeling & Translation

If a review does not have a rating, we deem it unannotated. Reviews with a rating of 1 or 2 are classified as negative, a rating of 3 is considered neutral, and a rating of 4 or 5 is classified as positive. Two manual experiments are carried out to validate the use of ratings as a measure of sentiment in product reviews. In the first experiment, around 10% of the reviews are randomly selected and annotated manually. The annotated labels are cross-checked with the original labels, resulting in a 96.7% accuracy in the corresponding labels. In addition, we consult the work of Wang et al. (2020) that explored the issue of incongruous sentiment expressions with regard to ratings. Specifically, the study scrutinized two categories of reviews: high ratings lacking a positive sentiment, and low ratings lacking a negative sentiment. We perform an analysis to identify such inconsistencies within our dataset and discovered that only a minuscule 3.41% of the samples exhibited this pattern. This figure is relatively insignificant when considering the substantially large scale of our dataset.

![Figure 1: Class Distribution of BANGLABOOK.](image)

After discarding the unannotated reviews, we curate a final dataset of 158,065 annotated reviews. Of these, 89,371 are written entirely in Bangla. The remaining 68,694 reviews were written in Romanized Bangla, English, or a mix of languages. They are translated into Bangla with Google Translator and a custom Python program using the googletrans library. The translations are subsequently subjected to manual review and scrutiny to confirm their accuracy. The majority of inaccurate translations primarily comprise spelling errors and instances where English words remain untranslated within samples containing a combination of Bangla and English text. The meticulous evaluation process of untranslated samples involves a thorough assessment by post-graduate native Bangla speakers, who critically compare the translated text against the original untranslated text to ascertain the correctness of the translation.

3 Statistical Analysis

Tables 1 and 2 provide an overview of the statistical properties of the BANGLABOOK dataset. The sentiment doughnut chart in Figure-1a illustrates the proportion of positive, neutral, and negative reviews, while the rating doughnut chart in Figure-1b displays the percentage of reviews that corre-
Upon analyzing the sentiment chart, it appears that the majority of the reviews (124,084 + 17,503 = 141,587 samples) are positive, with a significant portion also being negative (2,728 + 6,946 = 9,674 samples). A relatively small fraction of the reviews are neutral (6,804 samples). This suggests that overall, the books have been well received by the readers, with the majority expressing favorable opinions. The distribution of the dataset is representative of real-world scenarios and it tests itself well with previous content analysis works on book reviews (Lin et al., 2005; Sorensen and Rasmussen, 2004). In Figure 2, we can visualize an illustration of the sentiment distribution among the 5 most frequently reviewed categories of books. We can gain some salient insights from the popularity of these genres. Contemporary novels are bestsellers as they reflect current events, social issues, and trends, making them relatable and thought-provoking for the readers while self-help and religious books provide guidance, inspiration, and a sense of purpose, catering to individuals’ quest for personal growth and spiritual fulfillment.

4 Developing Benchmark for BANGLABOOK

A series of baseline models with combinations of different lexical and semantic features are chosen to evaluate the BANGLABOOK dataset. An overview of the models, evaluation metrics, results, and analysis of the experimental results are provided in this section.

4.1 Baseline Models & Features

For the lexical features, we extract bag-of-words (BoW), char n-grams (1-3), and word n-grams (1-3) from the reviews as these representations have performed well in different classification tasks (Islam et al., 2022). After extracting the features, they are vectorized using TF-IDF and count vectorizer and trained on a series of ML models such as Random Forest (Breiman, 2001), XGBoost (Chen and Guestrin, 2016), linear SVM (Cortes and Vapnik, 1995), Logistic Regression (le Cessie and van Houwelingen, 1992) and Multinomial Naïve Bayes (John and Langley, 1995). We choose LSTM (Hochreiter and Schmidhuber, 1997) with GloVe (Pennington et al., 2014) embedding for its ability to understand context along with recent dependency. We also fine-tuned two available transformer-based models in Bangla: Bangla-BERT (base-uncased) (110M parameters) (Sarker, 2020) and Bangla-BERT (large) (2.5B parameters) (Bhattacharjee et al., 2022), due to the recent success of BERT (Devlin et al., 2019) in various downstream NLP tasks. We select F1-score and weighted average F1-score to evaluate the models because the dataset has an un-
Table 4: Most frequent word unigrams conveying the strongest sentiments of each class with English translation. The colors respectively denote Positive, Neutral and Negative sentiments.

<table>
<thead>
<tr>
<th>Class</th>
<th>10 Most Frequent Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>ভােলা (good), সুন্দর (nice), অসাধারণ (extraordinary), দুদর্ন্ত (splendid), সেরা (best), সহজ (facile), চমৎকার (beautiful), আশা (hope), মনন (gratefulness), আলমুদ্যান (gratitude)</td>
</tr>
<tr>
<td>Neutral</td>
<td>ভােলা (good), সুন্দর (nice), ধারাপ (bad), মেটিমুট (average), কম (low), দুর্নীত (messy), তুল (fault), আশা (hope), সহজ (facile), অসাধারণ (extraordinary)</td>
</tr>
<tr>
<td>Negative</td>
<td>ভােলা (good), পাজে (trash), ধারাপ (bad), কম (low), সুন্দর (nice), আশা (hope), নষ্ট (waste), আশা (hope), নষ্ট (waste)</td>
</tr>
</tbody>
</table>

Figure 3: Confusion Matrix for Bangla-BERT

4.2 Results & Findings

Table 3 summarizes the experimental results for BANGLABOOK. Results show that Bangla-BERT (large) outperforms all other models by a clear margin. Also, the combination of word/char 2-gram and word/char 3-gram perform exceptionally well with respective classifiers. Our hypothesis is that these two features result in a large number of unique word and character combinations, aiding the models’ ability to generalize effectively across categories. Islam et al. (2022, 2021); Taher et al. (2018) concur on the same verdict by implying that the task predominantly relies on word units, with minimal dependence on subword level information and the nature of the Bangla language itself. Furthermore, Majumder et al. (2002) outlined the suitability of n-gram approaches in generating language profiles for Indo-European languages in their work. The bag-of-words (BoW) feature is inept at classifying the corresponding categories because of its inability to capture critical contextual information and nuance (Zheng and Casari, 2018). Although word 1-gram does not outperform word 2-gram and word 3-gram, it does predict the ‘Neutral’ class well. Both the pre-trained Bangla-BERT models perform fairly consistently across all categories on the BANGLABOOK dataset, demonstrating the usefulness of contextual understanding and transfer learning in classification tasks even in low-resource languages like Bangla. The LSTM model with GloVe embedding recognizes the ‘Negative’ and ‘Positive’ classes very marginally and fails completely to identify the ‘Neutral’ category. It is also notable that, SVM with bigram and trigram achieves perfect scores in the ‘Positive’ class.

To summarize, the utilization of pre-trained models (i.e. Bangla-BERT) that undergo training on extensive corpora, leading to exposure to extensive general language knowledge, has significantly contributed to their superior classification performance compared to other models and word embeddings. Additionally, models trained on hand-crafted features also perform significantly well. It should be noted that Bangla pre-trained models are currently undergoing development, and further training on expansive corpora has the potential to enhance their ability to generalize and achieve even more impressive results.

4.3 Error Analysis

In the ‘Positive’ class, all the models produce excellent classification results. While some models perform reasonably well on the ‘Negative’ class, nearly all of the models perform poorly on the ‘Neutral’ class. The class imbalance of the dataset, as shown in Figure 1, is one obvious cause of this fluctuation in results. The confusion matrix for
Bangla-BERT on our dataset, presented in Figure-3, reveals that most of the ‘Negative’ and ‘Neutral’ samples are misclassified as ‘Positive’ samples by our classifiers. To further analyze the misclassifications, we examine the W1 (word unigrams) of these three classes. We find 124,796 unique W1 for the ‘Positive’ class, 77,57% of the W1 from the ‘Neutral’ class, and 79.83% of the W1 from the ‘Negative’ class are found in the ‘Positive’ class. Table 4 depicts the most frequent W1 conveying the strongest sentiments for each class. With only one distinct ‘Neutral’ W1 and even the ‘Negative’ class having multiple positive W1, the dominance of ‘Positive’ sentiment W1 over the other two classes is evident. This may have contributed to the lack of distinctive words in the ‘Negative’ and ‘Neutral’ classes, which inevitably prevented the feature-based models from generalizing.

5 Morphology and Negation Patterns of Bangla

Understanding the morphology and negation patterns of a language holds paramount importance in the realm of sentiment analysis because negation can alter the meaning of words and phrases, thereby affecting the overall sentiment conveyed by a text. We provide a concise yet insightful recapitulation of the topic in the case of Bangla accompanied by review samples from our dataset BANGLA BOOK as the respective examples. From the linguistic typological standpoint, Bangla is categorized as a subject-object-verb (SOV) language because the subject, object, and verb generally adhere to said order in its sentential structure (Ramchand, 2004). The most common juxtaposition of polarity from positive to negative is the use of ni (নি) as a tensed negative. For example,

মাসি তার অনুরাগী হওয়ায় আমি এই বইটি কোনা
 থেকে নিজেকে গ্রাহিতের কথা পারিনি !!!!
Translation: As I am a fan of his I couldn’t resist myself from buying this book!!!!

Another negational feature is expressed by placing na (না) prior to the non-finite verb and after the finite verb in a sentence (although there are some exceptions). For example,

আর হুমায়ুন আহেমদ লিখেছেন এই বইটির
 উপর তিনি নিজেও পড়েনি তা।
Translation: Of course, Humayun Ahmed wrote that he himself is not satisfied with this book.

The Bangla language consists of no negative adverbs or pronouns (Thompson, 2006). This is why the negative element responsible for the reversal of polarity transcends from the word-level to the sentence-level rendering the occurrences of almost all negations in Bangla manifest on the syntactic level (Thompson, 2006).

In the cases of double negatives, we see the involvement of lexical negation, a morphological feature that works with negative affixes (prefixes and suffixes) attached to a root word. The prefixes in Bangla have two different phonetic variations or allophones depending on whether the prefix precedes a vowel or a consonant. The same is true for prefixes that imbue a negative connotation to a root word, e.g. o (অ) and on (অন) For example,

বিষ্ণু এই বইটি এই অপূণর্তা
 থেকে ফেলেছে।
Translation: But this book has covered up this incompleteness.

এমন তারিখের কাজে কিছু বই অন্যতম লেখকের বই,
 যাদের কোন কথা দেই।
Translation: In the words of Omar Khayyam, some books are books of never-ending youth, which have no decay.

Another negative prefix that precedes a root word to invert its polarity is nir (নির). For example,

লেখকের নিক্ষিপ্ত রাম লেখায় ফুটে উঠেছে।
Translation: The relentless effort of the author is reflected in the writing.

On the contrary, the suffix hin (হিন) succeeds a root word to convert it to the corresponding negative form. For example,

এককম বিভিন্ন সমাপ্তিক কথা নিয়ন্ত্রণ না পড়িতে
 তালো।
Translation: It is better for children not to read such baseless fictional stories.

The expression of negative sentiment is, therefore, very nuanced in the Bangla language as every occurrence of negative is intertwined with features like the tense, hierarchy of syntax, verb status, case-specific issues, and sequential arrangement of words (Thompson, 2006).

6 Conclusion

This paper introduces BANGLA BOOK, the largest Bangla book review dataset with 158,065 samples, each labeled with 1 of 3 user sentiments. We provide extensive statistical analysis and strong baselines facilitating the utility of the dataset. Given its massive size and fine-grained sentiment distribution, BANGLA BOOK has the potential to alleviate the resource scarcity in Bangla language research.
7 Limitations

Many of the reviews that were gathered for constructing BANGLABOOK are discarded because they lack a corresponding rating. A manual annotation process would have yielded a much larger dataset, which was not feasible due to resource constraints. Moreover, one of the challenges for validating the dataset is the lack of statistical models and word-embeddings pre-trained on the Bangla language. Some pre-trained Bangla-BERT models, yet to be trained on extensive corpora, have only recently been proposed. Improving transformer-based models for Bangla can enhance sub-word level contextual understanding which will consequently help in more accurate identification of the sentiments in BANGLABOOK (Islam et al., 2022).

References


Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.


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. IEEE.


Shamsul Araf in Mahtab, Nazmul Islam, and Md Mahmuzdur Rahman. 2018. Sentiment analysis on bangladesh cricket with support vector machine. In 2018 international conference on Bangla speech and language processing (ICBSLP), pages 1–4. IEEE.


A Appendix
<table>
<thead>
<tr>
<th>Dataset</th>
<th>Sentiment Classification</th>
<th>Sentiment Distribution</th>
<th>Total # of Samples</th>
<th>Availability</th>
<th>Type of Content</th>
<th>Source(s)</th>
<th>Baseline Models</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Tabassum and Khan, 2019)</td>
<td>Positive</td>
<td>-</td>
<td>1,050</td>
<td>Closed</td>
<td>Posts, comments, reviews</td>
<td>Facebook</td>
<td>RF</td>
</tr>
<tr>
<td>(Chowdhury and Chowdhury, 2014)</td>
<td>Positive</td>
<td>-</td>
<td>1,300</td>
<td>Closed</td>
<td>Posts, comments, reviews</td>
<td>Twitter</td>
<td>-</td>
</tr>
<tr>
<td>(Nabi et al., 2016)</td>
<td>Positive</td>
<td>-</td>
<td>1,500</td>
<td>Closed</td>
<td>Posts, comments, reviews</td>
<td>Social Media</td>
<td>-</td>
</tr>
<tr>
<td>(Mahbub et al., 2018)</td>
<td>Positive</td>
<td>-</td>
<td>1,601</td>
<td>Closed</td>
<td>Comments, reviews, reviews</td>
<td>Predhom Alo</td>
<td>SVM, DT, NB</td>
</tr>
<tr>
<td>(Akter and Aziz, 2016)</td>
<td>Positive, Neutral, Negative</td>
<td>-</td>
<td>3,600</td>
<td>Closed</td>
<td>Posts, comments, reviews</td>
<td>Facebook</td>
<td>NB</td>
</tr>
<tr>
<td>(Rahman and Kumar Dey, 2018)</td>
<td>Positive</td>
<td>-</td>
<td>4,700</td>
<td>Open</td>
<td>Comments, reviews, reviews</td>
<td>Facebook pages, SVM, LR, KNN, DT, DT, LSTM Prothom Alo, NB, CNN Facebook, Twitter, YouTube, News Portal</td>
<td></td>
</tr>
<tr>
<td>(Dey and Sarker, 2019)</td>
<td>Positive</td>
<td>2,600</td>
<td>5,200</td>
<td>Closed</td>
<td>Comments, reviews, reviews</td>
<td>YouTube</td>
<td>NB, XGB</td>
</tr>
<tr>
<td>(Khanam and Rabeya, 2022)</td>
<td>Positive, Negative</td>
<td>-</td>
<td>5,500</td>
<td>Closed</td>
<td>Comments, reviews, reviews</td>
<td>Social Media</td>
<td>-</td>
</tr>
<tr>
<td>EMOIC (Iqbal et al., 2022)</td>
<td>Anger</td>
<td>-</td>
<td>7,000</td>
<td>Open</td>
<td>Posts, comments, reviews</td>
<td>Facebook, YouTube, Online blogs, Bangla story books, novels, newspapers, discourse</td>
<td></td>
</tr>
<tr>
<td>(Tahsin et al., 2019)</td>
<td>Happy, Tense, Excited, Sad, Angry, Scared</td>
<td>-</td>
<td>7,500</td>
<td>Closed</td>
<td>-</td>
<td>-</td>
<td>NB, Tropical Method</td>
</tr>
<tr>
<td>(Akter et al., 2021)</td>
<td>Positive</td>
<td>-</td>
<td>7,905</td>
<td>Closed</td>
<td>Product reviews</td>
<td>Daraz</td>
<td>RF, LR, SVM, KNN, XGB</td>
</tr>
<tr>
<td>(Hasan et al., 2016)</td>
<td>Positive</td>
<td>-</td>
<td>9,337</td>
<td>Closed</td>
<td>Comments, reviews, reviews</td>
<td>Facebook, Twitter, YouTube, News Portal</td>
<td>LSTM</td>
</tr>
<tr>
<td>(Rahb et al., 2022)</td>
<td>Impartial, Curious, Gratitude</td>
<td>3,800</td>
<td>10,581</td>
<td>Closed</td>
<td>Comments, reviews, reviews</td>
<td>Social Media</td>
<td>SVM, RF, CNN, LSTM</td>
</tr>
<tr>
<td>(Al Kaiser et al., 2021)</td>
<td>Positive</td>
<td>-</td>
<td>1,909</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Sazzed, 2020a)</td>
<td>Positive, Negative</td>
<td>-</td>
<td>11,006</td>
<td>Closed</td>
<td>Comments, reviews, reviews</td>
<td>Facebook</td>
<td></td>
</tr>
<tr>
<td>(Sazzed, 2020b)</td>
<td>Positive</td>
<td>-</td>
<td>12,000</td>
<td>Closed</td>
<td>Comments, reviews, reviews</td>
<td>YouTube</td>
<td></td>
</tr>
<tr>
<td>(Islam et al., 2021)</td>
<td>Positive</td>
<td>4,959</td>
<td>17,852</td>
<td>Open</td>
<td>Comments, reviews, reviews</td>
<td>Predhom Alo</td>
<td>CNN, LSTM, BERT, GRU, fastText</td>
</tr>
<tr>
<td>(Islam et al., 2022)</td>
<td>Positive</td>
<td>4,202</td>
<td>10,620</td>
<td>Closed</td>
<td>Comments, reviews, reviews</td>
<td>Predhom Alo</td>
<td>CNN, LSTM, BERT, GRU, fastText</td>
</tr>
<tr>
<td>(EmoNoBa (Islam et al., 2022)</td>
<td>Product</td>
<td>-</td>
<td>141,587</td>
<td>Closed</td>
<td>Book reviews, reviews</td>
<td>Rokomari, Watlife</td>
<td>RF, LSTM, LR, GRU, MNB, SVM, XGB Bangla-BERT</td>
</tr>
</tbody>
</table>

Table 5: Comparison of notable Bangla Sentiment Analysis datasets sorted in ascending order of size. The abbreviations respectively denote, Random Forest (RF), Support Vector Machine (SVM), Decision Tree (DT), Naïve Bayes (NB), Logistic Regression (LR), K-Nearest Neighbors (KNN), Long Short-term Memory (LSTM), Convolutional Neural Network (CNN), Extreme Gradient Boost (XGB), Multinomial Naïve Bayes (MNB), Radial Basis Function (RBF), Extreme Random Tree (ET), Recurrent Neural Network (RNN), Bidirectional Encoder Representations from Transformers (BERT), Gated Recurrent Unit (GRU), Bi-LSTM (Bidirectional LSTM). All the publicly available datasets are hyperlinked. Open denotes the redaction of the link for anonymity.
ACL 2023 Responsible NLP Checklist

A  For every submission:

☑ A1. Did you describe the limitations of your work?
  Left blank.

☐ A2. Did you discuss any potential risks of your work?
  Not applicable. Left blank.

☑ A3. Do the abstract and introduction summarize the paper’s main claims?
  Left blank.

☒ A4. Have you used AI writing assistants when working on this paper?
  Left blank.

B  ☒ Did you use or create scientific artifacts?
  Left blank.

☐ B1. Did you cite the creators of artifacts you used?
  Not applicable. Left blank.

☐ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
  Not applicable. Left blank.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
  Not applicable. Left blank.

☐ B4. Did you discuss the steps taken to check whether the data that was collected/used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect/anonymize it?
  Not applicable. Left blank.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
  Not applicable. Left blank.

☑ B6. Did you report relevant statistics like the number of examples, details of train/test/dev splits, etc. for the data that you used/created? Even for commonly-used benchmark datasets, include the number of examples in train/validation/test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
  Left blank.

C  ☑ Did you run computational experiments?
  Left blank.

☒ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
  Left blank.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.
C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
   *Left blank.*

C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
   *Left blank.*

C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
   *Left blank.*

D. Did you use human annotators (e.g., crowdworkers) or research with human participants?
   *Left blank.*

D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
   *Not applicable. Left blank.*

D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants’ demographic (e.g., country of residence)?
   *Not applicable. Left blank.*

D3. Did you discuss whether and how consent was obtained from people whose data you’re using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
   *Not applicable. Left blank.*

D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
   *Not applicable. Left blank.*

D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
   *Not applicable. Left blank.*