ISLAMIC UNIVERSITY OF TECHNOLOGY (IUT)

Entity Level Sentiment Analysis from Online Bangla Reviews

By

Obayed Bin Mahfuz (161041013)

A thesis submitted in partial fulfilment of the requirements
for the degree of M.Sc. in Computer Science and Engineering

Academic Year: 2023-2024

Department of Computer Science and Engineering (CSE)
Islamic University of Technology (IUT).
A Subsidiary Organ of the Organization of Islamic Cooperation (OIC).
Dhaka, Bangladesh.

January 15, 2024

Declaration of Authorship

I, Obayed Bin Mahfuz, declare that this thesis titled, 'Entity Level Sentiment Analysis from Online Bangla Reviews' and the work presented in it is my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Any part of this thesis has not been submitted for any other degree or qualification at this University or any other institution.
- Where I have consulted the published work of others, this is always clearly attributed.

Submitted By:

_____

(Signature of the Candidate)
Obayed Bin Mahfuz- 161041013
January 2024

Entity Level Sentiment Analysis from Online Bangla Reviews

Approved By:

Dr. Hasan Mahmud
Thesis Supervisor,
Associate Professor,
Department of Computer Science and Engineering (CSE),
Islamic University of Technology (IUT).

Dr. Md. Kamrul Hasan
Professor,
Department of Computer Science and Engineering (CSE),
Islamic University of Technology (IUT).

Dr. Md. Hasanul Kabir
Head of the Department and Professor,
Department of Computer Science and Engineering (CSE),
Islamic University of Technology (IUT).

Dr. Mohammad Rezwanul Huq
Associate Professor
Department of Computer Science and Engineering (CSE),
East West University, Dhaka, Bangladesh.

Abstract

Extracting sentiment orientation from texts is known as sentiment analysis or opinion mining. The evaluation of consumer sentiment through reviews offers valuable insights into product quality. Entity Level Sentiment Analysis (ELSA) works on the specific entity of a product e.g. electronic accessories, clothing, food, fashion items, groceries, sports accessories, etc. Thus, these analyses can infer more specific information related to a product like marketing strategy development, product quality estimation, and service evaluation, etc. Sentiment analysis has been extensively studied in popular languages like English, Arabic, French, Chinese, etc. However, the Bangla language, which ranks as the sixth most widely spoken language globally, has received relatively less attention in this area. This limited focus can be attributed to the scarcity of relevant data and challenges related to cross-domain adaptability, resulting in a small number of works available for Bangla sentiment analysis. Entity Level Sentiment Analysis (ELSA) is the sentiment extracted from a specific entity of the text. To date, no studies have been conducted on ELSA in Bangla text. To address this gap, we present an entity-level sentiment analysis conducted on a dataset of 10,000 reviews consisting of book reviews, women's clothing reviews, and health product reviews in Bangla text. The dataset comprises 300 book reviews collected from online bookshops namely Rokomari and Wafilife and 9700 samples of women's clothing and health product reviews collected from the Daraz online e-commerce site. The reviews are manually annotated for entity identification with their corresponding sentiment by 2 native annotators. The sentiment is categorized into three main groups: positive, negative, and neutral. The inter-rater reliability (IRR) between the annotators is performed using Cohen Kappa's score. To establish baselines we used pre-trained language models. For product entity identification, we used mbert-bengali-ner language model and for sentiment analysis, we used the bangla-bert-base language model. The results of our proposed methodology for Named Entity Recognition (NER) and Sentiment Analysis (SA) are promising. The F1-score of NER is 87.91% and SA is 86.74% respectively. Our data collection web crawler code[1], constructed data[2] and baseline analysis code[3] are publicly available.

Keyword - Sentiment Analysis; Entity Level Sentiment; Named Entity Recognition; Large Language Model; BERT

[1] https://github.com/obayedsiam/WebScrappingPython.git

[2] https://drive.google.com/drive/folders/11re2e4buv2jnuJuPlehCcI_81EBDDVE3?usp=drive_link

[3] https://colab.research.google.com/drive/1grp1YHyv5eZ2Y5uIf-JH8zu2oU072xlz?usp=sharing

Acknowledgements

I would like to express my whole-hearted gratitude to my Lord Allah Subhanu Wata'ala for giving me the strength and divine guidance to complete this study, and for His ever-present support when no one else was by my side.

I extend my heartfelt appreciation to my dedicated supervisor, Dr. Hasan Mahmud, for his unwavering mentorship and invaluable support throughout this journey. I am profoundly grateful to Dr. Md. Kamrul Hasan for his consistent motivation, support, and insights that have been instrumental in the successful completion of this study.

Furthermore, I express my special thanks to Mohsinul Kabir, and Ferdous Hridoy for helping and guiding me throughout the thesis journey with study materials, guidelines, and suggestions.

Finally, I would like to thank my parents and my wife for their constant motivation and support.

This work is partially supported by

Systems and Software Lab (SSL)
Department of Computer Science and Engineering (CSE)
Islamic University of Technology (IUT)

Contents

List of Figures

List of Tables

Abbreviations

Abbreviation	Meaning
NLP	Natural Language Processing
SA	Sentiment Analysis
ELSA	Entity Level Sentiment Analysis
TSA	Targeted Sentiment Analysis
SLSA	Sentence Level Sentiment Analysis
DOCSA	Document Level Sentiment Analysis
ABSA	Aspect Based Sentiment Analysis
NER	Named Entity Recognition
LLM	Large Language Model
BERT	Bidirectional Encoder Representations from Transformers
MLM	Masked Language Model
NSP	Next Sentence Prediction
RoBERTa	Robustly Optimized BERT Pre-training Approach
GPT	Generative Pre-trained Transformers
GRU	Gated Recurrent Units
IRR	Inter Rater Reliability
RNN	Recurrent Neural Network
CNN	Convolutional Neural Network
LSTM	Long Short Term Memory
SVM	Support Vector Machine
RF	Random Forest
MNB	Multinomial Naive Bayes
XGBoost	Extreme Gradient Boosting
K-NN	K-Nearest Neighbor

LR	Linear Regression
LogR	Logistic Regression
DT	Decision Tree
BoW	Bag of Words
TF-IDF	Term Frequency - Inverse Document Frequency
POS	Parts of Speech
LDD	Lexicon Data Dictionary
VADER	Valence Aware Dictionary and Sentiment Reasoner
ROC	Receiver Operating Characteristic
AUC	Area Under the ROC Curve
TP	True Positive
FP	False Positive
FPR	False Positive Rate
TPR	True Positive Rate

Dedicated to my parents and siblings for their lifelong dedicated support to my education …

Chapter 1

Introduction

Sentiment analysis (SA) is the process of extracting subjective opinions like sentiment polarity from texts using natural language processing, text analysis, and computational linguistics [1, 2]. The origins of SA in written documents trace back to World War II, with a primary focus on political aspects. However, it gained significant research attention in the mid-2000s, leveraging Natural Language Processing (NLP) to extract subjective information from diverse online content sources [2]. SA is also termed sentiment mining, sentiment extraction, opinion mining, or opinion extraction [3]. When SA focuses on specific entities within a text then it is referred to as Entity Level Sentiment Analysis (ELSA) [4] or Targeted Sentiment Analysis (TSA) [5]. The number of online news portals, blogs, social media, and e-commerce sites is rapidly growing. Individuals express their opinions and provide reviews on products, news articles, and stories on various platforms. As a result, the virtual world has become overwhelmed with opinionated texts [6]. These texts hold significant value in the field of natural language processing (NLP), particularly in ELSA[7] . Additionally, they contribute to prompt decision-making in business, the development of robust marketing strategies, analysis of product quality, evaluation of services, examination of customer behavior, and much more [5]. Furthermore, ELSA proves advantageous from a customer's perspective too. Customers can obtain the collective viewpoint of previous customers on a specific entity, aiding their decision-making process and allowing them to compare products [5]. Thus, ELSA is highly beneficial for numerous purposes. This type of analysis effectively identifies the positive and negative aspects of an entity within a given text. The wealth of online product reviews serves as a pivotal resource for conducting such research.

The primary distinction between general SA and ELSA lies in ELSA's ability to specifically identify sentiments related to individual products. This feature is highly valuable

for explicitly pinpointing product sentiments, which general SA cannot achieve. From both a business and customer perspective, this capability is immensely beneficial.

For instance, consider a product review : "আমার শ্যাম্পুটি একদমই ভালো লাগেনি কিন্তু ডেলিভারি সময় মত পেয়েছি ।" (I did not like the shampoo at all but received it on time). In this review, there is a negative sentiment regarding the product but a positive one about the delivery. Now, a general SA would provide an overall neutral sentiment, while ELSA would correctly interpret it as a negative sentiment specifically related to the product. Thus, by using general SA, we would overlook the product's issues and lose the exact product review completely. Hence, if we apply general SA and try to make a sales and marketing strategy, we will not consider this product to be improved, potentially leading to incorrect product selection and marketing strategies.

Now, consider another example: "সাবান টি কিনেছি । ব্যবহার করার পরে বলতে পারবো কেমন হয়েছে । কিন্তু সেলারের ব্যবহার খুব খারাপ" (Bought the soap. Will provide a review after using it. But the seller behavior was very bad). Here, the product receives no review or a neutral one, while the seller's behavior is viewed negatively. Thus, general SA would attribute an overall negative sentiment to the review, potentially leading to an inaccurate evaluation of the product. In contrast, ELSA would provide a neutral sentiment, offering a more accurate representation. Now, in the context of business strategy, we would refrain from considering this product for immediate improvement. Instead, we would keep it in a queue for future decision-making based on further review, allowing for a more precise and informed business strategy. In this way, ELSA proves to be a valuable asset in shaping business strategies, making informed product selections, and guiding sales and marketing decisions effectively.

1.1 Problem Statement

While general SA has gained considerable attention, the exploration of ELSA remains notably underdeveloped. In particular, the Bangla language has yet to receive the comprehensive scrutiny it deserves within the realm of ELSA[7]. This observed lack of focus can be attributed to multiple factors, ranging from the scarcity of annotated data to the intricate challenges associated with cross-domain adaptability [8]. Remarkably, the domain of Bangla ELSA remains largely undiscovered, with very limited existing research to date. The absence of exhaustive research and effective methodologies for Bangla ELSA presents a series of formidable obstacles. Mainly, the lack of suitable datasets obstructs the refinement and assessment of robust ELSA models tailored for

Bangla text. Additionally, the complex language features and subtle context of the Bangla language create difficult problems. These complexities always affect how well we can figure out and thoroughly understand the feelings expressed about particular entities. Furthermore, the relatively unexplored landscape of Bangla ELSA curtails its potential for diverse applications across domains such as customer feedback analysis, market research, and product development. Addressing these challenges presents an exciting opportunity for future research to unlock the full spectrum of Bangla ELSA and its far-reaching implications.

Therefore, it is imperative to meticulously discern the limitations and challenges associated with the Bangla text ELSA through extensive research and experimentation. This study endeavors to confront several prominent challenges posed by ELSA by meticulously constructing annotated datasets focused on product entities and their associated sentiment orientations.

1.2 Motivation & Scopes

The driving force behind researching ELSA arises from the recognition of the significance of understanding sentiments toward specific entities in textual data. Traditional SA methods often overlook the subtle nuances associated with sentiments expressed towards individual entities such as products, services, or public figures. By focusing on ELSA, we aim to fill this gap and provide a more specific analysis that enables a deeper comprehension of how individuals perceive and respond to specific entities.

The following points best summarize the motivation of this research:

1. Firstly, opinionated online text reviews are increasing due to the increased number of online platforms. These texts hold significant information related to different entities. Researchers in the field of NLP can use these data to analyze the sentiments of these entities.

2. Secondly, ELSA is useful for individuals, businesses, and organizations. Furthermore, analyzing entity sentiment can provide valuable insights into customer opinions, market trends, brand perception, etc. This information can guide decision-making processes, improve products and services, and enhance customer satisfaction.

3. Finally, an annotated dataset and entity-specific sentiment analysis methodology can foster ELSA research and create more research scopes in the field of sentiment analysis.

1.3 Research Challenges

The exploration of ELSA for Bangla texts consists of several challenges. When we perform SA on a whole text, the sentiment labeling typically covers the entire content. However, in the case of ELSA, we not only have to label the overall sentiment but also annotate the specific entities and their associated sentiment orientation [7]. Unfortunately, there is a significant challenge in this research as there is a lack of high-quality, extensive, and annotated datasets that include both entities and sentiment annotations.

Moreover, the ever-changing context in which entities and their sentiments are discussed adds another layer of complexity to accurately evaluate ELSA. Machine learning algorithms, with their feature extraction methods, struggle to grasp the nuanced contextual meanings of words within the text [8]. This presents yet another significant challenge in our research. Furthermore, the ELSA framework stands apart from the typical SA framework because it deals with both entities and sentiment. However, the way this framework is applied in English [4] doesn't directly apply to Bangla due to the unique characteristics of the Bangla language. Consequently, finding the right approach for Bangla ELSA that also fits with the annotation process is another challenge that this research addresses. Hence, addressing these multifaceted challenges calls for innovative approaches that combine linguistic expertise with advanced data analysis techniques.

1.4 Research Contribution

By addressing the limitations of existing research, this study aims to develop a methodology for the advancement of ELSA in the Bangla language. Our contribution to this research can be divided into two parts. In the first part, we collected and annotated the dataset for Bangla ELSA. This annotation requires a named entity as well as sentiment annotation corresponding to the entity. The creation of a labeled dataset and the annotation will contribute to the development of accurate ELSA models for Bangla. In the second part, we created a pre-trained language model-based methodology to analyze the sentiments of the specific entity mentioned in the text.

Overall, this research aims to advance the field of ELSA in the realm of the Bangla language, providing valuable insights and tools for comprehending sentiment towards specific entities in Bangla textual data. The key contribution of this thesis can be outlined as follows :

- **Labeled dataset construction for Bangla ELSA:** Despite advancements in overall SA, there is a lack of research in Bangla ELSA. Furthermore, there are limited available Bangla-labeled datasets for ELSA. As we collected our review samples from online e-commerce products, our entity estimation is confined to product-type entities only.
- **A pre-trained language model-based methodology development for Bangla ELSA:** SA methodology designed for overall sentiment detection may not perform well in the context of entity-specific sentiment analysis. Bangla ELSA research can focus on specific domains such as e-commerce, social media, or customer reviews in Bangal text to provide more accurate and tailored sentiment analysis solutions on specific entities.

1.5 Thesis Outline

In Chapter 1 the objective of the study has been discussed concisely. Chapter 2 deals with the necessary background study & literature review for this study. In Chapter 3, the proposed methodology, data collection and annotation procedure, and the quality control mechanism have been discussed in detail. The summary statistic of the constructed dataset for this study is described in Chapter 4. The baseline classification model for this dataset and evaluation metrics are presented in Chapter 5. Chapter 6 discusses the classification results, potential sources of bias in the data, and the necessary aspects to consider while conducting additional research in this domain. Chapter 7 concludes the current study and discusses the limitations of this research and future directions. The final segment of this study contains all the references and credits used.

Chapter 2

Background Study

Online platforms have expanded significantly in recent years. These platforms have fostered widespread interaction among individuals who share their perspectives, post reviews on various products, and express opinions on current events and news. According to Statista as of October 2022, there were approximately 5.03 billion active Internet users worldwide, representing about 63% of the global population, with 93% of them actively using social media [2]. Thus these platforms are overwhelmed with opinionated texts. These texts hold information related to products, organizations, content, persons, etc. A wealth of information can be extracted by analyzing the sentiment orientation of these texts [9]. Furthermore, this information can help in business decision-making, strategy development, product quality improvement, market analysis, and product selection process of the customers [3, 10]. Therefore, delving into the realm of SA can prove valuable both from a business and a customer's standpoint, as different subdomains within sentiment analysis can be applied across various levels, each requiring distinct methodologies.

2.1 Overview of Sentiment Analysis

SA as a field, encompasses a diverse array of interrelated topics that collectively contribute to our understanding of this domain. An essential starting point in comprehending SA is to delve into these constituent topics. Among these, the discussion of the specific tasks that SA can effectively address stands out as a critical consideration. These tasks are notably influenced by the different levels of SA, which provide the foundational framework for their execution. Moreover, an exploration of the various approaches already undertaken for these specific tasks and related topics is paramount in this journey.

Understanding the current research scopes and the evolving landscape of SA research is integral in grasping the subject's nuances and its contemporary relevance. Alongside these key areas, this overview seeks to shed light on other interconnected themes that enrich the discourse around SA, contributing to a comprehensive understanding of the field and its ongoing research endeavors.

2.1.1 Tasks of Sentiment Analysis

SA can be utilized in several tasks. Within the last two decades, numerous researches have been conducted in this area of natural language processing. Among them, subjectivity classification, sentiment classification, opinion spam detection, implicit language detection, aspect extraction, cross-language classification, cross-domain classification, and polarity detection are the most important subfields of SA. [10]. The essential and most extensively researched fields of SA tasks are outlined below.

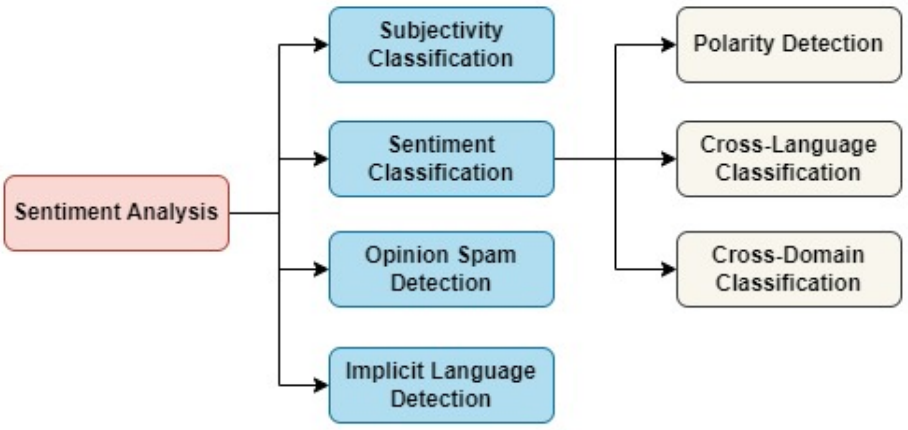

Figure 2.1: Sentiment Analysis Tasks

2.1.1.1 Subjectivity Classification :

Subjectivity classification plays an integral role within the realm of SA[11]. This is because not every text inherently possesses a sentiment orientation. For instance, the sentence, "I will go home tomorrow" lacks a clear sentiment orientation. Consequently, before embarking on any form of SA, conducting subjectivity classification can prove to be a valuable preparatory step, potentially leading to improved analytical outcomes [10]. Most of the research in subjectivity classification is in English. For instance, Karamibekr et al. [12] introduced an approach that takes into account the role of verbs in subjectivity analysis, particularly in social contexts. Additionally, they have developed an

architecture capable of performing subjectivity and sentiment classification tasks. Furthermore, Al Hamoud et al. [13] assess and compare six deep learning techniques for subjectivity classification. These methods encompass Long Short-Term Memory Networks (LSTM), Gated Recurrent Units (GRU), bidirectional GRU, bidirectional LSTM, LSTM with attention, and bidirectional LSTM with attention. Savinova et al. [14] treat subjectivity analysis as a regression task and evaluate the performance of a transformer RoBERTa model in assigning subjectivity scores to online news content, including news sourced from social media. In contrast, the Bangla language lacks research in subjectivity classification. In our research, we found most of the Bangla SA research is sentiment classification-based. However, Das et al. [15] demonstrates the effectiveness of a Conditional Random Field (CRF) based subjectivity detection approach. It has been tested on both English and Bengali corpora from various domains, highlighting its utility in a multi-domain context. Furthermore, a limited amount of research has been carried out in subjectivity analysis in Arabic [16, 17, 18], Hindi [19], Chinese [20] and Japanese [21, 22] languages.

2.1.1.2 Sentiment Classification :

Sentiment classification is one of the widely researched fields of SA. In this domain, the sentiment of texts is derived based on several criteria. The automatic assigning of the positive, negative, and neutral sentiment towards a text is mainly referred to as sentiment classification [23]. Furthermore, cross-language classification is another subdomain in sentiment classification. In this domain, the polarity of a text is determined by training the dataset of another language [10]. Cross-language classification is mostly used for under-resourced languages. Many non-English languages have unannotated dataset problems. These problems can be alleviated using cross-language classification. Different techniques for this task include machine translation, parallel corpora, bilingual sentiment lexicon, pre-trained language models based on cross-lingual word embedding, etc. [24]. Nonetheless, a significant challenge in this domain is the ambiguity surrounding word polarity. While research in sentiment classification spans multiple languages, the vast majority of studies are conducted in English, indicating a substantial imbalance. For a comprehensive discussion of sentiment classification, its methodologies, and associated challenges, please refer to section 2.3.

2.1.1.3 Opinion Spam Detection :

Another important task of SA includes opinion spam detection which is used to identify fake reviews, comments, and opinions [25]. Reviews and opinions hold substantial influence, particularly when it comes to consumer decisions on products, hotel booking, institution selection, restaurant selection, etc. Although the web is proliferated with reviews and opinions, fake texts and opinions are also found to manipulate customer decisions. In these scenarios, opinion spam detection can be used to identify fake and real reviews. In the English language, a substantial amount of research has been undertaken in this domain. Rastogi et al. [26] conducted a comprehensive analysis and categorization of existing literature on opinion spamming, focusing on three distinct detection targets: opinion spam, opinion spammers, and collusive opinion spammer groups. Additionally, the research classifies opinion spamming into three categories, taking into account textual and linguistic features, behavioral patterns, and relational aspects. Mukherjee et al. [27] introduced an unsupervised method for identifying opinion spam. The paper presents an innovative generative model designed to detect deception by leveraging both linguistic and behavioral cues that spammers leave behind. Mewada et al. [28] provides a comprehensive survey on opinion spam detection in English. The researchers have systematically categorized and classified methods for detecting spam reviews based on review features, reviewer characteristics, and characteristics of the spammers' groups.

In the Bangla language, there have been some endeavors in the realm of opinion spam detection. For instance, Amin et al. [29] have created a system for detecting spam emails in Bangla. They have additionally compiled datasets of Bangla spam emails to facilitate the training and testing of their system. This paper delves into the application of six supervised machine-learning techniques. The classification outcomes indicate that Random Forest (RF) exhibited the highest performance, achieving an accuracy rate of 93.60%. Furthermore, Uddin et al. [30] aims to detect Bengali spam SMS using both traditional Machine Learning algorithms, LSTM and GRU. The study compares the performance of all these algorithms to identify the most effective one. Notably, both LSTM and GRU achieved the highest testing accuracy rates. Islam et al. [31] employed the Multinomial Naive Bayes (MNB) classifier, a supervised machine learning algorithm with feature extraction, to detect spam in Bangla text at the sentence level. Their methodology identifies spam by considering the polarity of each sentence associated with it. Nonetheless, a substantial amount of research has been dedicated to spam and fake review detection in Hindi [32, 33] and Arabic [34, 35] language also.

2.1.1.4 Implicit Language Detection :

Sarcasm is characterized by the use of statements that convey a meaning or sentiment opposite to what is expressed. It is often employed to mock, irritate, or for humor. The inclusion of sarcasm within a document poses a challenge for sentiment analysis, as conventional methods struggle to identify and account for sarcastic expressions [2]. Sarcasm, irony, and humor are commonly categorized as forms of implicit language [10]. These kinds of languages are very difficult to classify even by humans. For example consider the sentence ”Oh, perfect timing! Another traffic jam, just what I needed!”. Here, ”perfect timing” and ”just what I needed” express positive sentiment. But the middle part gives the whole sentence a sarcastic meaning. For this reason, the overall sentiment of the sentence becomes negative. Furthermore, examining emoticons, laughter expressions, and the frequent use of punctuation marks represents more traditional methods for identifying implicit language [10].

Most of the research in this field is conducted in English. For example, Potamias et al. [36] introduced a novel transformer-based approach that utilizes the pre-trained RoBERTa model in conjunction with a recurrent convolutional neural network to address figurative language in social media. They compared their network across four distinct benchmark datasets. Furthermore, Misra et al. [37] presented a substantial and high-quality dataset consisting of news headlines sourced from both a sarcastic news website and a legitimate news website. After highlighting the distinctive features of the dataset, they provided a comparative analysis of its characteristics with existing benchmark datasets for sarcasm detection. Additionally, they employed a Hybrid Neural Network architecture to gain insights into the elements that define sarcasm in textual content. A study by Tan et al. [2] suggests a mutual benefit between sarcasm detection and sentiment classification, demonstrating a positive correlation between these two tasks. To leverage this correlation and enhance overall sentiment analysis, the paper introduces a multi-task learning framework based on deep neural networks. This novel approach surpasses existing methods, achieving a significant 3% improvement in performance with an impressive F1-score of 94%.

In Bangla also some research has been conducted on implicit language detection. For instance, Anan et al. [38] introduced a BERT-based system that attains a remarkable accuracy of 99.60%, surpassing the performance of traditional machine learning algorithms, which achieve only 89.93%. Furthermore, they have utilized a newly curated Bangla sarcasm dataset, BanglaSarc, designed explicitly for this study’s evaluation. This

dataset comprises recent instances of both sarcastic and non-sarcastic comments, primarily sourced from Facebook and YouTube comment sections. Ghosh et al. [39] in another research curated a Bengali irony detection dataset encompassing 1500 labeled Bengali tweets gathered from Twitter. Their research provides a comprehensive dataset description and presents findings based on various established machine learning algorithms, including Naïve Bayes, Support Vector Machine (SVM), K-Nearest Neighbor (K-NN), and RF. Lora et al. [40] introduces "Ben-Sarc", a substantial Bengali corpus for sarcasm detection, comprising 25,636 comments from public Facebook pages. The research outlines a comprehensive approach that leverages various models, including traditional machine learning, deep learning, and transfer learning, to identify sarcasm within Bengali text using the Ben-Sarc corpus. The study concludes with a comparative analysis of these models' performance on the Ben-Sarc dataset. In addition to the English and Bangla languages, the detection of irony and sarcasm has been extensively investigated in other languages, including Hindi [41], Arabic [42] and Japanese [43].

Our analysis of existing studies reveals that SA spans a wide spectrum of research areas. It is apparent from prior research that a significant portion of SA investigations across diverse domains has been primarily carried out in the English language. In contrast, subjectivity classification in most languages remains a relatively underexplored domain.

In the context of Bangla, the bulk of research efforts has been concentrated on Sentiment Classification. While there are a few notable studies in the domains of opinion spam detection and implicit language detection, these areas, along with subjectivity classification, hold significant potential for further research opportunities. It's important to note that SA research extends beyond English and includes languages such as Arabic, Hindi, Chinese, and Japanese, where various investigations have also been conducted.

2.1.2 Levels of Sentiment Analysis

Sentiment analysis can be applied at various granular levels, ranging from an entire paragraph down to individual words. The choice of the analysis level will depend on the specific requirements of the task. Nevertheless, the primary categories for sentiment analysis are typically categorized into broader levels, including document-level, sentence-level, entity-level, and aspect-level analysis, as outlined by Liu (2012)[11]. Our analysis of the prevailing trend in Sentiment Analysis reveals a historical progression. Initially, researchers emphasized classical machine learning approaches, transitioning to deep learning algorithms for enhanced contextual comprehension, driven by the quest for higher accuracy. In recent developments, SA research predominantly leans

towards leveraging pre-trained language models, reflecting a pursuit of improved accuracy and superior contextual understanding. However, a detailed discussion of different SA levels and their corresponding research is presented below.

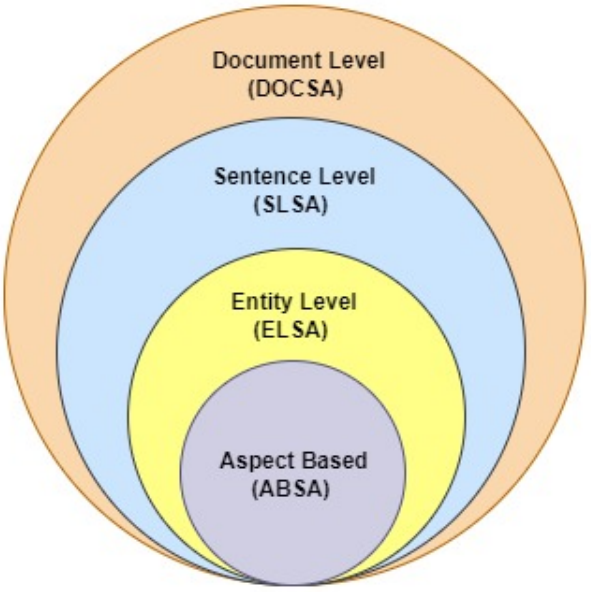

Figure 2.2: Sentiment Analysis Levels

2.1.2.1 Document Level Sentiment Analysis :

In Document Level Sentiment Analysis (DOCSA), it is assumed that the whole content is about a single object written by the author and a generic sentiment is assessed based on the entire document [9]. This type of SA is generally used in chapters or pages of books to assign sentiment polarity. However, two important aspects of DOCSA are cross-domain and cross-language issues. DOCSA is very much domain sensitive. When dealing with documents containing multiple languages, DOCSA encounters challenges in ascertaining sentiment polarity whereas domain-specific DOCSA showed promising results [10]. However, DOCSA can be conducted for both small and large documents, and it is specifically concerned with analyzing sentiment across multiple sentences, regardless of the document's size. The key focus is on capturing sentiment patterns that span beyond individual sentences.

Our research suggests that most DOCSAs are performed on small documents such as reviews, news, emails, SMS, etc. However, documents containing comparatively large texts have also been done. In DOCSA most works are in English. For example, Cao et al. [44] developed a deep learning-based SA on online product reviews. Deng et al. [45] introduced two models, DABS and ODABS, for DOCSA, leveraging four key features.

Rodriguez-Ibanez et al. [46] provided an in-depth exploration of SA in social networks, highlighting the significance of temporal aspects and causal effects. It investigates their applications in diverse contexts like stock markets, politics, and cyberbullying within educational settings. Hao et al. [47] focuses on investigating a sentiment analysis method for recognizing and analyzing official document text using the BERT neural network model. Alshuwaier et al. [9] presented a systematic literature review of deep learning methods for DCOSA, exploring various text features.

DOCSA research in Bangla: Although several research on DOCSA has been conducted in the Bangla language, most of them are on small documents. For instance, Purba et al. [48] conducted a DOCSA and introduced a new Bangla dataset annotated with three emotions (Happy, Sad, Angry). It employs two feature extraction methods, Bag of Words (BoW) utilized by Logistic Regression (LR) and MNB classifiers, and Word Embedding employed by Artificial Neural Network (ANN) and Convolutional Neural Network (CNN) classifiers. Islam et al. [49] introduced an annotated dataset comprising 22,698 Bangla public comments extracted from diverse social media platforms, spanning 12 distinct domains, including Personal, Politics, and Health. Rahman et al. [50] employs Deep Learning-based approaches, specifically CNN and LSTM, for the classification of Bangla text documents. Islam et al. [51] introduce an annotated SA dataset consisting of informally written Bangla texts. The dataset includes public comments from social media across various domains such as politics, education, and agriculture. Rahman et al. [52] classified Bangla text documents using the latest transformer or attention mechanism-based models, specifically BERT (Bidirectional Encoder Representations from Transformers) and ELECTRA (Efficiently Learning an Encoder that Classifies Token Replacements Accurately).

2.1.2.2 Sentence Level Sentiment Analysis :

A document generally contains multiple sentences. When a generic sentiment is extracted from that document, the sentiments related to specific sentences are overlooked. To address this issue, the sentiment is determined for each sentence in sentence-level sentiment analysis (SLSA) [53]. SLSA is highly useful in texts, reviews, and comments where isolated sentences are used. The typical approach involves aggregating the sentiment orientation of individual words within a sentence or phrase to determine the overall sentiment of the sentence or phrase [54]. This methodology aims to categorize a sentence as expressing a positive, negative, neutral, or mixed sentiment, or as being either subjective or objective. Several works on SLSA have been conducted in English. Shirsat et al. [55] targets sentence-level negation identification in news articles and blogs using

a two-step approach involving pre-processing (stop word removal, punctuation mark removal, etc.) and post-processing (sentiment identification and score calculation). Sun et al. [56] proposed a semi-supervised approach that focuses on SLSA by integrating the SenticNet lexicon and a recursive autoencoder with an attention mechanism. Experimental results showcase its superiority over RAE and other existing models. Su et al. [53] introduced a supervised solution for SLSA using gradual machine learning (GML).

SLSA research in Bangla: In Bangla explicitly single sentence level SA has very few works. For example, Hossain et al. [57] introduced a sentence-level SA which employs a unique dataset of diverse comments from online sources, labeled for positive and negative emotions. They aimed to develop a Bengali context-aware system using machine-learning algorithms such as K-NN, DT, LR, SVM, MNB, and RF. However, extensive research has been conducted on multiple sentence-based SA, primarily utilizing data collected from online reviews and comments. Bhowmik et al. [1] conducted restaurant and cricket data SA. Durga et al. [3] introduced a sentiment analysis dataset derived from YouTube comments. These studies are predominantly categorized as short DOCSA.

2.1.2.3 Entity Level Sentiment Analysis :

Analyzing a given text can become intricate when it references multiple distinct entities. These entities may be mentioned repeatedly, both directly and indirectly, and could be associated with various opinions. When sentiment analysis is carried out specifically for those entities then it is called Entity Level Sentiment Analysis (ELSA) [4]. Consequently, addressing the complexities of ELSA may involve multiple sub-tasks. These sub-tasks include Named Entity Recognition (NER) and pinpointing sentiment targets and/or aspects along with their respective polarities [7]. Nonetheless, it's important to note that the approach and granularity of SA can vary, contingent on the particular nature and objectives of sentiment extraction. In the context of ELSA, the identification of entities is a pivotal task [58]. ELSA, a specialized form of SA, leverages techniques from both SA and NER to facilitate the evaluation of sentiment at the entity level. For a comprehensive discussion of existing works on ELSA, NER and its associated challenges, please refer to section 2.4.

2.1.2.4 Aspect Based Sentiment Analysis :

Traditional SA typically centers on categorizing the overall sentiment conveyed in a text without delving into the specifics of what the sentiment pertains to. While this approach

serves its purpose in many cases, it may fall short when the text concurrently addresses multiple subjects or entities, often expressing varying sentiments toward these distinct aspects [59]. Recognizing and discerning sentiments linked to specific aspects within a text represents a more intricate endeavor, known as Aspect Based Sentiment Analysis (ABSA). [60]. ABSA comprises three primary stages: aspect extraction, polarity classification, and aggregation [60]. The process of ABSA commences with the extraction of aspects, a pivotal step that sets it apart from traditional SA. Aspects are typically identified using a predefined set of criteria, which should be thoughtfully tailored to the specific domain of the application [10].

Several researchers have conducted ABSA studies in English. Do et al. [61] proposed a paper concentrated on refining granularity at the aspect level, with a dual focus on aspect extraction in product reviews and sentiment classification of target-dependent tweets. Zhang et al. [62] introduced a new ABSA taxonomy based on sentiment elements, highlighting recent advances in compound ABSA tasks. It explores the use of pre-trained language models. Mowlaei et al. [63] introduced extensions to two lexicon generation methods for ABSA.

ABSA research in Bangla: The field of ABSA in the Bangla language is currently underdeveloped in comparison to English. Only a few research work has been found. Naim et al. [60] proposed a new technique called PSPWA (Priority Sentence Part Weight Assignment) for aspect category or term extraction on publicly available datasets (Cricket and Restaurant). Sultana et al. [64] introduced a Bangla ABSA model, utilizing 4012 Bangla text comments related to cricket, drama, movie, and music from YouTube. Ahmed et al. [65] introduced BAN-ABSA, a manually annotated high-quality Bengali dataset for ABSA. The dataset, annotated by three native Bengali speakers, includes 2619 positive, 4721 negative, and 1669 neutral samples from 9009 unique comments sourced from popular Bengali news portals. Samia et al. [66] proposed a manually annotated Bengali dataset with five aspects and corresponding sentiment. The baseline evaluation utilized the Bidirectional Encoder Representations from Transformers (BERT) model.

2.2 General Procedure of Sentiment Analysis

The general procedure for conducting Sentiment Analysis involves several key steps, including data selection, data collection, data processing, the selection of a suitable approach for sentiment analysis, and the completion of specific analysis tasks. [54]. This

general procedure serves as a foundational framework for the majority of SA research undertakings [10].

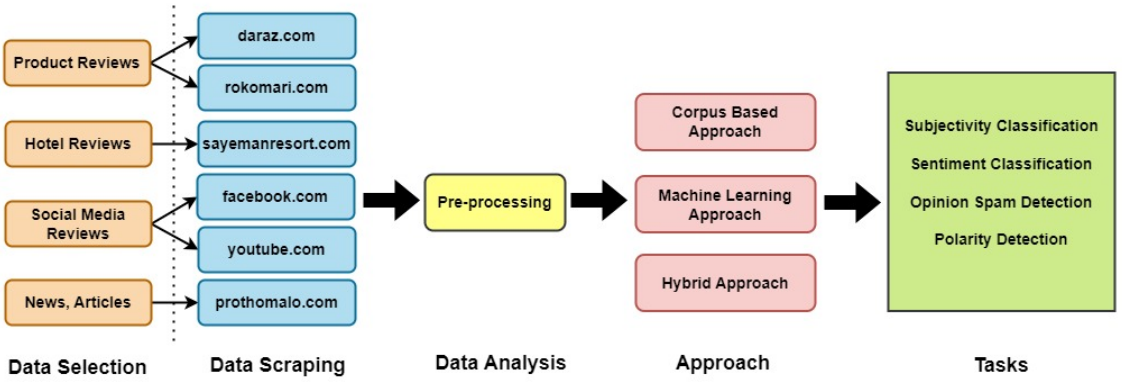

Figure 2.3: General Procedure of Sentiment Analysis

Data Collection: The initial and pivotal stage in any SA task is the selection of domain-specific data. This involves a thoughtful process of handpicking data from a diverse array of online platforms. The choice of the most suitable online platform is contingent upon the specific objectives of the task at hand. These platforms encompass a wide spectrum, including social media giants like Facebook and Twitter, YouTube for video reviews, various online portals, reputable newspapers, official restaurant and hotel websites, and an array of e-commerce platforms, and many more. Following the selection of suitable data, the next step involves its collection from the chosen online platforms. This can be achieved through both manual and automated approaches. The manual approach entails navigating the sites manually and collecting the desired data. However, this method is often time-consuming and costly. Therefore, for handling large volumes of data, it is advisable to develop a web scraper or crawler. A web scraper or crawler is a program designed to systematically traverse the desired platform, gather the required data, and store it. Many pre-existing web scrapers are available online. In cases where a scraper for a specific platform is not readily available, a custom scraper must be developed for that particular platform.

Pre-processing and Feature Extraction: After the data collection phase, comprehensive data analysis becomes a crucial step. For data analysis, the data has to be taken into a standard format which involves tokenization, normalization, stemming, and lemmatization. In addition, selecting appropriate feature extraction methods and subsequently deriving relevant features from the collected data is another important phase. Some of the common feature extraction techniques encompass Bag of Words (BoW), word embeddings like Word2Vec and GloVe, N-grams, term frequency-inverse document frequency (TF-IDF), NLP-based features etc.[67]. It's important to note that the choice

of feature extraction techniques and specific approaches may vary depending on the specific objectives of the analysis.

Classification: Furthermore, a suitable SA approach must be chosen after data preprocessing and feature extraction. There are mainly three popular types of SA approach namely Lexicon based approach, Machine Learning approach, and hybrid approach. We can choose one method based on requirements. Different approaches have different characteristics and result in variations. Ultimately, the chosen approach is applied to fulfill the desired tasks. which can include subjectivity classification, sentiment classification, opinion spam detection, polarity detection, and more.

2.3 Sentiment Analysis from Online Reviews

While SA encompasses a multitude of approaches, they can be broadly categorized into three main categories [10]: Lexicon-Based Approach, Machine Learning Approach, and Hybrid Approach. It's worth noting that within each of these categories, there exist several subtypes and variations. However, only a limited number of survey-based studies have been conducted in the field of Bangla SA in recent years. For example, Sen et al. [68] analyze 75 BNLP research papers, categorizing them into 11 areas, including Information Extraction, Machine Translation, NER, Parsing, and more. They cover papers from 1999 to 2021, with a significant portion published after 2015. The paper discusses Classical, Machine Learning, and Deep Learning approaches, addresses limitations, and outlines current and future BNLP trends. Hira et al. [69] presents an overview of the current landscape of sentiment analysis (SA) and offers a sequence of improved research findings compared to existing work. The study utilized the TOPSIS method to create this sequence and delves into the challenges that need to be addressed to enhance sentiment analysis. Islam et al. [49] begin their research with a comprehensive review of available resources, tasks, and tools in the Bangla NLP domain. It proceeds to benchmark datasets across nine NLP tasks, employing transformer-based models. The research compares monolingual and multilingual models of different sizes, presenting results for individual and consolidated datasets. A review of 108 papers and 175 experiments showcases promising performance with transformer-based models while emphasizing the associated computational costs.

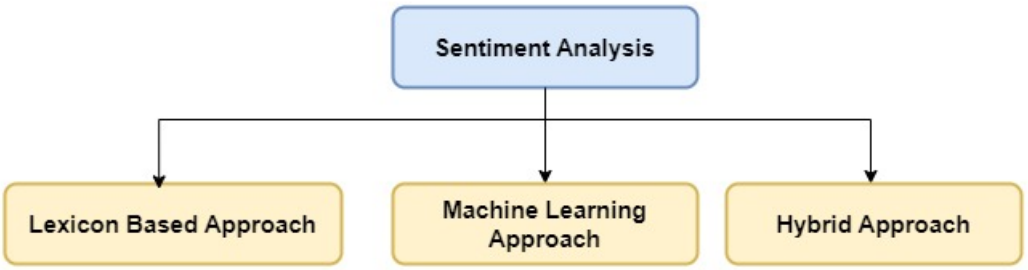

Figure 2.4: Sentiment Analysis Approaches

2.3.1 Lexicon Based Approach

Sentiment lexicons are the terms that convey opinions. Lexicons can be words, phrases, or other expressions labeled with sentiment polarity, typically categorized as positive or negative, along with polarity strength [70]. Lexicon plays a significant role in sentiment classification. Sentiment lexicons are invaluable in scrutinizing the essential subjective elements of texts, including opinions and attitudes. In the lexicon-based SA approach, the general framework typically begins with the collection of data in the form of text sentences. Then words are tokenized. Dividing the texts into wordlists is called tokenization. Then normalization is done by removing punctuation marks, special characters, and stopwords from the tokens [1]. After that comes the Stemming part. Stemming means originating the root word from the tokens. Extra letters are removed from the beginning or end of the normalized tokens. After that, the tokens will go through Parts of Speech (POS) tagging. Each normalized token will be tagged with a part of speech. POS tagger amplifies the weight score of the word. Hota et al. [71] investigate sentiment analysis (SA) across six countries—India, USA, Spain, Italy, France, and the UK—employing a Lexicon-based approach. Utilizing Twitter data from March 15 to April 15, 2020, the study categorizes sentiments as Negative, Neutral, or Positive through both Lexicon-based and Valence Aware Dictionary for Sentiment Reasoning (VADER)-based methods. Furthermore, Hasan et al. [72] propose a different approach for calculating the sentiment. They calculate positive, negative, and neutral scores by adding all negative, positive, and neutral scores. After that, they calculate the percentage of positive, negative, and neutral sentiments by dividing the respective sentiment by the total sentiment. Finally, the sentence's sentiment is assigned, with the sentiment class having the greater percentage. Pre-processing techniques and finding the total sentiment score of a sentence differs based on research methodology [6].

Data Dictionary : The most important aspect of the lexicon-based approach is the data dictionary which holds the labeled sentiment words and phrases. The main purpose

Table 2.1: Existing Bangla Data Dictionary for Sentiment Analysis

Proposed By	No of Words	Availability
Chowdhury et al. [75]	737	Unavailable
Sazzed et al. [74]	1000	Unavailable
Mahmudun et al. [76]	1500	Unavailable
Bhowmik et al. [1]	5,061	Unavailable
Dey et al. [77]	5,100	Unavailable
Akter et al. [78]	9,000	Unavailable
Rahman et al. [79]	12,696	Unavailable
Ali et al. [80]	61,582	Available [81]

of a data dictionary is to assign a sentiment polarity to a word which will lead to the calculation of the sentiment of the sentence [73]. General-purpose, domain-specific, and aspect-based data dictionaries are important for sentiment analysis [74]. However, in Table 2.1, a list of existing Bangla data dictionaries and their number of words are shown.

Several works in Bangla combine SA with data dictionaries. For example, Bhowmik et al. introduced their algorithm, BTSC (Bangla Text Sentiment Score), for sentence score calculation [1]. Additionally, in another study by Bhowmik, Arifuzzaman and Mondal et al. [6], the algorithm developed by [1] is utilized. In this algorithm, the value of a word differs based on the tagged parts of speech and the words' orientation (positive, negative, and neutral). The negative words are calculated and multiplied by orders of -1. Finally, the final sentiment is calculated by multiplying all the words' scores by the value of the score of the negative word. Their data dictionary consists of 5,061 words. Akter et al. [78] follow a very trivial approach to calculate the sentiments to classify 3600 Facebook statuses in Bangla text. They built a data dictionary of 9000 sentiment words. After pre-processing, the words are assigned with polarity. Then the total sentiment score is calculated by adding all the scores. Dey et al. [77] applied summation of all the word scores to calculate the polarity of the sentences. They created a data dictionary of 5100 words. They maintained a list of boost words to amplify the word's score next to the boost word. Furthermore, they manually made a wordlist of negative words in Bangla sentences. Rather than assigning value to those words, they multiplied the summation of all word scores with -1 to calculate the final score. For each negative word, the multiplication increased one time. Finally, they normalized the sentiment score to bring it to their desired level. Iqbal et al. [82] created a bangla sentiment corpus with 7000 texts from social media, newspapers, and portals. They categorized the texts into six categories namely anger, fear, surprise, sadness, joy, and disgust. Rahman et al. [79] collected more than 10,000 sentences from the Prothomalo online news portal. Their major news section was sports review. From this dataset, they prepared a data dictionary

of 12,996 words with negative, positive, and neutral sentiments. Ali et al. [80] proposed the largest Bangla data dictionary named ”BanglaSenti” with a number of 61,582 words. This is the largest data dictionary among the existing Bangla dictionary we found.

Major limitations of the existing Bangla data dictionaries:

The Bangla SA data dictionary lacks annotated data, tools, and resources [79, 83, 84]. Although a few Bangla data dictionaries have been created within the last decade using different methods, they have many limitations [85]. The major limitations of the existing data dictionaries are discussed below.

- Most of the work on SA in Bangla is machine learning-based [68]. Thus, a limited number of data dictionaries have been created till now. Moreover, most of the researchers do not create original dictionaries. Rather they translate their tokens using English data dictionaries [72]. All the existing Bangla data dictionaries have a very small number of datasets [85].
- Data dictionary for POS tagger words plays a vital role in enhancing polarity [85]. But Bangla SA lacks POS tagger dictionaries [74]. Only a few Bangla POS tagger data dictionaries have been developed. For example, [1] proposes a POS tagger extended LDD for reviews of cricket and restaurants.
- Banglish means a mixture of Bangla and English texts. Romanized Bangla means Bangla words written in English. At present, the common trend in social media platforms is Banglish and Romanized writing style. But to date, no such dictionary exists in Bangla [86].
- Most data dictionaries are developed manually or by machine translation for domain-specific datasets. For example, [1], [72], and [80] used these procedures. Furthermore, a few researchers chose a different method. In [74], a corpus-based SA is developed. Due to these reasons, verifying the existing data dictionary’s fairness and reliability becomes difficult [85].
- Most of the lexicon-based SA uses translated data dictionaries [72]. Hence, the real data dictionary is not created. In [1], a manual approach is used to develop extended LDD. Moreover, the Human-based process of developing a data dictionary consumes time, needs expert intervention and efficient translation tools, etc. [74]. Although all these requirements are not necessary at the time to develop a dictionary, the process is very difficult and costly.

2.3.2 Machine Learning Based Approach

The majority of research in SA is predominantly based on Machine Learning techniques. Interestingly, most of these works are conducted in the English language. Nevertheless, there are also limited studies on SA in other languages such as Hindi, Chinese, Arabic, and Bangla. However, when compared to the extensive body of work in English, the research in these languages remains relatively limited. In the early stages of applying machine learning techniques to SA, many researchers primarily relied on classical ML algorithms, such as LR, Logistic Regression (LogR), NB, SVM, Decision Trees (DT), RF, XGBoost, and K-NN, among others. As time progressed, there has been a noticeable shift towards an increased utilization of deep learning models in SA. For instance, the widespread adoption of models like RNN, LSTM, and BERT has been observed. Additionally, some researchers have conducted studies that involve a comparison between deep learning and classical machine learning approaches. These comparative analyses aim to assess the performance and suitability of these different methodologies for SA tasks. The field of SA continues to evolve, with both classical and deep learning techniques playing essential roles in advancing our understanding and capabilities in this domain.

Studies Utilizing Classical Machine Learning Algorithms:

Mahtab et al. [87] analyze the sentiment polarity of the news portal and social media platform texts as positive, negative, neutral, praise, criticism, and sadness. They preprocessed their labeled data with the TF-IDF feature extraction technique. They used SVM, DT, and NB machine learning models for classification with their data collected from sites and ABSA (Dataset for Bangladesh Cricket and Restaurant Related Comments). The ABSA dataset shows better results because of the large dataset. Tuhin et al. [88] uses two machine learning techniques, the NB and Tropical method are proposed to classify emotion into six categories. These categories include happy, tender, excited, sad, angry, and scared. They analyzed sentence-level and article-level text and compared these two methods. The topical method showed better performance. Khatun et al. [83] proposed an ML-based method with 5500 book reviews collected from Social Media for positive and negative polarity. They used AdaBoost, DT, SVM, LightGBM, and RF. RF showed the highest accuracy with 98.39%. Akter et al.[89] propose a Bangla e-commerce review SA of 7905 datasets from Daraz e-commerce online platform. They use RF, LogR, SVM, K-NN, and XGBoost algorithms with this dataset. K-NN performs better and obtains a higher accuracy of 96.25% and a better f1-score of 96%. Kaiser et al. [90] utilized 11,006 Facebook comments for hate speech detection. They categorized the comments into eight categories. They used LR, DT, RF, MNB, K-NN, and

SVM classical ML algorithms. MNB showed better results. Hence, it can be stated that studies employing classical machine learning algorithms for SA have delved into several traditional models.

Studies Utilizing Deep Learning Algorithms:

A Bangla and Romanized Bangla (Bangla written in English) textual dataset (BRBT) has been developed and multi-validated by Hassan et al. [86]. This dataset comprises text collected from various sources, including Facebook, YouTube, Twitter, online news portals, and product review pages. To assess its performance, this dataset was subjected to testing using the Deep Recurrent Model, specifically a LSTM network. The processed Bangla dataset demonstrated the highest accuracy of 70%. Durga et al. [3] use RNN to classify YouTube video comments on Bangla and Romanized Bangla texts. Rahib et al. [91] collected 10,581 COVID-19-related comments from Facebook and YouTube and used SVM, RF, CNN, and LSTM for classification. They are classified into three categories namely insightful, curious, and gratitude. LSTM outperformed all other algorithms and showed 84.92% accuracy. Naim et al. [60] proposed an aspect-based SA technique named PSPWA (Priority Sentence Part Weight Assignment) with CNN algorithm using Cricket and Restaurant data. Banik et al. [92] discussed a survey on Bangla Text SA and its future scope. Karim et al. [93] developed BengFastText, the largest Bengali word embedding model, based on 250 million articles. They conducted three experiments involving document classification, SA, and hate speech detection, using BengFastText in a Multichannel Convolutional-LSTM (MConv-LSTM) network. Results show that BengFastText outperforms baseline models, achieving high F1-scores of up to 92.30% in document classification, 82.25% in SA, and 90.45% in hate speech detection during 5-fold cross-validation tests. Hoq et al. [94] developed four models using a hybrid of CNN and LSTM with various Word Embeddings (Embedding Layer, Word2Vec, Glove, and CBOW) to detect emotions in Bangla text, including words and sentences. These models accurately identify basic emotions (happiness, anger, and sadness) and outperform classical Machine Learning techniques such as SVM, NB, and K-NN, using Facebook Bangla comments as the dataset. The best-performing model, which combines Word2Vec embedding with a CNN-LSTM hybrid, achieved an impressive accuracy of 90.49% and an F1 score of 92.83%.

Furthermore, the Large Language Models (LLMs) employed for Sentiment Analysis are powerful pre-trained models that excel in understanding complex linguistic structures. Leveraging extensive language knowledge, it enhances sentiment analysis by capturing

nuanced contextual cues and adapting to diverse textual expressions. To gain a comprehensive understanding of current research in sentiment analysis, including the utilization of pre-trained language models and their methodologies please refer to Section 2.5. This section provides an insightful overview, offering valuable insights into the evolving landscape of sentiment analysis research.

2.3.3 Hybrid Approach

In the hybrid approach, the lexicon and machine learning approaches are combined for SA. Bashar et al. [95] utilize machine-learning models namely NB, LSVM, LR, and RF, and a lexicon-based approach named VADER for sentiment analysis of COVID-19 public sentiment data. Bhowmik et al. [1] provided a rule-based SA algorithm for Bangla text (BTSC). The authors added an Extended Lexicon Data Dictionary and a machine-learning approach to classify the polarity of text at the sentence level. In addition, two feature matrices using TF-IDF and sentiment scores were included using the BTSC algorithm. The algorithm was tested with SVM, LR, and K-NN classification methods, and SVM showed an accuracy of 82.21%. Bhowmik et al. [6] proposed a Deep Learning method with fine-tuning for Bangla Text SA. They used the BTSC algorithm with an extended data dictionary to find the sentiment score. The proposed LSTM model, namely HAN-LSTM, Bi-LSTM, and BERT-LSTM, show 78.52%, 80.82%, and 84.18% accuracy, respectively. Tabassum et al. [96] combined lexicon and ML methods with 1050 reviews collected from Twitter. They preprocessed data using tokenization, normalization, and POS tagging. They used negative and positive labeled word lists for training. Then trained using the RF algorithm. They found an accuracy of 87%. Hasan et al. [97] propose Bangla Aspect Based SA techniques using RF, SVM, CNN, FastText, BERT, XLM, and RoBERTa. BERT showed better results.

2.4 Entity Level Sentiment Analysis (ELSA)

2.4.1 Existing Works in ELSA

Several ELSA researches have been conducted in English. For instance, Fu et al. [4] showcases the development of an ELSA system for English telephone conversation transcripts in contact centers. Two approaches are presented: one solely based on the transformer-based Distil-BERT model, and another incorporating a CNN along with heuristic rules. Ronningstad et al. [7] investigates ELSA for longer texts, specifically

exploring sentiment toward volitional entities (persons and organizations). The study annotates professional reviews to assess overall sentiment towards each entity, finding that existing tasks and models do not sufficiently address ELSA. The experiments analyze document-level, sentence-level, and target-level SA contributions, revealing shortcomings and the need for further research on sentiment-relevant relations to volitional entities. The paper also includes a survey of previous relevant work. Toledo et al. [5] proposes a multi-domain Targeted Sentiment Analysis (TSA) system that enhances a training set with diverse weak labels from various domains. Weak labels are obtained through self-training on the YELP reviews corpus. Extensive experiments across different domains demonstrate the effectiveness of this approach on three evaluation datasets. Ding et al. [98] proposed an ELSA, creating a dataset of 3,000 labeled issue comments from 10 GitHub projects. They introduce SentiSW, a tool for SA and entity recognition, classifying issue comments into <sentiment, entity> tuples. Furthermore, Manman et al. [99] presented the Negative Sentiment Smoothing Model (NSSM) for ELSA, assessing sentiments toward the 45th President in news paragraphs. NSSM adjusts sentiment scores based on Negative Associated Entities (NAEs), evaluated across CNN, FOX, and NPR news data over three months. Huang et al. [100] proposed an ELSA research in the Chinese language. This research focuses on entity-level sentiment analysis in financial text using a pre-trained language model, specifically Bidirectional Encoder Representations from Transformers (BERT). However, to the best of our knowledge, we found only a few ELSA research in Bangla.

Named Entity Recognition (NER): ELSA is mainly comprised of two tasks. Entity identification and sentiment analysis correspond to the entity. Hence, for ELSA task completion it is necessary to identify entities. Identifying entity-type objects from texts is known as entity recognition. Entities can be a person, organization, place, time, product, etc. A vast body of work in NER has been conducted in English, covering diverse domains. For instance, Jehangir et al. [101] conducted a comprehensive analysis of Named Entity Recognition (NER) methodologies, encompassing unsupervised, rule-based, supervised, and deep learning approaches. The study explores relevant datasets, tools, and deep learning techniques, including CNNs, RNNs, Bidirectional LSTM, and transfer learning. The paper also discusses the challenges faced by NER systems and outlines future directions in the field. VeeraSekharReddy et al. [102] proposed AL-CRF model leverages Conditional Random Field (CRF) and Active Learning for NER. It efficiently clusters samples, utilizes stratified sampling, and employs entropy-based selection to iteratively enhance the NER model, achieving improved results with fewer manually marked training samples. Ullah et al. [103] developed the MedNER model,

utilizing domain-specific embeddings and Bi-LSTM which enhances NER in biomedical text mining. Achieving a remarkable 98% F1-score on a Covid-related scientific publications dataset, the model outperforms previous approaches. This underscores the effectiveness of our deep learning-based approach in accurately recognizing and classifying biomedical named entities, offering promising avenues for future advancements in biomedical text mining.

2.4.2 Research Challenges

Bangla ELSA presents unique challenges stemming from the linguistic intricacies and data limitations specific to the Bangla language. Tackling these challenges is crucial for advancing SA capabilities in the Bangla text domain. Below are some prominent research challenges in the pursuit of enhancing ELSA for Bangla content.

1. **Limited Labeled Data:** The availability of labeled data for Bangla ELSA is often limited. Annotating data with entity-level sentiment labels is a resource-intensive task.
2. **Entity Recognition Accuracy:** Accurate recognition of entities in Bangla text is challenging due to the complex linguistic structure of the language. Existing entity recognition tools may not perform well on Bangla text.
3. **Entity-Level Sentiment Annotation Guidelines:** Developing comprehensive guidelines for annotating sentiment at the entity level in Bangla text is crucial. Establishing clear criteria for sentiment labeling enhances the consistency of annotated data.
4. **Cross-Domain Adaptation:** Models trained on one domain may not generalize well to other domains. Adapting models to different domains without substantial labeled data is a challenge in Bangla ELSA.
5. **Resource-Scarce Environment:** Limited availability of computational resources and specialized tools for Bangla sentiment analysis research can hinder the development and evaluation of sophisticated models.

Researchers and practitioners addressing these challenges contribute to the advancement of Bangla ELSA, enabling more accurate and context-aware sentiment understanding in diverse applications.

2.4.3 ELSA in Bangla Language

Although extensive research has been conducted on SA in Bangla, English, French, and Arabic, the same level of attention has not been given to ELSA in these languages [104]. Bangla is the 7th most widely spoken language worldwide with a global speaker base of 265 million [68]. Even though a few research on ELSA have been conducted in English, to the best of our knowledge, very few ELSA research has been found in Bangla to this date. [7]. Thus it deserves more focus in this area of natural language processing. Moreover, the domain of ELSA in Bangla lacks adequate research due to limited annotated datasets. Consequently, there is a pressing need for further exploration and advancement in ELSA techniques specific to Bangla.

However, several research efforts have contributed to Bangla NER. For instance, Hoang et al. [105] presented NER systems for SemEval-2023 in English and Bangla, addressing challenges with fine-grained named entity types. They use data augmentation based on BabelNet concepts and Wikipedia redirections, leveraging the mDeBERTa language model. The augmented systems outperform baselines, achieving macro-f1 scores of 52.64% and 64.31%, indicating improvements of 2.38% and 11.33% for English and Bangla, respectively. Haque et al. [106] introduced the B-NER dataset, a novel Bangla NER dataset comprising 22,144 manually annotated Bangla sentences from newspapers and Wikipedia. It includes 9,895 unique words categorized into eight entity types, addressing a significant limitation in Bangla NER research. Mukherjee et al. [107] presented MuRILCRF Bangla, a system utilizing a pre-trained language model for complex entity identification and type recognition in the low-resource language Bangla. The model achieves a notable macro average F-score of 76.27% for the sequence labeling task.

2.5 Large Language Models for SA and NER

Large Language Models (LLMs) play a pivotal role in enhancing the efficiency of various machine-learning tasks. As elucidated earlier, the preparation of datasets constitutes a fundamental aspect of machine learning research, with the quality and quantity of data directly influencing the efficacy of machine learning models. However, the creation of extensive datasets and the subsequent training of models incur significant time and cost. To address this challenge, LLMs have been devised, undergoing pre-training with substantial volumes of data. Leveraging a transfer learning strategy, these models are pre-trained on diverse data, enabling the fine-tuning of specific datasets to yield improved

results. This approach, characterized by adapting pre-existing knowledge to new tasks, has positioned LLMs as invaluable tools in contemporary research endeavors.

Numerous studies have been undertaken, leveraging pre-trained language models. For instance, Islam et al. [108] collected 17,852 comments from prothom alo online portal and classified using BERT, Word2Vec, and Fasttext. BERT performed better than other models. Bhattacharjee et al. [109] introduce a variant of the pre-trained language model BERT namely BanglaBERT and BanglishBERT. Prottasha et al.[110] harnessed BERT's transfer learning capabilities to enhance a CNN-BiLSTM model's performance in sentiment analysis. The research also introduced transfer learning to classical machine learning algorithms for performance comparison and explored various word embedding techniques like Word2Vec, GloVe, and fastText, comparing their effectiveness with BERT's transfer learning approach. Kabir et al. [8] developed the largest Bangla SA dataset containing 1,58,065 book reviews. They utilized LR, RF, XGB, MNB, SVM, LSTM, and Bangla-BERT. Bangla-BERT performed better than other classification algorithms. In addition, they provided a details list of existing SA datasets and a detailed error analysis.

In the realm of sentiment classification and entity recognition, several prominent LLMs have emerged as widely adopted solutions. A few of these significant models are discussed below.

2.5.1 BERT

The Transformer architecture, introduced in the groundbreaking paper "Attention is All You Need" by Vaswani et al. [111] in 2017, laid the foundation for significant advancements in natural language processing. Leveraging this architecture, Google AI Language unveiled BERT (Bidirectional Encoder Representations from Transformers) in 2019 [112], a pre-trained language model that has since become a cornerstone in various NLP applications. Unlike traditional language models that read text in one direction, BERT utilizes a bidirectional approach, considering both the left and right context in all layers. This bidirectional understanding allows BERT to capture intricate language patterns and relationships. BERT follows an unsupervised learning approach.

BERT is pre-trained on diverse unlabeled text data, drawing from the BooksCorpus (800 million words) and English Wikipedia (2,500 million words) for an extensive and varied linguistic foundation [112]. This training process involves the model mastering the prediction of missing words within sentences, leveraging the Masked Language Model

(MLM) technique. Furthermore, BERT is equipped to excel in next-sentence prediction (NSP) tasks during its training phase.

Following this comprehensive training regimen, BERT showcases its versatility by effectively addressing more than eleven distinct natural language processing (NLP) tasks. These tasks include but are not limited to, sentiment analysis, named entity recognition, and text classification, demonstrating the model's robustness and adaptability across a multitude of linguistic challenges. The architecture of the BERT model is visually depicted in Figure 2.5.

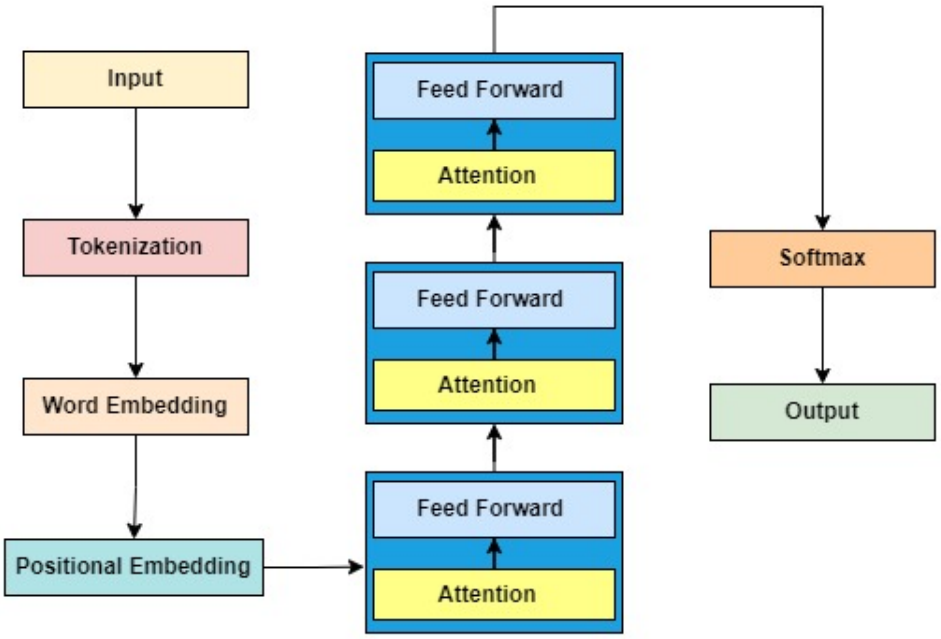

Figure 2.5: BERT Model Architecture

The key steps in the BERT model, include input processing, tokenization, embedding, encoder stack, attention mechanism, feed-forward network, softmax, and output. These phases are described below in brief.

- **Input:** BERT processes input text, which may consist of one or more sentences.
- **Tokenization:** Each sentence is tokenized into subword or word pieces.
- **Word Embedding:** The tokenized sentences are then converted into embeddings, representing each token with a vector. BERT employs embedding layers to convert the tokenized input into dense vectors that capture contextual information.
- **Positional Embedding:** Positional embeddings are added to the token embeddings to provide information about the position of tokens in the sequence.

- **Encoder Stack:** The model utilizes a stack of encoder layers to capture hierarchical contextual information from the embeddings.
- **Attention:** Multi-Head Self-Attention Mechanism helps BERT attend to different parts of the input sequence while processing each token.
- **Feed-Forward:** After attention, the model passes the information through feed-forward neural networks to refine the features.
- **Softmax:** The final layer employs softmax activation to generate probabilities for different classes or tokens.
- **Output:** BERT produces an output that captures contextual understanding and can be used for various natural language processing tasks.

Two primary variants of the BERT model were developed by Devlin et. al [112], namely BERT-base and BERT-large. Details regarding the associated parameters for each variant are presented in table 2.2.

Table 2.2: BERT-Large vs BERT-Base Architecture Details

Model	Layer Number	Hidden Layers	Attention Heads	Parameters
BERT Large	24	1024	16	340M
BERT Base	12	768	12	110M

The model's architecture, Transformer, enables parallelization, making it efficient for training on large datasets. During fine-tuning, BERT can be adapted to specific tasks with a relatively small amount of labeled data. Several variants of BERT have been developed to address specific tasks and languages. Notable variants include RoBERTa, DistilBERT, ALBERT, and many task-specific models fine-tuned from BERT. BERT's efficiency lies in its ability to capture contextual information, making it adept at various natural language processing tasks, including sentiment analysis, named entity recognition, and question-answering. Its pre-trained nature reduces the need for extensive task-specific labeled data, making it applicable to diverse domains.

BERT Model for Bangla SA and NER:

1. **Bangla-BERT-Base**
   Sarkar et al. [113] introduced Bangla-Bert-Base pretrained language model designed for the Bengali language in 2020. The training corpus was derived from the Bengali Common Crawl Corpus downloaded from OSCAR[1] and the Bengali

[1] https://oscar-project.org/

Wikipedia Dump dataset[2]. The corpus underwent preprocessing to adhere to the BERT format. To facilitate training, the BNLP[3] package was utilized for training Bengali sentence pieces, with a vocabulary size of 102,025. The training was performed on a single Google Cloud GPU over 1 million steps. The architecture employed for Bangla-Bert-Base follows the bert-base-uncased architecture. This model demonstrates utility across various NLP tasks, including Bangla text classification, sentiment analysis, named entity recognition, machine translation, and document summarization.

2. **mBERT-Bengali-NER**
   Additionally mBERT-Bengali-NER, an NER model was introduced for Bangla texts. This transformer-based model is constructed using the bert-base-multilingual-uncased[4] model and Wikiann Datasets. The training of mBERT-Bengali-NER involved the utilization of the Wikiann datasets[5], and the training process employed the transformers-token-classification script. The model was trained for a total of 5 epochs, and the training was executed on a Kaggle GPU. The bert-base-multilingual-uncased model, on which mBERT-Bengali-NER is based, is pre-trained on the top 102 languages with the largest Wikipedia, utilizing a MLM objective.

Limitations of BERT: Despite its successes, BERT has limitations. It requires substantial computational resources for pre-training, and its large model size can be challenging to deploy on resource-constrained devices. In summary, BERT's bidirectional approach, efficient pre-training, and adaptability to various tasks contribute to its prominence in natural language processing. However, researchers continue to explore enhancements and address limitations to advance language models further.

2.5.2 GPT

GPT, or Generative Pre-trained Transformer, is an advanced language model built on transformer architecture [114]. Developed by OpenAI, GPT is pre-trained on vast amounts of diverse data, enabling it to generate human-like text and perform various natural language processing tasks. The model excels in understanding context and generating coherent, context-aware responses. GPT's success lies in its ability to capture long-range dependencies in language, making it effective for a wide range of applications, from text completion to language translation [115]. However, like any model, GPT has its

[2]https://dumps.wikimedia.org/bnwiki/latest/
[3]https://github.com/sagorbrur/bnlp
[4]https://huggingface.co/bert-base-multilingual-uncased
[5]https://huggingface.co/datasets/wikiann

limitations, including potential biases in generated content and the need for substantial computational resources for training and fine-tuning.

2.5.3 XLNet

XLNet is a state-of-the-art language model that belongs to the transformer-based architecture family. Introduced by Google AI and Carnegie Mellon University researchers, XLNet stands out for its bidirectional context learning, overcoming some limitations of traditional models like BERT [116]. XLNet employs a permutation language modeling objective, enabling it to consider all possible permutations of words in a sentence, enhancing its understanding of context and relationships between words. This approach improves the model's performance in capturing bidirectional dependencies and makes it effective for various natural language processing tasks [117]. However, like other large language models, XLNet requires significant computational resources for training and can be challenging to fine-tune for specific applications.

2.5.4 SpaCy

SpaCy is an open-source NLP library designed for efficient and high-performance processing of textual data. Developed by Explosion AI, SpaCy is written in Python and is widely used for various NLP tasks, including tokenization, POS tagging, NER, and dependency parsing. It features pre-trained models for multiple languages, making it convenient for researchers and developers working on multilingual applications [118]. SpaCy is known for its speed, accuracy, and ease of use, providing a user-friendly interface for common NLP tasks. It also offers customizable pipelines and supports integration with other machine learning frameworks, making it a valuable tool in the NLP community.

2.5.5 Flair

Flair is an open-source NLP library developed by Zalando Research. It is designed to facilitate state-of-the-art NLP tasks and offers a unique approach to contextual string embeddings [119]. Flair is written in Python and provides pre-trained models for various NLP tasks, including part-of-speech tagging, named entity recognition, and text classification. What sets Flair apart is its use of contextual string embeddings, capturing word meanings based on their surrounding context. This contextual information enhances the

performance of NLP models. Flair is user-friendly, supports multiple languages, and allows researchers and developers to easily integrate advanced NLP capabilities into their applications.

Chapter 3

Proposed Approach

Our proposed approach for Bangla ELSA aims to identify ”Product” entities and their corresponding sentiment orientation in Bangla online reviews. The study focuses on three product types: Books, health products, and women’s clothing, with reviews collected from Rokomari and Wafilife online bookstores, as well as the Daraz e-commerce site. For faster data collection we developed a web crawler. After data collection, we filtered reviews to include reviews that contained only Bangla words. Afterward, the entities were tagged by the annotators. In addition, entity sentiments were also annotated by them. After dataset preparation, a pre-trained language model-based method was used for entity identification and SA.

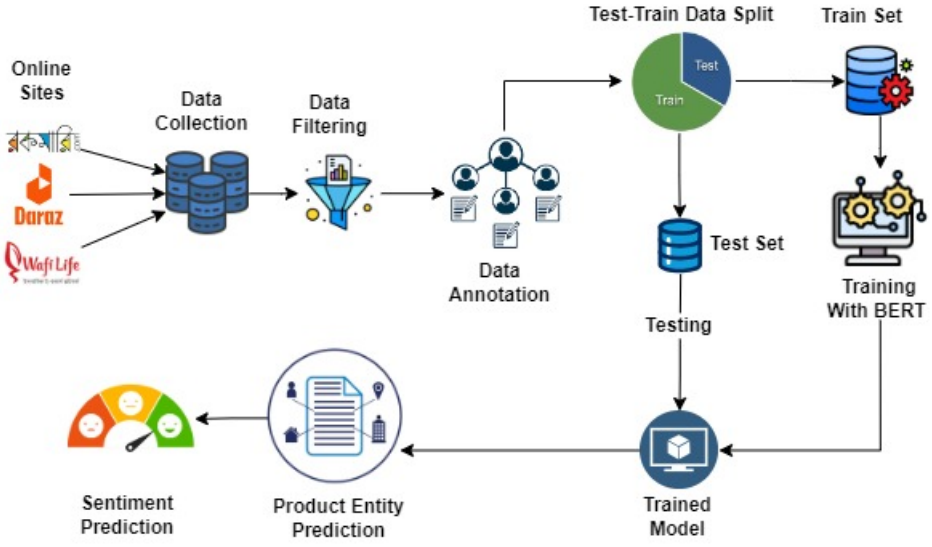

Figure 3.1: Proposed Approach for Entity Level Sentiment Analysis

For Bangla ELSA in online reviews, a pipeline approach was followed. We combined NER identification followed by SA. Two BERT variants pre-trained for Bangla were

selected: bangla-bert-base model for SA and mbert-bengali-ner for NER. The annotated data underwent fine-tuning with the pre-trained models, resulting in our ELSA outcomes.

Hence, our proposed methodology can be primarily divided into 2 sections namely dataset preparation and ELSA model development. However, these two phases can be subdivided into the following subtopics.

- **Dataset Construction**
    1. Data Collection
    2. Data Filtering
    3. Data Annotation
    4. Data Splitting
- **ELSA Model Development**
    1. Data Formating
    2. Training
    3. Model Development
    4. Testing

3.1 Dataset Construction

The dataset forms the core foundation of any machine learning research endeavor. With utmost dedication, we invested substantial effort in crafting a sizable and standardized dataset. Despite facing time and cost constraints, we meticulously filtered the data to curate a robust dataset. Our dataset creation process unfolds through distinct phases, encompassing data collection, language identification, filtering, and annotation. Refer to Figure 3.2 for an overview of this meticulous process, and delve into the subsequent sections for a concise exploration of each dataset creation phase.

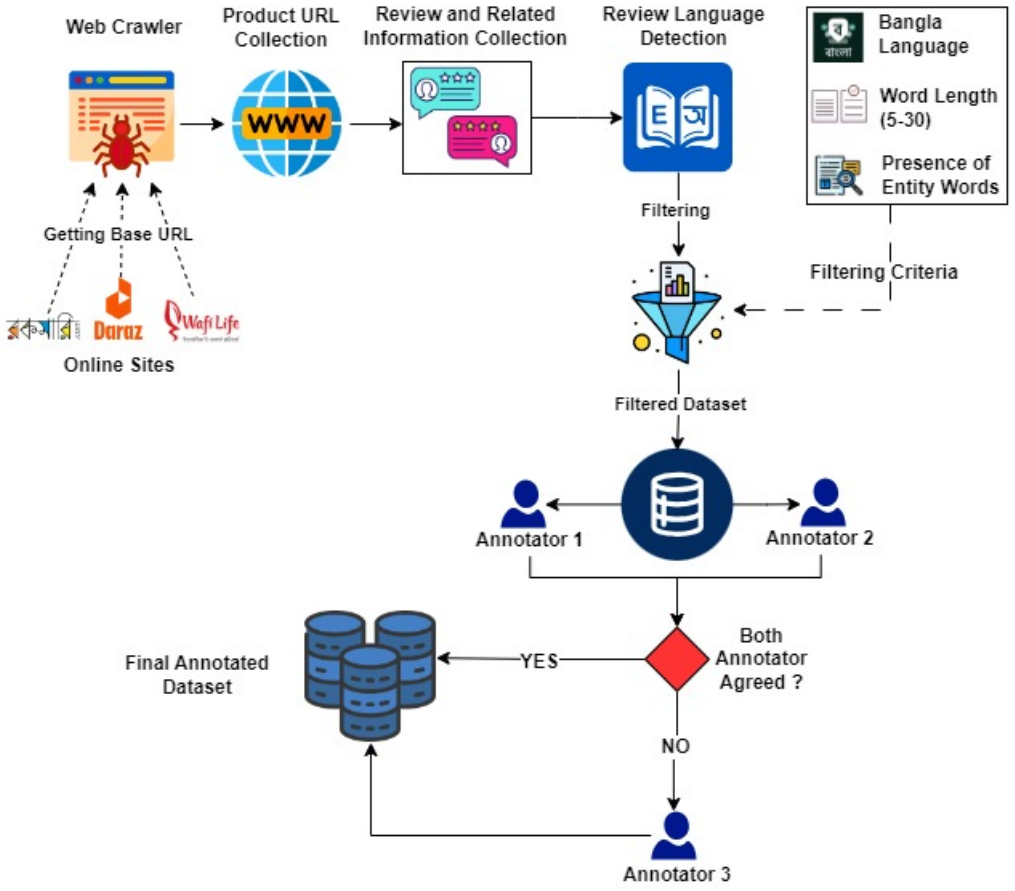

Figure 3.2: Overview of the Dataset Creation Process

3.1.1 Data Collection

For data collection, we chose three prominent online platforms namely Rokomari[1] and Wafilife[2], well-known online bookshops in Bangladesh, where we gathered book reviews. Additionally, we selected Daraz[3], a popular e-commerce site in Bangladesh, for collecting reviews on health products and women's clothing. In-depth statistical analysis, annotation accuracy, and properties of our collected dataset are discussed in Chapter 4. However, the detailed process of data collection is outlined in the following sections.

1. **Web Crawler Development:** These three platforms house an extensive dataset, and manually collecting reviews from such a vast volume would be time-consuming. To expedite the data collection process, we designed a web crawler using the Python programming language. Employing various libraries such as BeautifulSoup, Selenium, Pandas, Openpyxl, and Webdriver facilitated URL connection

[1]https://www.rokomari.com
[2]https://www.wafilife.com
[3]https://www.daraz.com.bd

and data scraping. However, occasional site restrictions prompted us to implement a proxy IP obtained from "https://free-proxy-list.net/" to overcome request blocks resulting from multiple attempts in a short period. Figure 3.3 shows the data collection flow of our developed web crawler.

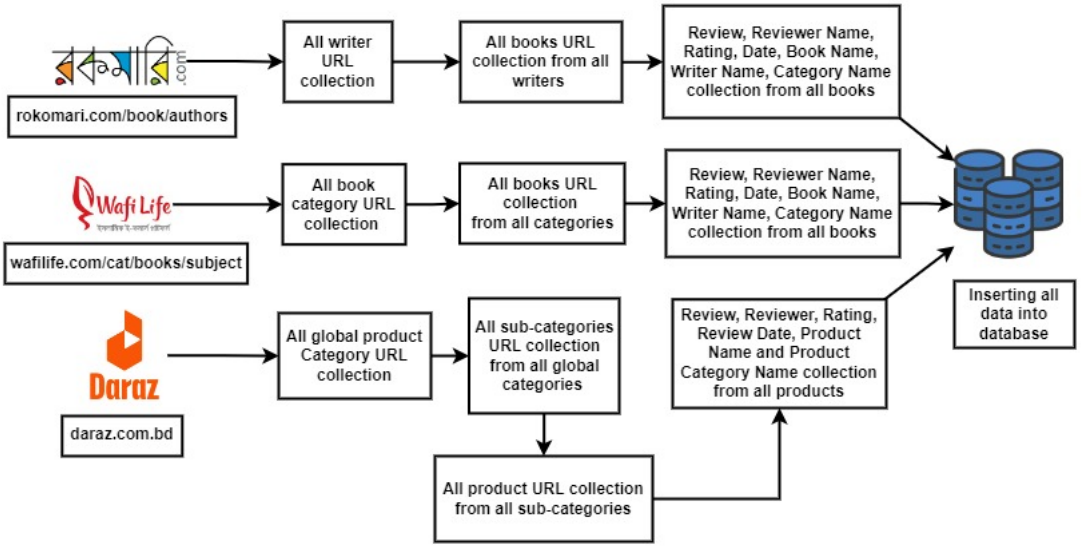

Figure 3.3: Web Crawler Architecture

2. **URL Collection:** Following the development of the web crawler, our next step involved collecting URLs for books and products. For book reviews from Rokomari and Wafilife, we initially gathered the URLs of the authors and categories respectively. Subsequently, we iterated through each author and category URLs, collecting the URLs of their corresponding books for review. In the case of Daraz, we initially collected global product category URLs and then, based on these global categories, collected sub-category URLs. Finally, these URLs were stored in a MySQL database for further iteration.

3. **Review and Related Data Collection:** After collecting the URLs, we meticulously extracted reviews along with additional details such as the reviewer's name, rating and date of the review. For book products, we also gathered information on the book's category and name. Similarly, for health and clothing products, we collected the category name and sub-category name.

4. **Review Language Detection:** Additionally, we conducted language analysis on the reviews, tagging them based on language categories such as Bangla, English, Bangla-English mixed, and Romanized Bangla. Finally, we inserted these collected values into the database.

3.1.2 Data Filtering

Primarily, we translated all non-Bangla reviews into Bangla using the Google Translate library. Due to time constraints, we didn't assess translation accuracy, prompting us to filter our data based on language and retain only Bangla reviews. Further scrutiny revealed reviews that were similar and comprised only a few words without any entity references. Consequently, we refined our dataset by including reviews with a word count between 5 and 30. As our objective involved ELSA, reviews devoid of entity words posed a challenge for annotation. To address this, we scrutinized the dataset, identifying frequently used words, and ultimately compiling a list of 60,597 words. From this extensive list, we selected words with a frequency exceeding 50, resulting in a refined set of 1,579 words. From this wordlist, we selected the top 107 entity words manually. Figure 3.4 visually illustrates this intricate filtering process.

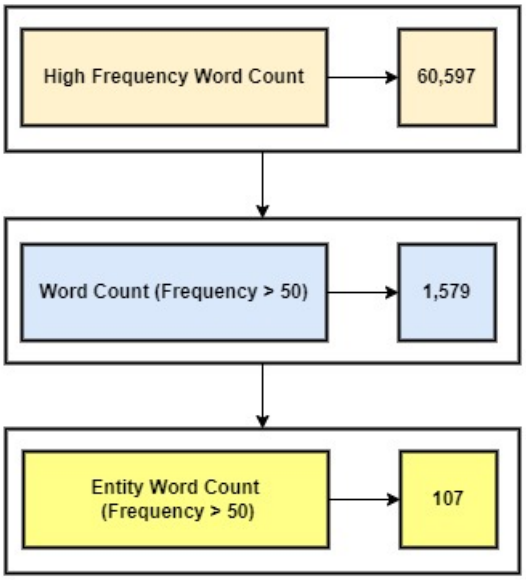

Figure 3.4: High-Frequency Entity Word Selection Process

Subsequently, our dataset underwent additional filtering to include only reviews featuring at least one of these identified entity words. This process culminated in the creation of our Bangla product review dataset, meeting the specified 5-30 word limit.

3.1.3 Data Annotation

After applying the filtering process, we gathered a dataset comprising 10,000 review samples. These reviews are composed of Bangla words, with lengths ranging from 5 words to 30 words. Furthermore, we proceeded to annotate our prepared data using two

native annotators, focusing on two primary aspects. Initially, they annotated the entity words and subsequently assigned sentiment orientations to the product reviews. The sentiment of the reviews was systematically annotated into three categories: positive, negative, and neutral. Additionally, to tag the entity words, we employed the "_NE_" marker, strategically placed before each entity word for clarity and categorization. A sample annotated review structure is shown in figure 3.5.

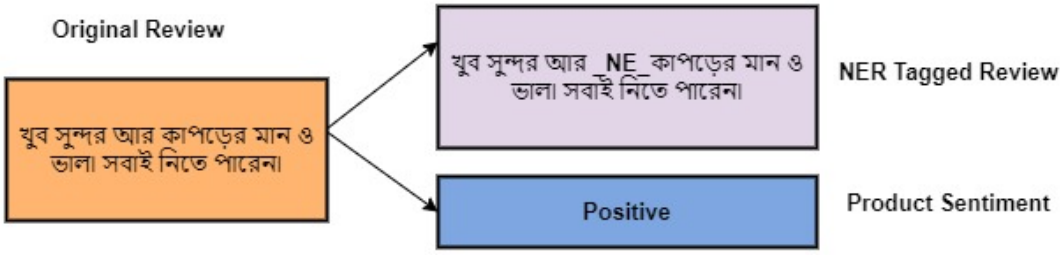

Figure 3.5: Structure of Annotated Data

After completing the initial annotation, we collected both the agreed and disagreed annotated reviews between the annotators. Our annotators reached an agreement on 84.18% of the data. The confusion matrix depicting the sentiment annotations by two annotators is illustrated in Figure 3.6. To reconcile disparities, a third annotator assessed the reviews with conflicting sentiments, and subsequently, we computed the inter-annotator agreement. Additionally, for product entity annotation, the matrix is presented in Figure 3.7. Similar to sentiment annotations, third-party annotation was employed for disputed entities, and the agreement level was calculated accordingly.

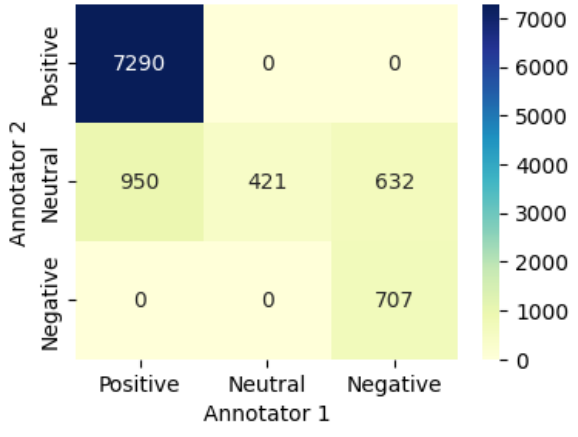

Figure 3.6: Sentiment Annotation Confusion Matrix

In our analysis, we employed Cohen's Kappa mathematical equations to calculate inter-rater reliability scores. The specific formulas used for the computation are detailed below:

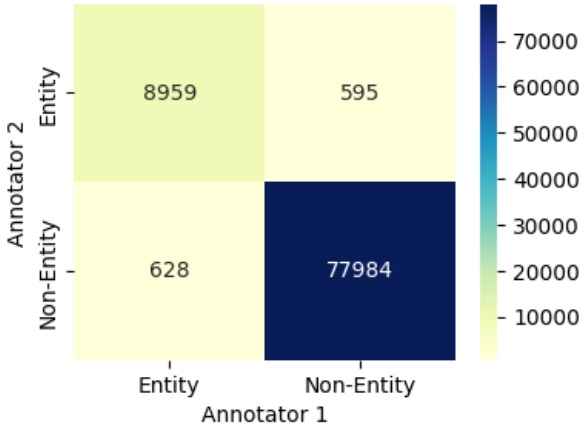

Figure 3.7: Entity Word Annotation Confusion Matrix

$$\text{Po}_i = \frac{\text{Total\_Agreed}_i}{\text{Total}_i} \quad \text{for each class } i$$

$$\text{Po} = \frac{\sum_i \text{Po}_i}{\text{Number of Classes}}$$

$$\text{Pe}_i = \left(\frac{\text{Total}_i}{\text{Total\_Samples}}\right)\left(\frac{\text{Total\_Agreed}_i}{\text{Total\_Samples}}\right) \quad \text{for each class } i$$

$$\text{Pe} = \frac{\sum_i \text{Pe}_i}{\text{Number of Classes}}$$

$$\text{Kappa} = \frac{\text{Po} - \text{Pe}}{1 - \text{Pe}}$$

Utilizing the Cohen Kappa score calculation formula described above, we determined the Kappa score for our annotated dataset. The resulting sentiment analysis score, **0.5936**, suggests a moderate level of agreement between annotators. Furthermore, The resulting entity annotation score, **0.9283**, suggests an almost perfect level of agreement between annotators. This assessment aligns with the criteria established by [120], as detailed in Table 3.1

Table 3.1: Kappa Score wise Agreement Level

Kappa Score Range	**Agreement Level**	**SA Score**	**NER Score**
Kappa ≤ 0	No agreement	–	–
$0 <$ Kappa ≤ 0.20	Slight agreement	–	–
$0.21 \leq$ Kappa ≤ 0.40	Fair agreement	–	–
$0.41 \leq$ Kappa ≤ 0.60	**Moderate agreement**	**0.5936**	–
$0.61 \leq$ Kappa ≤ 0.80	Substantial agreement	–	–
$0.81 \leq$ Kappa ≤ 1	**Almost perfect agreement**	–	**0.9283**
1	Perfect agreement	–	–

3.1.4 Data Splitting

In the realm of machine learning, the proper handling of data is crucial, encompassing the essential steps of training, validation, and testing. Therefore, post-annotation, we meticulously partitioned our dataset into three distinct sets—training, validation, and testing—utilizing a balanced ratio of 70-15-15. Given that our reviews encompass sentiments ranging from positive, negative, to neutral, it was imperative to ensure that the class distribution remained consistent across each set. Adhering to this meticulous approach, we successfully executed the data split, setting the stage for comprehensive model training and evaluation.

3.2 ELSA Model Development

The development of the Bangla ELSA model is another important phase of our research. In this phase, we collected our prepared dataset and used it for training and model development. The details of this phase are outlined below.

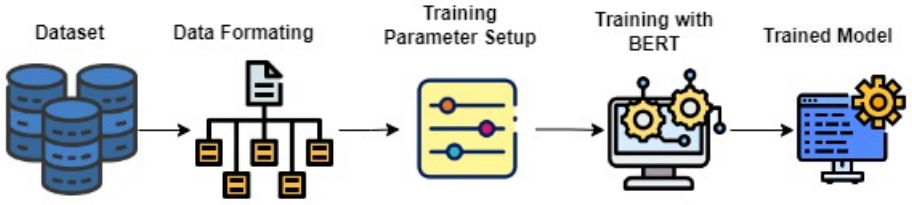

Figure 3.8: Trained Model Development Flow

3.2.1 Data Formatting

In the realm of classical machine learning, the process involves extracting features from the data and training the model based on these features. However, the paradigm shifts with pre-trained language models, eliminating the need for explicit feature vectors or selection methods. Instead, the focus shifts to formatting the data in a way that aligns with the language model's comprehension. The step-by-step data formatting process is illustrated in Figure 3.9.

The general procedure involves creating word tokens from the reviews, identifying entity words, and tagging them with numbers. In our case, where the sole entity category is "Product," the tag "1" is assigned when a product entity is detected. Otherwise, the words are tagged as "0". Subsequently, the word tokens undergo tokenization based on

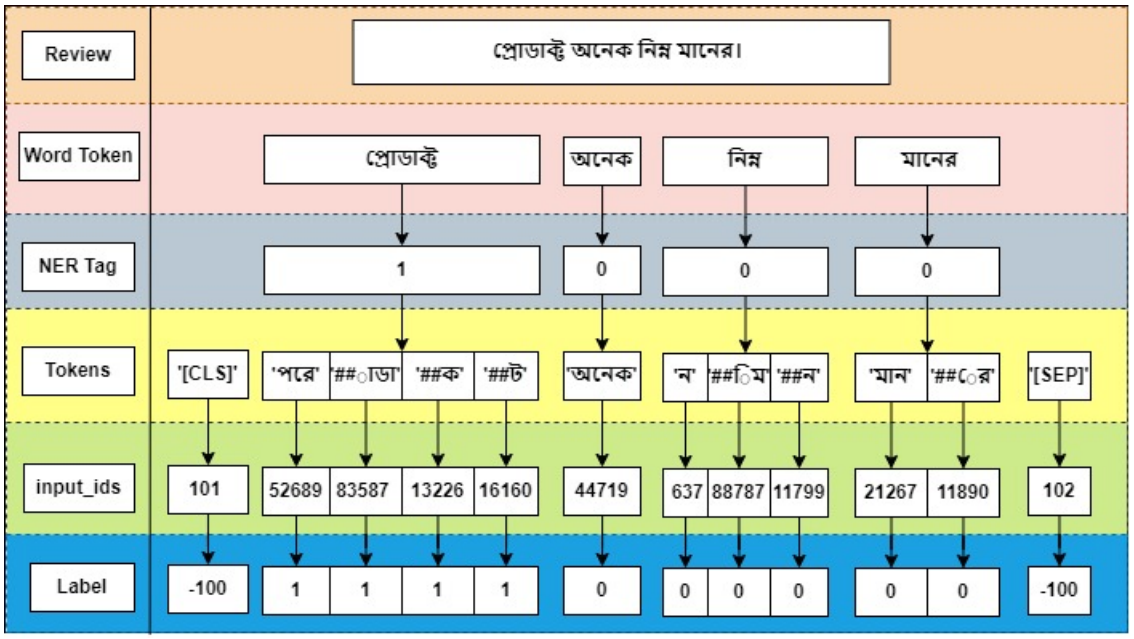

Figure 3.9: Dataset Formatting for NER Training

the pre-trained language model. This results in tokens and input tags, where each word is separated and assigned a numerical identifier. Labels are also generated, associating each input ID and token with the NER tag assigned earlier. This process ensures uniformity in the length of all reviews, achieved through either shortening or padding. When it comes to sentiment prediction, the transformation of reviews into corresponding input IDs and tokens is essential. The labels are assigned based on the annotated sentiment for the product. To maintain uniformity in length, a process of either shortening or padding is applied. This finalizes the structural preparation of the dataset for training and model development.

3.2.2 Training

Following the annotation and data split, our focus shifted to the crucial phase of training and validation for model development. In the context of ELSA, training is essential for both entity identification and entity sentiment.

1. **Training Parameter Setup:** For training the parameters related to training are very important. We used the following parameter values for training.

   A short description of the above-mentioned parameters is as follows.

Table 3.2: Parameter Setup for Taining

Name of the parameter	**Value**
Evaluation Strategy	epoch
Learning Rate	3×10^{-5}
Per Device Train Batch Size	16
Per Device Eval Batch Size	16
Number of Train Epochs	10
Weight Decay	0.01
Logging Steps	50

- **Evaluation Strategy:** This parameter determines when the model is evaluated during training. In this case, it is set to ”epoch,” indicating that the evaluation takes place after each epoch.
- **Learning Rate:** Specifies the step size during optimization, influencing the magnitude of updates to the model weights.
- **Per Device Train Batch Size:** Defines the number of training samples processed on each device in a single batch during training.
- **Per Device Eval Batch Size:** Specifies the number of evaluation samples processed on each device in a single batch during evaluation.
- **Number of Train Epochs :** Sets the total number of training epochs, representing the number of times the model goes through the entire training dataset.
- **Weight Decay:** Introduces regularization by penalizing large weights in the model, helping prevent overfitting.
- **Logging Steps:** Determines the frequency of logging training information, with updates displayed every 50 steps for monitoring progress.

These parameters collectively influence the training and performance of BERT-based models for NER and SA tasks. Adjusting them allows customization for optimal results based on the specific characteristics of the dataset and task requirements.

2. **Training:** Following the detailed steps of data formatting and configuring the training parameters, we start the training process. Using the pre-trained models aligned with the specified parameters, our data goes through the training phase. Initially, we utilized the pre-trained BERT language model for entity identification, leveraging the mbert-bengali-ner designed explicitly for Bangla NER. However, as the language model lacked training for Product-type entities, we proceeded to fine-tune our data specifically for product entity identification. For sentiment prediction, we employed the bangla-bert-base model, leveraging its capabilities to

understand and analyze sentiment in Bangla text. This model, designed for Bangla SA, played a crucial role in predicting the sentiment orientation of our annotated reviews, enhancing the overall effectiveness of our ELSA approach.

3.2.3 Model Development

After training, an essential validation step follows, resulting in the computation of key result metrics—precision, recall, F1-score, and accuracy for both NER and SA. This thorough approach ensures a comprehensive understanding of the model's performance. With the successful completion of these steps, our fine-tuned model is ready for testing and predictions, ready to provide valuable insights.

With our models in place, the next step is to put them to the test using sample reviews. This enables our models to showcase their capability in identifying entities and their associated sentiments within the given text. For example, when provided with a text, the model excels in recognizing product entity words and determining the sentiment orientation tied to those entities within the review. This testing phase ensures the practical applicability and effectiveness of our models in real-world scenarios.

Chapter 4

Dataset Properties and Analysis

In our comprehensive data collection process, we systematically gathered information from three leading online platforms, amassing a substantial dataset of 6,40,184 reviews. Through a meticulous three-step filtering approach, we carefully refined and annotated a final set of **10,000** reviews for ELSA. Each filtering step was strategically implemented to not only reduce the dataset size but also optimize the efficiency of the annotation process, minimizing both cost and time.

Figure 4.1: Product Category Word Cloud

4.1 Product Category

The primary focus of our data encompassed books, health products, and women's clothing, further branching into 149 categories. Representing the diversity among the three product types, we generated word clouds for 149 product categories selected from our extensive dataset. Figure 4.1 visually presents the word cloud analysis, offering insights into the prominent words associated with each product category.

4.2 Frequently Used Words

While the volume of data was significant, the unannotated nature of the reviews necessitated a judicious approach. Recognizing the complexity, time constraints, and cost implications associated with manual data annotation, we strategically employed filters. Initial emphasis was placed on language-based filtering, narrowing down the reviews to 2,59,513 by retaining only Bangla language reviews. To further streamline the dataset, we implemented a length filter, preserving reviews ranging from 5 to 30 words and resulting in 30,587 reviews. Following the word length filtering, we identified frequently used words in the reviews. We then selected the top words with a frequency greater than 50. Figure 4.2 showcases a word cloud illustrating these frequently used words.

Figure 4.2: Word Cloud of Frequently used Words in Reviews

4.3 Frequently Used Entity Words

Additionally, we analyzed frequently used words, identifying potential entity words from the list. Figure 4.3 presents a word cloud displaying the entity word list. Subsequently, we filtered our reviews using this word list, ensuring that each review contained at least one entity word. Annotators then utilized these entity words as a foundation for annotating entities within the reviews, supplementing the annotations with any additional entity words they identified.

Figure 4.3: Entity Words Word Cloud

4.4 Collected and Filtered Dataset Summary

Subsequently, we applied these selected words to filter our dataset, yielding a final subset of 12,429 reviews. From this filtered pool, a random selection of 10,000 reviews was made for annotation. The comprehensive statistical analysis and details of our collected and filtered dataset are meticulously presented in Table 4.1.

Table 4.1: Collected and Filtered Dataset Summary

	Rokomari & Wafilife	**Daraz**	**Total**
Product Type	Book	Health and Clothing	3 Types
Number of Sub-category	1	148	149
Number of Products	36,221	20,529	56,750
Total Review Collected	2,04,682	4,35,502	6,40,184
After Language Filtering	1,14,237	1,45,276	2,59,513
After word length Filtering	3,660	26,927	30,587
After Entity Word Filtering	1109	11,320	12,429
Final Selection	300	9700	**10,000**

- **Reason for selecting comparatively less data from Bookshops (Rokomari and Wafilife):** In our journey through SA Research in Bangla, we initially gathered Bangla reviews from Rokomari and Wafilife, leading to the publication of our research paper [8]. However, as we transitioned to ELSA in Bangla, we encountered a limitation. Book reviews exclusively featured a single product category, ”Book,” which posed challenges for our ELSA model in obtaining a diverse set of products. To address this, we expanded our data collection to Daraz, a comprehensive e-commerce platform, acquiring reviews spanning 148 categories, including 20,529 unique subcategories. This diversified dataset from Daraz provides a more robust foundation for the development of our ELSA model, prompting a deliberate

choice to limit the inclusion of reviews from online bookshops such as Rokomari and Wafilife.

4.5 Finally Annotated Dataset Summary

Our carefully selected and annotated dataset of 10,000 reviews now includes annotations for negative, positive, and neutral sentiments, and identified entity words. The detailed statistical analysis of this final annotated dataset is presented in Table 4.2.

Table 4.2: Annotated Dataset Summary

Annotated Review Dataset	10,000
Positive Sentiment Review	7,897
Negative Sentiment Review	980
Neutral Sentiment Review	1,123
Words in all Review	88,166
Highest Review Length (in words)	30
Lowest Review Length (in words)	5
Average Review Length (in words)	8.82
Highest No of Reviews for a Single Product	180
Highest No of Reviews' Product Name	Trimmer
No of Product	2,658
Entity Words	10,182
Non-Entity Words	77,984

4.6 Sentiment Distribution

The sentiment distribution extracted from our final dataset, as illustrated in Figure 4.4, distinctly reveals a prevalence of positive sentiments, constituting 78.97% of the reviews. Meanwhile, the remaining sentiments are almost evenly distributed between negative and neutral. This distribution pattern is a common occurrence in product reviews, where individuals tend to share their experiences more prominently when they hold positive sentiments. This behavior results in a higher frequency of positive reviews compared to negative and neutral ones.

Furthermore, Figure 4.9 illustrates the sentiment distribution of the dataset after undergoing the training, testing, and evaluation phases, following a 70-15-15 split ratio. This division ensures a balanced representation of sentiments across different sets, contributing to the robustness of our SA model.

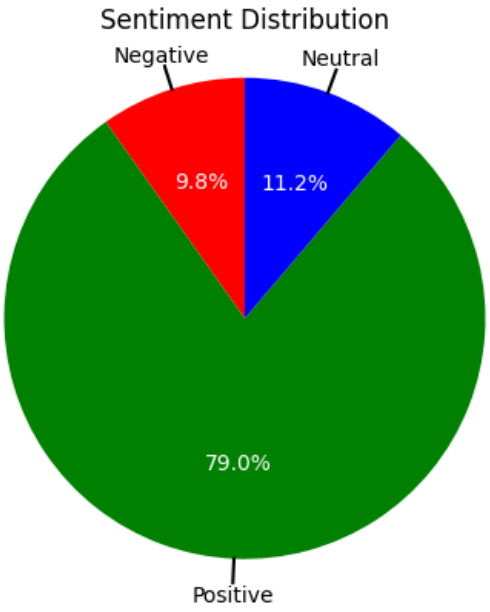

Figure 4.4: Sentiment wise Dataset Distribution

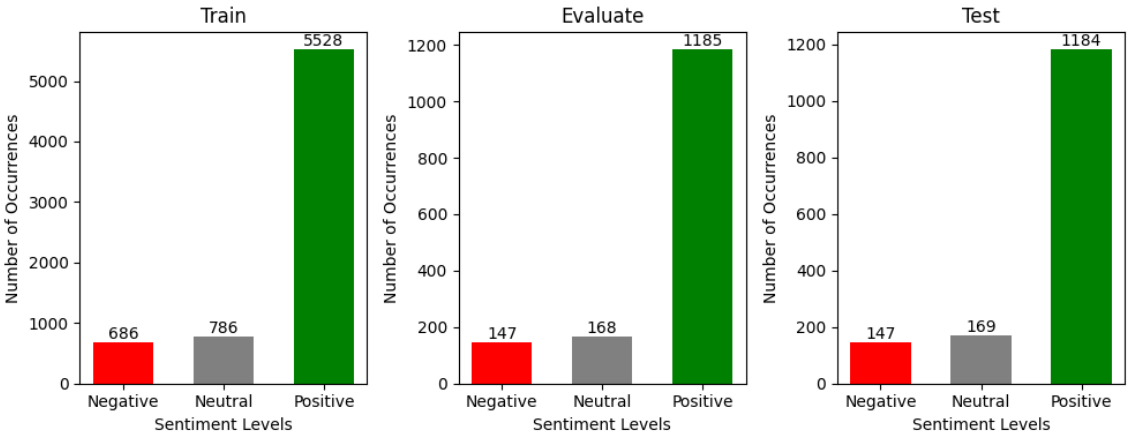

Figure 4.5: Data Split based on Sentiment

4.7 Entity and Non-Entity Word Distribution

Within the entirety of our review dataset, 11.55% of the words are identified as entity words, as depicted in Figure 4.6. This distribution reveals that the majority of the words within a given review are non-entity words. The prominence of non-entity words is a natural outcome, as individual reviews typically contain a higher proportion of general, non-specific language compared to specific entity-related terms.

Additionally, Figure 4.11 provides a visual representation of the distribution between entity and non-entity words after the dataset undergoes a split into a 70-15-15 ratio. This split ensures that the entity and non-entity words are proportionately represented across

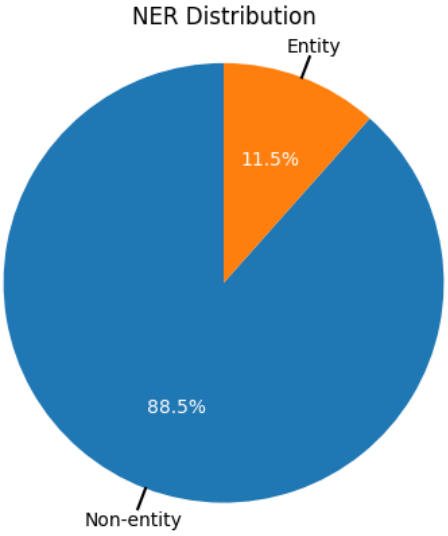

Figure 4.6: Entity and Non-Entity Word Distribution

the training, validation, and test sets, contributing to a balanced and comprehensive analysis of both types of words.

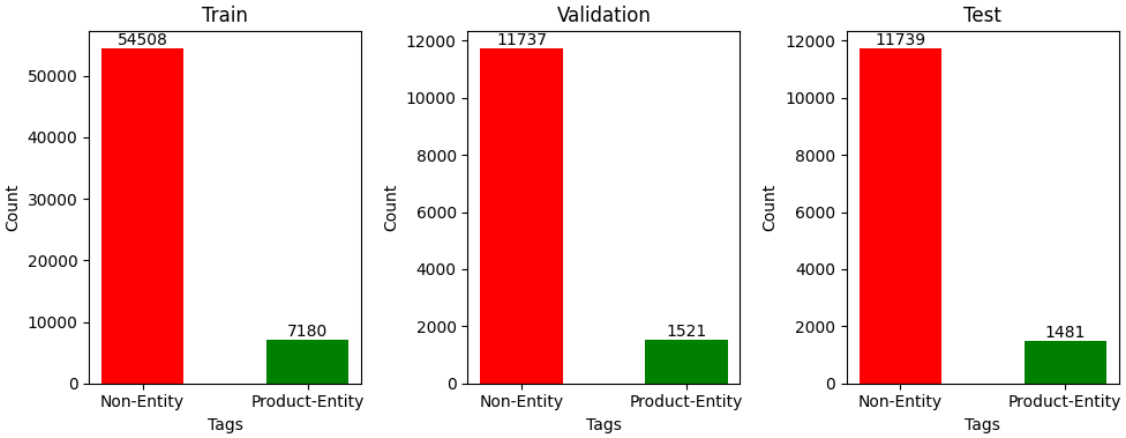

Figure 4.7: Data Split based on NER

4.8 Equitable Class Distribution via Downsizing

In light of our investigation, it became evident that our dataset exhibited a notable imbalance in class distribution. To address this issue comprehensively, we made a strategic decision to optimize the balance within our dataset through a meticulous process of data downsizing. In this endeavor, we carefully curated a dataset that ensures near equality in class distribution, thereby promoting a more representative and unbiased foundation

for our analysis. The ensuing figures depict the refined class distribution and entity allocation, underscoring the efficacy of our approach in achieving a more harmonized and equitable dataset.

4.8.1 Class Balanced Dataset Distribution

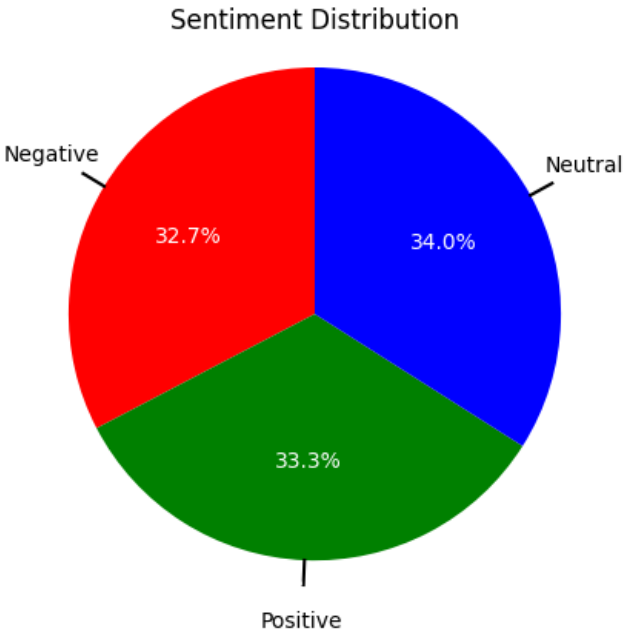

Figure 4.8: Class Balanced Dataset Sentiment Distribution

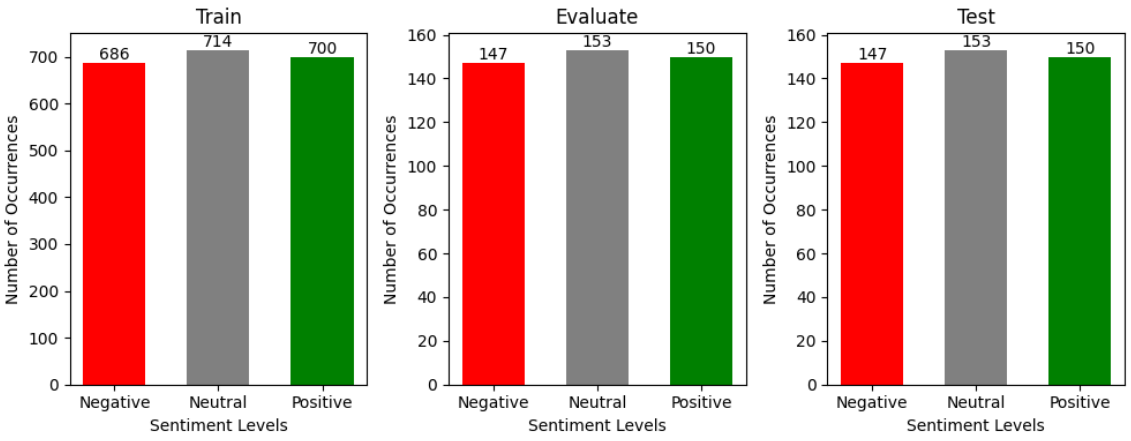

Figure 4.9: Data Split based on Sentiment (Class Balanced Dataset)

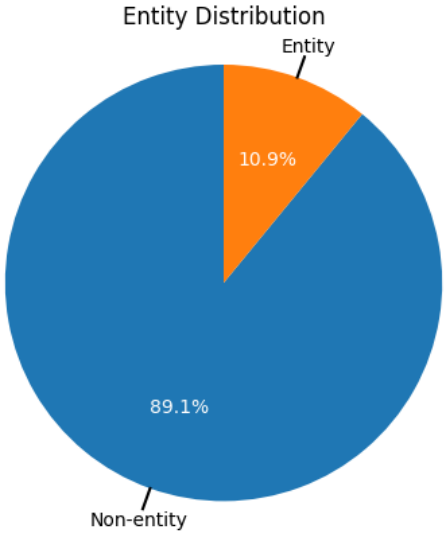

Figure 4.10: Class Balanced Dataset Entity Distribution

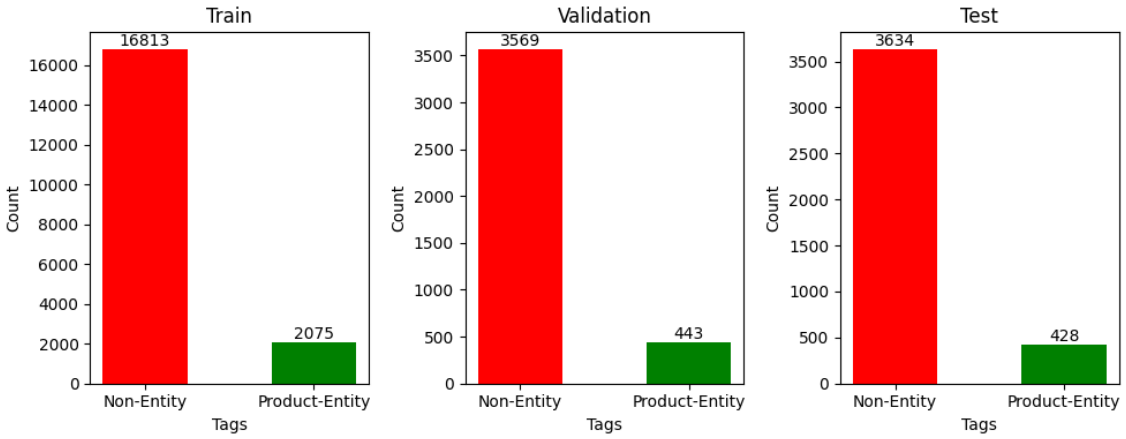

Figure 4.11: Data Split based on NER (Class Balanced Dataset)

4.8.2 Class Balanced Dataset Summary

The subsequent presentation encapsulates a concise overview of our meticulously balanced dataset, providing insights into both the sentiment distribution and the allocation of entities within the data.

Table 4.3: Class Balanced Dataset Summary

Class Balanced Dataset	3,000
Positive Sentiment Review	1,000
Negative Sentiment Review	980
Neutral Sentiment Review	1,020
Words in all Review	26,962
Entity Words	2,946
Non-Entity Words	24,016

This chapter comprehensively explores the statistical and numerical attributes of our dataset. From the initial stages of data collection, through the intricate processes of filtering, categorization, and identification of frequently used words and entity words, to the insightful distribution figures, each facet has been meticulously examined. The detailed analysis and findings presented in this chapter provide a clear and nuanced understanding of the essential properties inherent in our collected dataset, setting the foundation for subsequent stages of our research.

Chapter 5

Experimental Design

Following the construction of the dataset, a critical analysis is imperative to ensure the distinctiveness of dataset classes within machine learning frameworks and the effective recognition of product entities. To scrutinize these characteristics, we employed two pre-trained language models: bangla-bert-base for sentiment prediction and mbert-bengali-ner for entity identification. Additionally, the quality of the dataset is thoroughly examined through the utilization of evaluation metrics. This chapter delves into the thorough selection process of baseline models, pre-processing techniques, and the intricate discussion of evaluation metrics to provide a comprehensive overview of the methodology employed.

5.1 Baseline Models Selection

Baseline performance on the constructed dataset is evaluated using bangla-bert-base [113] and mbert-bengali-ner [121]. Both models are pre-trained using a large amount of unlabeled data in an unsupervised manner. The BERT-based models are chosen in this study for the following reasons:

- **Contextual Understanding:** Models based on BERT can capture the context of each word by considering both preceding and succeeding words in a sentence. Given the significance of accurate context comprehension for the sentiment classification and entity identification from reviews, these BERT-based models are expected to surpass conventional deep learning models like LSTM, BiLSTM, or unidirectional transformer models like OpenAI GPT [122], which only handle preceding tokens in the self-attention layers of the transformer.

- **State-of-the-Art Performance:** Prior research has indicated that achieving remarkable performance in diverse downstream tasks, such as sentiment classification and NER, can be accomplished through fine-tuning BERT-based models. These models, having been pre-trained on a substantial volume of unlabeled data using self-supervised learning, exhibit exceptional efficacy. In specific applications related to social media, such as SA of posts [123], categorization of spam and fake news [124], and identification of named entities [125], BERT-based models have showcased notable performance. Notably, these models address various limitations of prior state-of-the-art language models by adopting the transformer encoder instead of the recurrent neural network architecture.
- **Transfer Learning:** BERT leverages transfer learning, where knowledge gained from pre-training on a large corpus can be transferred to specific tasks. This enables the model to perform well on sentiment classification and NER even with limited labeled data for fine-tuning.

Hence, BERT's contextual understanding, pre-trained representations, ability to capture semantic relationships, and success in transfer learning make it a powerful choice for tasks like sentiment classification and named entity recognition.

5.1.1 Fine-tuning Classifiers

It is essential to adjust the pre-trained model weights in a task-specific way, particularly concerning tweet texts and their annotated labels. This adjustment is necessary to enhance the classification performance, given that the models are pre-trained using data from diverse sources. The subsequent section illustrates the fine-tuning process for input representation, followed by the experiment's training parameters.

5.1.1.1 Input Representation

Before giving input into pre-trained models for embedding, each tweet text undergoes a necessary formatting process. To facilitate the classification task, a singular vector representing the entire input sentence is essential. BERT-based models employ the WordPiece tokenizer, which divides the input sequence into either complete words or word pieces. In the case of complete words, one token string represents a word, while in the case of word pieces, a word is expressed through multiple token strings. The utilization

of word pieces aids in recognizing related words that share similar token strings, a pivotal factor for context comprehension. Additionally, special token strings are created during tokenization to signify task type, the beginning of the input sequence, mask, etc.

- '[SEP]' refers to the end of one input sequence and the beginning of another.
- '[CLS]' refers to the classification task.
- '[PAD]' is used to indicate the necessary padding.
- '[UNK]' stands for unknown token.

The classifiers utilized in this investigation necessitate uniform input sequence lengths, implying that each review should possess an identical number of tokens once converted into token strings. As a maximum token length of 512 is applied, if a review contains fewer than 512 tokens, additional [PAD]' tokens are appended to the end of the token sequence. During fine-tuning, there may be new input data not present in the pre-trained vocabulary. In such instances, the novel input substring is substituted with the [UNK]' token. Consequently, the final input vector for the models is crafted by converting the token strings into integer token IDs.

5.1.1.2 Hyper-parameters Selection

Fine-tuning and assessing the classifiers necessitated the division of the proposed dataset into three sets—train, validation, and test. For the train set, 70% of reviews from each class were randomly selected, while the remaining reviews were evenly distributed between the validation and test sets. Both the pre-trained models, with 768 hidden output states, were employed for fine-tuning. The fine-tuning process involved using the Categorical Cross-Entropy loss function with the AdamW optimizer [126], which employs fixed weight decay, unlike common Adam optimizer implementations [127]. The learning rate was set to 3×10^{-5}. Both models underwent supervised fine-tuning for 10 epochs with a training batch size of 16 on the proposed dataset to predict the sentiment analysis and product entities from reviews. The sentiment classification showed better results for mostly positive classes than negative and neutral. And the entity prediction model demonstrated good performance for any reviews. Figure 5.1 illustrates the process of sentiment classification and NER using the fine-tuned classifiers from a sample review.

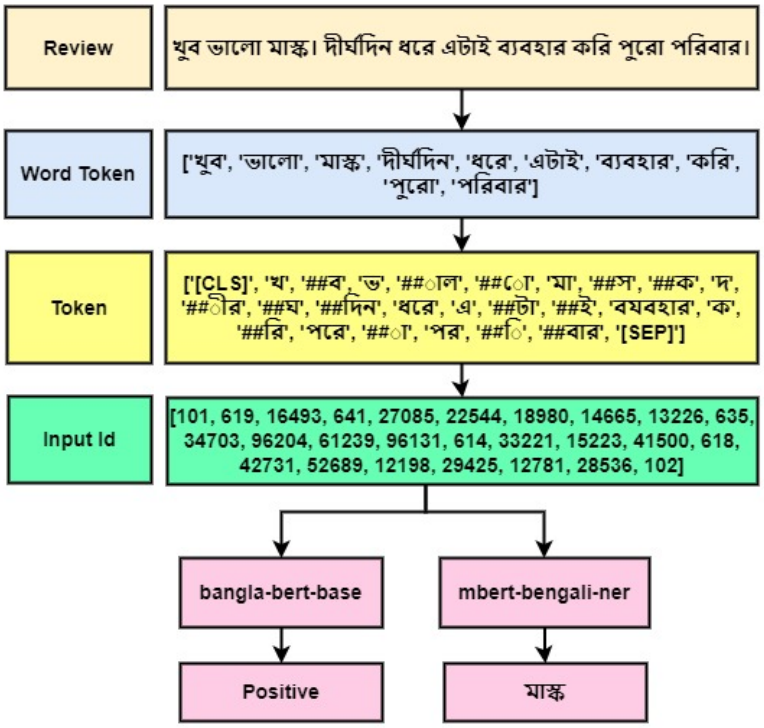

Figure 5.1: Sentiment and Entity Prediction from a Sample Review

5.2 Evaluation Metrics

The assessment metrics play a vital role in measuring the effectiveness of a predictive classifier, as noted by Sun et al. [128]. The selection of metrics is contingent on the dataset's characteristics, which can potentially result in misguided conclusions about the experiment. For instance, when evaluating an experiment on a significantly imbalanced dataset, metrics like accuracy, precision, or recall might yield conclusions that lack practical utility. In imbalanced datasets, achieving a high accuracy is possible without making meaningful predictions, as the majority of predictions may belong to the densely populated classes, as discussed by Leevy et al. [129].

Alternative commonly employed evaluation metrics such as precision and recall also come with inherent limitations. Precision, focused on the accuracy of the classification task, is determined solely by true positive and false positive values. Achieving a precision score of 1.0 is possible with just one accurate positive prediction. Conversely, recall, emphasizing completeness, relies solely on true positive and false negative values. Consequently, labeling all samples as positive would yield a recall of 1.0, even if precision remains quite low.

To address this challenge, this study employs the Receiver Operating Characteristic (ROC) curve and the area under the ROC curve (AUC) as evaluation metrics. These

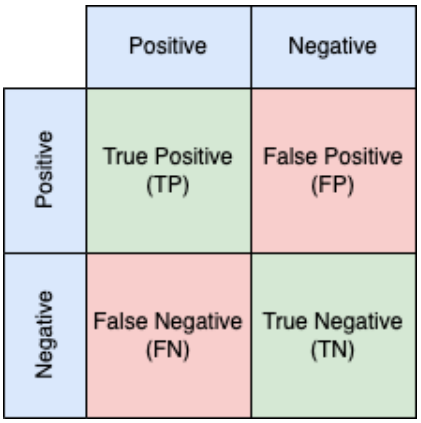

Figure 5.2: General Confusion Matrix

measures assess how effectively models differentiate between classes. The ROC curve, a diagnostic chart, computes the False Positive Rate (FPR) and True Positive Rate (TPR) for various predictions made by the model at different thresholds. This summary of the model's behavior helps analyze its capacity to distinguish between classes. The TPR indicates the proportion of correctly classified positive instances by the classifier, while the FPR signifies the proportion of incorrectly classified negative instances by the classifier. TPR and FPR are calculated as follows:

$$TPR = \frac{TP}{TP + FN} \tag{5.1}$$

$$FPR = \frac{FP}{TN + FP} \tag{5.2}$$

Definition of TP, FN, FP and TN can be derived from Figure 5.2.

The ROC curve is a probability curve that graphically depicts the True TPR against the FPR at different threshold values. It effectively distinguishes the 'signal' from the 'noise.' The AUC serves as a metric for the classifier's ability to differentiate between classes, providing a summary of the ROC curve. A model with no discriminatory power would be represented by a diagonal line between FPR 0 and TPR 0 (coordinate: 0,0) to FPR 1 and TPR 1 (coordinate: 1,1). Points below this line indicate models with less competence than none, while a flawless model would be situated in the upper-left corner of the plot.

Chapter 6

Results and Discussions

This chapter commences by showcasing baseline classification results obtained from both the class-imbalanced and balanced datasets. Following this, a brief comparative analysis is conducted to discern differences between the outcomes of these two datasets. The latter section of the chapter engages in a discussion, highlighting specific limitations of the study. Additionally, it explores potential avenues for future work that could further enhance and expand upon this research.

6.1 Classification Performance (Class Imbalanced Dataset)

The outcomes presented in Table 6.1 reveal the precision, recall, F1 score, and accuracy metrics for sentiment analysis (SA) and product entity recognition (NER) derived from our designed model using a class-imbalanced dataset. Notably, the results showcase that bangla-bert-base, applied to sentiment analysis, achieved an accuracy of 87.43% alongside an impressive F1 score of 86.74%. Additionally, employing mbert-bengali-ner for product entity recognition yielded noteworthy results, with an accuracy of 96.36% and a commendable F1 score of 87.92%. It is imperative to mention that these experiments were conducted within a computationally constrained environment, featuring relatively smaller batch sizes and fine-tuning limited to 10 epochs.

Table 6.1: Performance matrices of SA and NER (Imbalanced Dataset)

Task	Model	Precision	Recall	F1	Accuracy
SA	bangla-bert-base	0.8680	0.8743	0.8674	0.8743
NER	mbert-bengali-ner	0.8901	0.8708	0.8792	0.9636

Figure 6.1 displays the training loss and validation loss curve for sentiment analysis. The training and validation loss patterns provide insights over 10 epochs. Initially, the

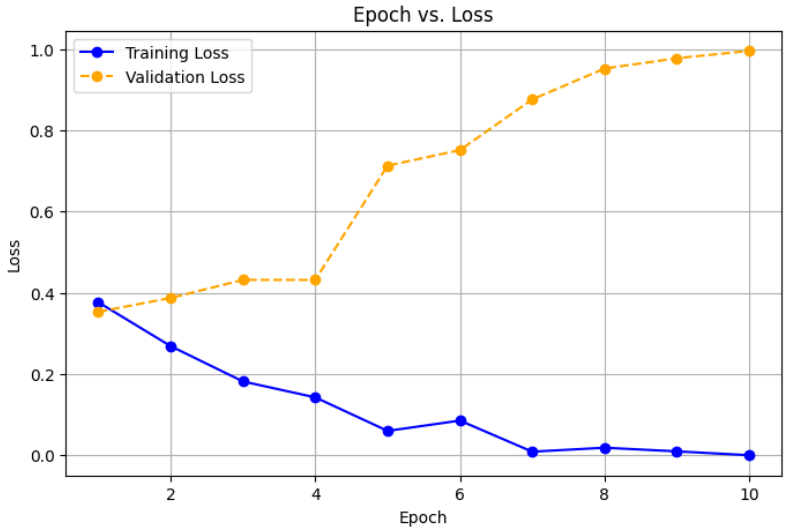

Figure 6.1: Loss Curves for Sentiment Analysis

model adapts well, but from Epoch 4, the increasing validation loss suggests potential overfitting, confirmed by a sharp drop in training loss at Epoch 7. Despite stable metrics in later epochs, the persistent rise in validation loss indicates ongoing overfitting.

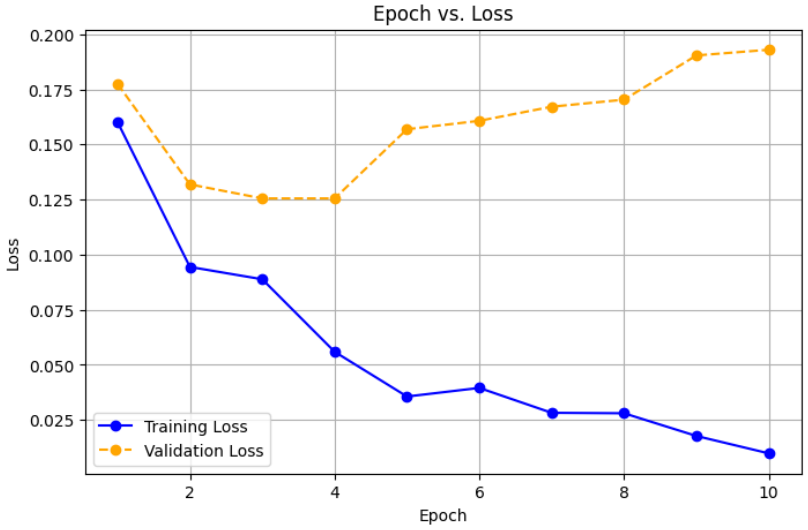

Figure 6.2: Loss Curves for Product Entity Identification

Figure 6.2 displays the training loss and validation loss curve for product entity identification. Over the 10 epochs, the training and validation loss dynamics are notable. In the initial stages, the model learns effectively with decreasing loss values. However, from Epoch 4, a rise in validation loss suggests potential overfitting, aligning with a sharp drop in training loss at Epoch 7. Subsequent epochs exhibit stable but high validation loss, indicating ongoing overfitting.

Table 6.2: AUC-ROC scores for SA

Model	Class Name	AUC-ROC Score
bangla-bert-base	Positive	0.9462
	Negative	0.9638
	Neutral	0.8320

The AUC-ROC score for the Negative sentiment class is very high, indicating excellent model performance in distinguishing between negative and non-negative sentiments. A score close to 1.0 suggests minimal false positives and false negatives for the Negative class. The AUC-ROC score for the Neutral sentiment class is good but slightly lower than for the Negative. This suggests that the model performs well in distinguishing between neutral and non-neutral sentiments, with some room for improvement. The AUC-ROC score for the Positive sentiment class is high, similar to the Negative class. This indicates strong performance in differentiating positive sentiments from non-positive sentiments, with a low rate of misclassification.

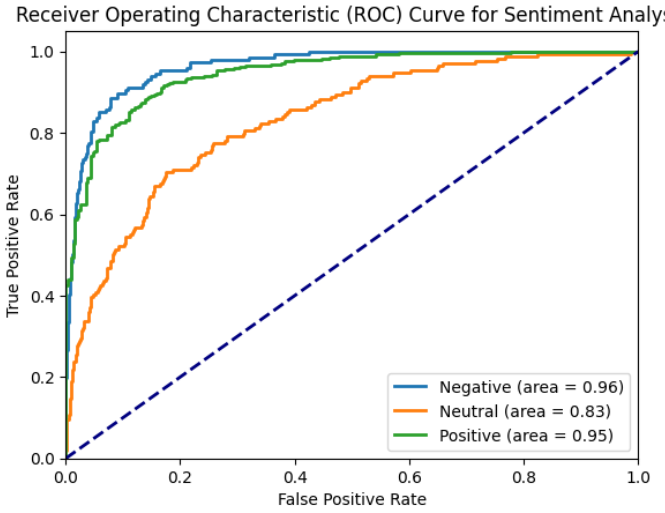

Figure 6.3: Class wise AUC-ROC curves for bangla-bert-base

In summary, the model shows excellent discriminatory power for both Negative and Positive sentiments, while its performance is slightly less robust for the Neutral sentiment class. Overall, these scores provide insights into how well the model separates each sentiment class based on the ROC curve, with higher scores indicating better performance.

In Table 6.3, a comprehensive display is presented, showcasing the true and predicted sentiments alongside associated entities for a selection of sample reviews. This table

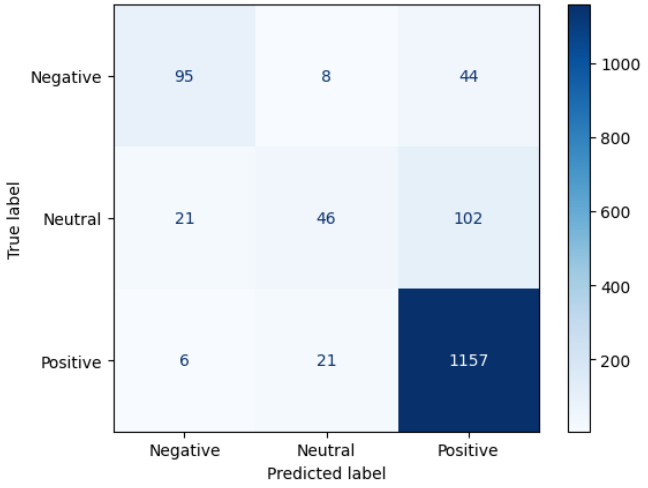

Figure 6.4: Sentiment Analysis Confusion Matrix

serves as a snapshot, offering insights into the sentiment analysis outcomes and the corresponding identified entities, providing a concise overview of the model's performance on these particular instances.

Table 6.3: Sentiment and Named Entity Prediction for Bangla Reviews

Review	**True Product Sentiment**	**True Product Entity Word**	**Predicted Product Sentiment**	**Predicted Product Entity Word**
এত বাজে জিনিসপত্র আপনারা ডেলিভারি দেন, যা কখনও ব্যবহার যোগ্য নয়। প্রতারক একটা দারাজ অনলাইন।	Negative	জিনিসপত্র	Negative	জিনিসপত্র
এই শপের পণ্য বরাবরই ভালো, দামে কম বাজারের চেয়ে। নিতে পারেন ।	Positive	পণ্য	Positive	পণ্য
ডেলিভারি ঠিক সময়ে হয় নাই কিন্তু কাপরের সাইজ একদম ঠিক আছে	Positive	কাপরের	Positive	কাপরের
টুপিটা কিনতেও পারেন, নাও কিনতে পারেন । রংটা সবাইকে মানাবে না ।	Neutral	টুপিটা	Positive	টুপিটা

6.2 Classification Performance (Class Balanced Dataset)

In light of the observed overfitting in our sentiment analysis and product entity recognition models, as depicted in Figures 6.1 and 6.2, there arises a critical need for a balanced dataset. The fluctuations in the validation loss curves, particularly the noticeable increase post-Epoch 3 in sentiment analysis, signify a challenge in the model's ability to generalize to unseen data. The abrupt improvement in training loss during Epoch 7 suggests potential overemphasis on the training set, resulting in a model that may not generalize well to real-world scenarios. Such overfitting highlights the limitations of

our experiments conducted with a class-imbalanced dataset, featuring relatively smaller batch sizes and a constrained fine-tuning period of 10 epochs.

Moving forward, achieving a more balanced distribution of classes in our dataset becomes imperative to enhance the model's reliability and foster robust generalization to diverse instances. We meticulously downsized our dataset to achieve balance, enhancing the reliability of our analyses. The outcomes from this balanced dataset reveal nuanced insights into sentiment analysis and product entity recognition. The refined dataset allows for better pattern recognition and improved overall performance, underscoring the importance of data preprocessing for reliable analyses.

The outcomes presented in Table 6.4 reveal the precision, recall, F1 score, and accuracy metrics for sentiment analysis (SA) and product entity recognition (NER) derived from our designed model using class balanced dataset.

Table 6.4: Performance matrices of SA and NER (Class Balanced Dataset)

Task	Model	Precision	Recall	F1	Accuracy
SA	bangla-bert-base	0.7466	0.7422	0.7437	0.7422
NER	mbert-bengali-ner	0.7809	0.7194	0.7489	0.9311

Figure 6.5 shows the training loss and validation loss curves for sentiment analysis.

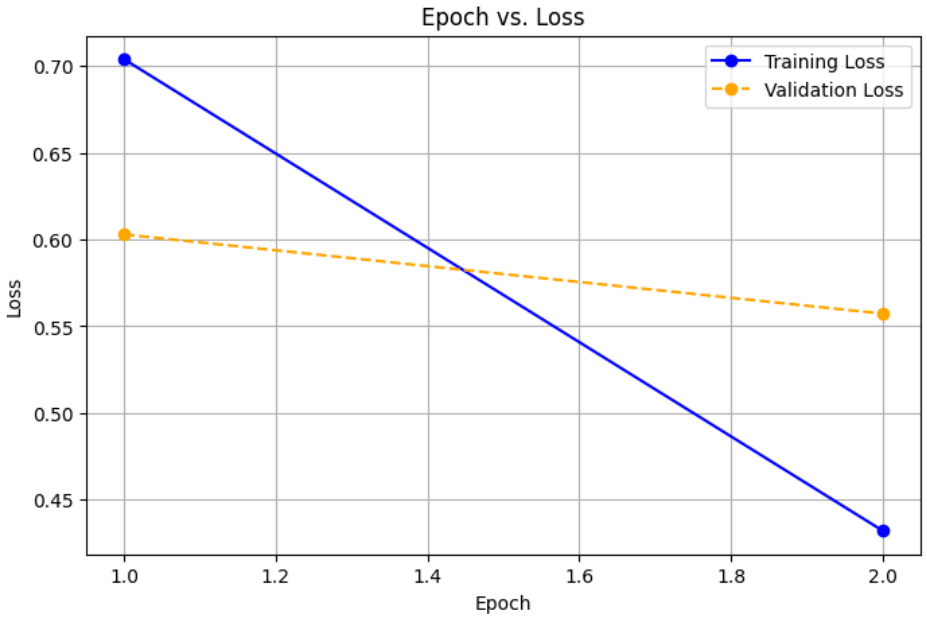

Figure 6.5: Loss Curves for Sentiment Analysis (Class Balanced Dataset)

Figure 6.6 shows the training loss and validation loss curves for product entity identification.

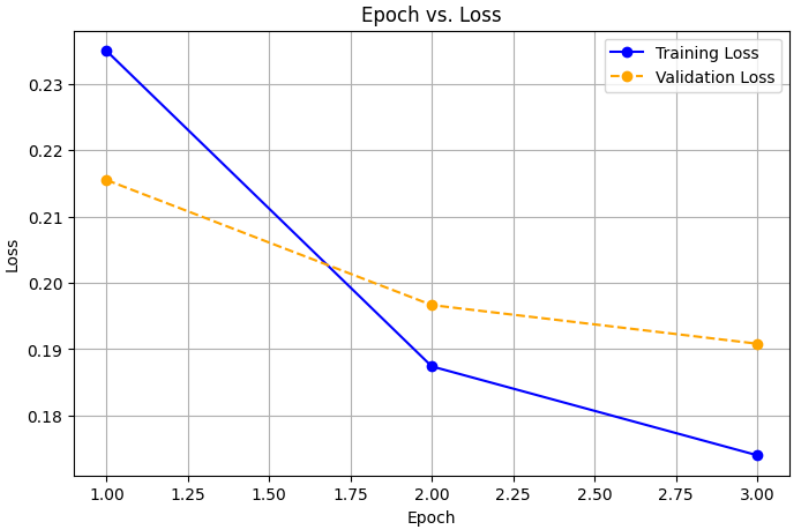

Figure 6.6: Loss Curves for Product Entity Identification (Class Balanced Dataset)

To prevent overfitting, the sentiment analysis model underwent training for only 2 epochs, while the named entity recognition model was trained for 3 epochs, terminating before the onset of overfitting.

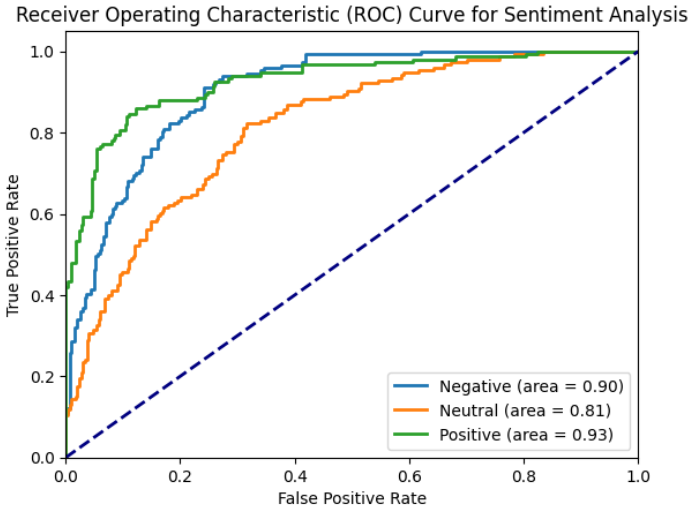

Figure 6.7: Class wise AUC-ROC curves for bangla-bert-base (Class Balanced Dataset)

Table 6.5: AUC-ROC scores for SA (Class Balanced Dataset)

Model	Class Name	AUC-ROC Score
bangla-bert-base	Positive	0.92706
	Negative	0.90260
	Neutral	0.81382

The AUC-ROC scores in table 6.5 and confusion matrix of Figure 6.8 reveal strong discriminative performance for the terminal sentiment classes (negative and positive),

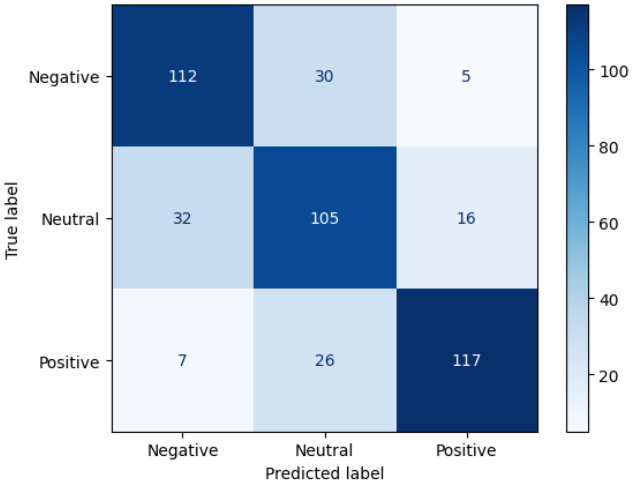

Figure 6.8: Sentiment Analysis Confusion Matrix (Class Balanced Dataset)

while the neutral class exhibits slightly lower but acceptable discriminative ability.

6.3 Imbalanced Vs Balanced Dataset Performance Analysis

The performance metrics for Sentiment Analysis (SA) and Named Entity Recognition (NER) under both imbalanced and balanced datasets are presented. Notably, the balanced dataset, achieved through downsizing and reduced epochs, displays a trade-off between precision, recall, F1 score, and accuracy compared to the imbalanced dataset.

Table 6.6: Imbalanced Vs balanced dataset performance matrices

Task	Model	Metric	Imbalanced Dataset	Balanced Dataset
SA	bangla-bert-base	Precision	0.8463	0.7466
		Recall	0.8653	0.7422
		F1	0.8461	0.7438
		Accuracy	0.8653	0.7422
NER	mbert-bengali-ner	Precision	0.9039	0.7810
		Recall	0.9533	0.7194
		F1	0.9279	0.7489
		Accuracy	0.9783	0.9311

For SA, the imbalanced dataset, utilizing bangla-bert-base, demonstrates higher precision (86.8%) and recall (87.4%), resulting in an F1 score of 86.74% and an accuracy of 87.4%. In contrast, the balanced dataset exhibits a decrease in precision (74.7%), recall (74.2%), F1 score (74.4%), and accuracy (74.2%).

Similarly, for NER with mbert-bengali-ner, the imbalanced dataset showcases precision (89%), recall (87.1%), F1 score (87.9%), and accuracy (96.4%). In contrast, the balanced dataset experiences a decline in precision (78%), recall (71.9%), F1 score (74.9%), and accuracy (93.1%).

While the imbalanced dataset may yield higher performance metrics, the balanced dataset provides a more reliable model by mitigating overfitting risks through downsizing and reduced epochs. The balanced dataset's generalization capability makes it a preferable choice despite the slight reduction in individual metrics. This ensures the model's robustness in handling unseen data, making it a more practical and dependable solution for real-world applications.

6.4 Limitations

This research encounters several limitations that merit consideration. Firstly, despite successfully addressing the initial challenge posed by imbalanced datasets, the transition to a balanced dataset introduces a distinct set of considerations. To rectify biases and enhance model robustness, a balanced dataset strategy was implemented. However, the consequential reduction in the overall dataset size emerges as a notable limitation. Downsizing to achieve balance inadvertently restricts the volume of training instances, potentially impacting the model's ability to generalize comprehensively.

Moreover, the dataset's exclusive reliance on Bangla reviews presents another limitation. The model's predictive performance is compromised when faced with reviews containing Banglish (a mix of Bangla and English) or Romanized Bangla (Bangla expressed in English letters). The model's proficiency in handling such linguistic variations may be limited, leading to potential inaccuracies in predictions.

Additionally, the study's focus on predicting product-type entities introduces a constraint. In scenarios where a single review encompasses multiple entities, the model prioritizes product entities, potentially overlooking other relevant entities and their associated sentiments. This limitation restricts the model's applicability to more diverse entity categories.

Furthermore, the annotation procedure employed in this research has certain limitations. Sentiment words are not explicitly annotated, as pre-trained language models used in this study are not designed to identify both sentiment words and entities concurrently.

If sentiment words were annotated and sentiments inferred based on these annotations, the sentiment orientation could be determined more accurately. This aspect represents an avenue for potential refinement in future research endeavors.

Chapter 7

Conclusion and Future Work

This study introduces a novel methodology for Entity-Level Sentiment Analysis (ELSA) on Bangla online product reviews. The framework includes a unique dataset that features labeled sentiment for reviews and labeled entities. Multiple annotators, with moderate inter-rater reliability, contributed to the dataset creation. The baseline classification results for the dataset were obtained through the fine-tuning of two contemporary pre-trained models. Specifically, bangla-bert-base was employed for sentiment analysis, while mbert-bengali-ner was utilized for product entity identification. This approach aims to fill the gap in ELSA datasets and research methodologies for the Bangla language. The detailed process and challenges in dataset creation may inspire researchers to gather similar corpora from various online platforms. The broader implications of Bangla ELSA include influencing business strategies, expediting marketing policy development, aiding product selection, and enhancing product improvement.

Future Works

In subsequent research endeavors in the future, several measures can be considered. One avenue involves expanding the dataset with additional samples, ensuring a more comprehensive representation of all classes even within the balanced dataset framework. This augmentation can contribute to a richer and more diverse training set, potentially leading to improved model generalization and robustness across various sentiments and named entities.

Leveraging more advanced transfer learning methods or exploring domain-specific pre-training could be beneficial. Fine-tuning not only on general-purpose pre-trained models but also on models trained specifically for sentiment analysis and product entity identification in the Bangla language might yield improved performance.

Exploring the realms of Banglish (Bangla-English mixed) and Romanized Bangla (Bangla word written in English letters) in future research presents intriguing opportunities to enhance the comprehensiveness of the study. For Banglish, the focus could delve into the intricate dynamics of code-switching, necessitating the development of models adept at interpreting contextual nuances arising from the fusion of Bangla and English. A crucial step involves creating a dedicated dataset reflecting the diverse linguistic patterns encountered in everyday Banglish communication.

Furthermore, addressing mixed-language data, where Bangla is written in English characters, calls for the development of hybrid models proficient in recognizing and processing both languages seamlessly. Integrating language identification mechanisms within these models is essential to distinguish between Bangla and English segments, ensuring targeted sentiment analysis and accurate entity identification. In essence, augmenting datasets to represent linguistic diversity, balancing language representations, and extending the study to incorporate multilingual models form integral components of future research endeavors in these linguistically diverse contexts.

It is pertinent to highlight that broadening the spectrum of annotated entity types holds the potential to significantly enhance the research's applicability. In doing so, the identification and annotation of additional entities, along with their correlated aspects, emerge as pivotal endeavors. This nuanced approach not only extends the applicability of the study but also enriches the depth of understanding by encompassing a more comprehensive range of entities and their interrelationships.

Appendix A

Appendix

Annotation Guidelines:

1. Annotate positive(2), negative(0) or neutral(1) sentiment as expressed for the NE tag in the review.
2. If there's no NE tag, please put one, the one that you think is most appropriate.
3. For multiple NE tags, only keep the most relevant one and remove others.
4. Kindly focus exclusively on annotating entities pertaining to product types, such as books, clothing, and health care products. In the presence of additional entities within the text, prioritize the annotation of the product entity and associate sentiments exclusively with the identified product entity.
5. Recognize and annotate sentiments expressed through sarcasm or irony. Acknowledge non-literal meanings that may indicate an opposite sentiment.
6. Annotate sentiments that are inferred but not explicitly stated. Interpret underlying emotions or attitudes based on context.
7. Refrain from making assumptions about unclear sentiments. Annotate based on information present in the text without personal interpretation.

Bibliography

[1] N. R. Bhowmik, M. Arifuzzaman, M. R. H. Mondal, and M. S. Islam, "Bangla text sentiment analysis using supervised machine learning with extended lexicon dictionary," *Natural Language Processing Research*, vol. 1, pp. 34–45, 2021. [Online]. Available: https://doi.org/10.2991/nlpr.d.210316.001

[2] Y. Y. Tan, C.-O. Chow, J. Kanesan, J. H. Chuah, and Y. Lim, "Sentiment analysis and sarcasm detection using deep multi-task learning," p. 2213–2237, Mar. 2023. [Online]. Available: http://dx.doi.org/10.1007/s11277-023-10235-4

[3] V. L. Durga and A. M. Sowjanya, "Sentiment analysis on bangla youtube comments using machine learning techniques," *Journal of Emerging Technologies and Innovative Research*, 2020.

[4] X.-y. Fu, C. Chen, M. T. R. Laskar, S. Gardiner, P. Hiranandani, and S. B. Tn, "Entity-level sentiment analysis in contact center telephone conversations," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*. Abu Dhabi, UAE: Association for Computational Linguistics, Dec. 2022, pp. 484–491. [Online]. Available: https://aclanthology.org/2022.emnlp-industry.49

[5] O. Toledo-Ronen, M. Orbach, Y. Katz, and N. Slonim, "Multi-domain targeted sentiment analysis," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Seattle, United States: Association for Computational Linguistics, Jul. 2022, pp. 2751–2762. [Online]. Available: https://aclanthology.org/2022.naacl-main.198

[6] N. R. Bhowmik, M. Arifuzzaman, and M. R. H. Mondal, "Sentiment analysis on bangla text using extended lexicon dictionary and deep learning algorithms," *Array*, vol. 13, p. 100123, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S259000562100059X

[7] E. Rønningstad, E. Velldal, and L. Øvrelid, "Entity-level sentiment analysis (ELSA): An exploratory task survey," in *Proceedings of the 29th International Conference on Computational Linguistics*. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, Oct. 2022, pp. 6773–6783. [Online]. Available: https://aclanthology.org/2022.coling-1.589

[8] M. Kabir, O. Bin Mahfuz, S. R. Raiyan, H. Mahmud, and M. K. Hasan, "BanglaBook: A large-scale Bangla dataset for sentiment analysis from book reviews," in *Findings of the Association for Computational Linguistics: ACL 2023*. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 1237–1247. [Online]. Available: https://aclanthology.org/2023.findings-acl.80

[9] F. Alshuwaier, A. Areshey, and J. Poon, "Applications and enhancement of document-based sentiment analysis in deep learning methods: Systematic literature review," *Intelligent Systems with Applications*, vol. 15, p. 200090, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2667305322000308

[10] M. Wankhade, A. C. S. Rao, and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," p. 5731–5780, Feb. 2022. [Online]. Available: http://dx.doi.org/10.1007/s10462-022-10144-1

[11] B. Liu, *Sentiment Analysis and Opinion Mining*. Springer International Publishing, 2012. [Online]. Available: http://dx.doi.org/10.1007/978-3-031-02145-9

[12] M. Karamibekr and A. A. Ghorbani, "Sentence subjectivity analysis in social domains," in *2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, vol. 1, 2013, pp. 268–275.

[13] A. Al Hamoud, A. Hoenig, and K. Roy, "Sentence subjectivity analysis of a political and ideological debate dataset using lstm and bilstm with attention and gru models," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 10, Part A, pp. 7974–7987, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1319157822002415

[14] E. Savinova and F. Moscoso Del Prado, "Analyzing subjectivity using a transformer-based regressor trained on naïve speakers' judgements," in *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, J. Barnes, O. De Clercq, and R. Klinger, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 305–314. [Online]. Available: https://aclanthology.org/2023.wassa-1.27

[15] A. Das, S. Bandyopadhyay, and J. Univers, "Subjectivity detection in english and bengali: A crf-based approach," 2009. [Online]. Available: https://api.semanticscholar.org/CorpusID:18335013

[16] M. Korayem, D. Crandall, and M. Abdul-Mageed, "Subjectivity and sentiment analysis of arabic: A survey," p. 128–139, 2012. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-35326-0_14

[17] M. Abdul-Mageed, M. Diab, and M. Korayem, "Subjectivity and sentiment analysis of Modern Standard Arabic," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, D. Lin, Y. Matsumoto, and R. Mihalcea, Eds. Portland, Oregon, USA: Association for Computational Linguistics, Jun. 2011, pp. 587–591. [Online]. Available: https://aclanthology.org/P11-2103

[18] A. AlKameli and M. Liakata, "Subjectivity analysis of arabic-english wikipedia," 2021. [Online]. Available: http://dx.doi.org/10.1049/icp.2021.0857

[19] V. Jha, N. Manjunath, P. D. Shenoy, and K. R. Venugopal, "Hsas: Hindi subjectivity analysis system," *2015 Annual IEEE India Conference (INDICON)*, pp. 1–6, 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:22309810

[20] Z. Zhang, Q. Ye, R. Law, and Y. Li, "Automatic detection of subjective sentences based on chinese subjective patterns," in *Cutting-Edge Research Topics on Multiple Criteria Decision Making*, Y. Shi, S. Wang, Y. Peng, J. Li, and Y. Zeng, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 29–36.

[21] H. Suzuki, Y. Miyauchi, K. Akiyama, T. Kajiwara, T. Ninomiya, N. Takemura, Y. Nakashima, and H. Nagahara, "A Japanese dataset for subjective and objective sentiment polarity classification in micro blog domain," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, and S. Piperidis, Eds. Marseille, France: European Language Resources Association, Jun. 2022, pp. 7022–7028. [Online]. Available: https://aclanthology.org/2022.lrec-1.759

[22] W. Li, "Subjectivity in japanese: A corpus-linguistic study," p. 202, Aug. 2019. [Online]. Available: http://dx.doi.org/10.5539/ijel.v9n5p202

[23] M. M. R. Mamun, O. Sharif, and M. M. Hoque, "Classification of textual sentiment using ensemble technique," Nov. 2021. [Online]. Available: http://dx.doi.org/10.1007/s42979-021-00922-z

[24] Y. Xu, H. Cao, W. Du, and W. Wang, "A survey of cross-lingual sentiment analysis: Methodologies, models and evaluations," p. 279–299, Jun. 2022. [Online]. Available: http://dx.doi.org/10.1007/s41019-022-00187-3

[25] A. Singh and K. Chatterjee, "A comparative approach for opinion spam detection using sentiment analysis," in *Proceedings of First International Conference on Computational Electronics for Wireless Communications*, S. Rawat, A. Kumar, P. Kumar, and J. Anguera, Eds. Singapore: Springer Nature Singapore, 2022, pp. 511–522.

[26] A. Rastogi and M. Mehrotra, "Opinion spam detection in online reviews," p. 1750036, Nov. 2017. [Online]. Available: http://dx.doi.org/10.1142/S0219649217500368

[27] A. Mukherjee and V. Venkataraman, "Spam detection : An unsupervised approach using generative models," 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:19034731

[28] A. Mewada and R. K. Dewang, "A comprehensive survey of various methods in opinion spam detection," p. 13199–13239, Sep. 2022. [Online]. Available: http://dx.doi.org/10.1007/s11042-022-13702-5

[29] R. Amin, M. M. Rahman, and N. Hossain, "A bangla spam email detection and datasets creation approach based on machine learning algorithms," Dec. 2019. [Online]. Available: http://dx.doi.org/10.1109/ICECTE48615.2019.9303525

[30] M. M. Uddin, M. Yasmin, M. S. H. Khan, M. I. Rahman, and T. Islam, "Detecting bengali spam sms using recurrent neural network," p. 325–331, 2020. [Online]. Available: http://dx.doi.org/10.12720/jcm.15.4.325-331

[31] T. Islam, S. Latif, and N. Ahmed, "Using social networks to detect malicious bangla text content," in *2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT)*, 2019, pp. 1–4.

[32] I. Tamhankar and A. Chaturvedi, "Classification of spam categorization on hindi documents using bayesian classifier," p. 8–13, Jan. 2019. [Online]. Available: http://dx.doi.org/10.14445/22312803/IJCTT-V66P102

[33] S. Kumar and T. D. Singh, "Fake news detection on hindi news dataset," *Global Transitions Proceedings*, vol. 3, no. 1, pp. 289–297, 2022, international Conference on Intelligent Engineering Approach(ICIEA-2022). [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666285X2200019X

[34] R. M. Saeed, S. Rady, and T. F. Gharib, "An ensemble approach for spam detection in arabic opinion texts," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 1, pp. 1407–1416, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1319157819307414

[35] A. M. Alkadri, A. Elkorany, and C. Ahmed, "Enhancing detection of arabic social spam using data augmentation and machine learning," *Applied Sciences*, vol. 12, no. 22, 2022. [Online]. Available: https://www.mdpi.com/2076-3417/12/22/11388

[36] R. A. Potamias, G. Siolas, and A. G. Stafylopatis, "A transformer-based approach to irony and sarcasm detection," p. 17309–17320, Jun. 2020. [Online]. Available: http://dx.doi.org/10.1007/s00521-020-05102-3

[37] R. Misra and P. Arora, "Sarcasm detection using news headlines dataset," *AI Open*, vol. 4, pp. 13–18, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666651023000013

[38] R. Anan, T. S. Apon, Z. T. Hossain, E. A. Modhu, S. Mondal, and M. G. R. Alam, "Interpretable bangla sarcasm detection using bert and explainable ai," in *2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC)*, 2023, pp. 1272–1278.

[39] A. Ghosh and K. Sarkar, "Irony detection in bengali tweets: A new dataset, experimentation and results," in *Computational Intelligence in Data Science*, A. Chandrabose, U. Furbach, A. Ghosh, and A. Kumar M., Eds. Cham: Springer International Publishing, 2020, pp. 112–127.

[40] S. K. Lora, G. M. Shahariar, T. Nazmin, N. N. Rahman, R. Rahman, M. Bhuiyan, and F. M. shah, "Ben-sarc: A corpus for sarcasm detection from bengali social media comments and its baseline evaluation," Jan. 2022. [Online]. Available: http://dx.doi.org/10.31224/osf.io/7yb4c

[41] A. Aggarwal, A. Wadhawan, A. Chaudhary, and K. Maurya, ""did you really mean what you said?" : Sarcasm detection in Hindi-English code-mixed data using bilingual word embeddings," in *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, W. Xu, A. Ritter, T. Baldwin, and

A. Rahimi, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 7–15. [Online]. Available: https://aclanthology.org/2020.wnut-1.2

[42] A. Rahma, S. S. Azab, and A. Mohammed, "A comprehensive survey on arabic sarcasm detection: Approaches, challenges and future trends," *IEEE Access*, vol. 11, pp. 18 261–18 280, 2023.

[43] Y. Okimoto, K. Suwa, J. Zhang, and L. Li, "Sarcasm detection for japanese text using bert and emoji," in *Database and Expert Systems Applications: 32nd International Conference, DEXA 2021, Virtual Event, September 27–30, 2021, Proceedings, Part I*. Berlin, Heidelberg: Springer-Verlag, 2021, p. 119–124. [Online]. Available: https://doi.org/10.1007/978-3-030-86472-9_11

[44] J. Cao, J. Li, M. Yin, and Y. Wang, "Online reviews sentiment analysis and product feature improvement with deep learning," p. 1–17, Aug. 2023. [Online]. Available: http://dx.doi.org/10.1145/3522575

[45] X. Deng, P. Zhang, Y. Xu, W. Zhou, D. Luo, Y. Shi, Z. Huang, and R. Jie, "Object-dependent document-level sentiment analysis based on sentence features," in *2023 2nd International Joint Conference on Information and Communication Engineering (JCICE)*, 2023, pp. 172–178.

[46] M. Rodríguez-Ibánez, A. Casánez-Ventura, F. Castejón-Mateos, and P.-M. Cuenca-Jiménez, "A review on sentiment analysis from social media platforms," *Expert Systems with Applications*, vol. 223, p. 119862, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417423003639

[47] S. Hao, P. Zhang, S. Liu, and Y. Wang, "Sentiment recognition and analysis method of official document text based on bert–svm model," Mar. 2023. [Online]. Available: http://dx.doi.org/10.1007/s00521-023-08226-4

[48] S. A. Purba, S. Tasnim, M. Jabin, T. Hossen, and M. K. Hasan, "Document level emotion detection from bangla text using machine learning techniques," Feb. 2021. [Online]. Available: http://dx.doi.org/10.1109/ICICT4SD50815.2021.9397036

[49] K. Islam, T. Yuvraz, M. S. Islam, and E. Hassan, "Emonoba: A dataset for analyzing fine-grained emotions on noisy bangla texts," in *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Association for Computational Linguistics, 2022, pp. 128–134.

[50] M. M. Rahman, R. Sadik, and A. A. Biswas, "Bangla document classification using character level deep learning," Oct. 2020. [Online]. Available: http://dx.doi.org/10.1109/ISMSIT50672.2020.9254416

[51] K. I. Islam, S. Kar, M. S. Islam, and M. R. Amin, "SentNoB: A dataset for analysing sentiment on noisy Bangla texts," in *Findings of the Association for Computational Linguistics: EMNLP 2021*. Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 3265–3271. [Online]. Available: https://aclanthology.org/2021.findings-emnlp.278

[52] M. Rahman, M. Pramanik, R. Sadik, M. Roy, and P. Chakraborty, "Bangla documents classification using transformer based deep learning models," *Proceedings of the IEEE*, pp. 1–6, 02 2021.

[53] J. Su, Q. Chen, Y. Wang, L. Zhang, W. Pan, and Z. Li, "Sentence-level sentiment analysis based on supervised gradual machine learning," Sep. 2023. [Online]. Available: http://dx.doi.org/10.1038/s41598-023-41485-8

[54] M. Bordoloi and S. K. Biswas, "Sentiment analysis: A survey on design framework, applications and future scopes," p. 12505–12560, Mar. 2023. [Online]. Available: http://dx.doi.org/10.1007/s10462-023-10442-2

[55] V. Shirsat, R. Jagdale, K. Shende, S. N. Deshmukh, and S. Kawale, "Sentence level sentiment analysis from news articles and blogs using machine learning techniques," p. 1–6, May 2019. [Online]. Available: http://dx.doi.org/10.26438/ijcse/v7i5.16

[56] J. Sun and M. Zhao, "Attention-based recursive autoencoder for sentence-level sentiment classification," in *2023 International Conference on Pattern Recognition, Machine Vision and Intelligent Algorithms (PRMVIA)*, 2023, pp. 272–276.

[57] T. Hossain, A. A. Nahian Kabir, M. Ahasun Habib Ratul, and A. Sattar, "Sentence level sentiment classification using machine learning approach in the bengali language," in *2022 International Conference on Decision Aid Sciences and Applications (DASA)*, 2022, pp. 1286–1289.

[58] M. Z. Haque, S. Zaman, J. R. Saurav, S. Haque, M. S. Islam, and M. R. Amin, "B-ner: A novel bangla named entity recognition dataset with largest entities and its baseline evaluation," *IEEE Access*, vol. 11, pp. 45 194–45 205, 2023.

[59] M. Hoang, O. A. Bihorac, and J. Rouces, "Aspect-based sentiment analysis using BERT," in *Proceedings of the 22nd Nordic Conference on Computational*

Linguistics. Turku, Finland: Linköping University Electronic Press, Sep.–Oct. 2019, pp. 187–196. [Online]. Available: https://aclanthology.org/W19-6120

[60] F. A. Naim, "Bangla aspect-based sentiment analysis based on corresponding term extraction," in *2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD)*, 2021, pp. 65–69.

[61] H. H. Do, P. Prasad, A. Maag, and A. Alsadoon, "Deep learning for aspect-based sentiment analysis: A comparative review," *Expert Systems with Applications*, vol. 118, pp. 272–299, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417418306456

[62] W. Zhang, X. Li, Y. Deng, L. Bing, and W. Lam, "A survey on aspect-based sentiment analysis: Tasks, methods, and challenges," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 11, pp. 11 019–11 038, 2023.

[63] M. E. Mowlaei, M. Saniee Abadeh, and H. Keshavarz, "Aspect-based sentiment analysis using adaptive aspect-based lexicons," *Expert Systems with Applications*, vol. 148, p. 113234, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417420300609

[64] N. Sultana, R. Sultana, R. I. Rasel, and M. M. Hoque, "Aspect-based sentiment analysis of bangla comments on entertainment domain," in *2022 25th International Conference on Computer and Information Technology (ICCIT)*, 2022, pp. 953–958.

[65] M. Ahmed Masum, S. Junayed Ahmed, A. Tasnim, and M. Saiful Islam, "Banabsa: An aspect-based sentiment analysis dataset for bengali and its baseline evaluation," in *Proceedings of International Joint Conference on Advances in Computational Intelligence*, M. S. Uddin and J. C. Bansal, Eds. Singapore: Springer Singapore, 2021, pp. 385–395.

[66] M. M. Samia, A. Rajee, M. R. Hasan, M. O. Faruq, and P. C. Paul, "Aspect-based sentiment analysis for bengali text using bidirectional encoder representations from transformers (bert)," 2022. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2022.01312112

[67] R. Ahuja, A. Chug, S. Kohli, S. Gupta, and P. Ahuja, "The impact of features extraction on the sentiment analysis," *Procedia Computer Science*, vol. 152, pp. 341–348, 2019, international Conference on Pervasive Computing Advances and Applications- PerCAA 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050919306593

[68] O. Sen, M. Fuad, M. N. Islam, J. Rabbi, M. Masud, M. K. Hasan, M. A. Awal, A. Ahmed Fime, M. T. Hasan Fuad, D. Sikder, and M. A. Raihan Iftee, "Bangla natural language processing: A comprehensive analysis of classical, machine learning, and deep learning-based methods," *IEEE Access*, vol. 10, pp. 38 999–39 044, 2022.

[69] S. Hira, M. R. Akhond, S. Chowdhury, A. K. Dipongkor, and S. M. Galib, "A systematic review of sentiment analysis from bengali text using nlp," *American Journal of Agricultural Science, Engineering, and Technology*, vol. 6, no. 3, pp. 150–159, 2022.

[70] S. Sazzed, "Bengsentilex and bengswearlex: creating lexicons for sentiment analysis and profanity detection in low-resource bengali language," *PeerJ Computer Science*, vol. 7, p. e681, 2021.

[71] H. Hota, D. K. Sharma, and N. Verma, "Lexicon-based sentiment analysis using twitter data," p. 275–295, 2021. [Online]. Available: http://dx.doi.org/10.1016/B978-0-12-824536-1.00015-0

[72] K. M. A. Hasan, M. Rahman, and Badiuzzaman, "Sentiment detection from bangla text using contextual valency analysis," in *2014 17th International Conference on Computer and Information Technology (ICCIT)*, 2014, pp. 292–295.

[73] V. Bonta, N. Kumaresh, and N. Janardhan, "A comprehensive study on lexicon based approaches for sentiment analysis," *Asian Journal of Computer Science and Technology*, vol. 8, no. S2, pp. 1–6, Mar 2019.

[74] S. Sazzed, "Development of sentiment lexicon in bengali utilizing corpus and cross-lingual resources," in *2020 IEEE 21st International Conference on Information Reuse and Integration for Data Science (IRI)*. IEEE, Aug 2020.

[75] S. Chowdhury and W. Chowdhury, "Performing sentiment analysis in bangla microblog posts," in *2014 International Conference on Informatics, Electronics Vision (ICIEV)*, 2014, pp. 1–6.

[76] M. Mahmudun, M. Tanzir, and S. Ismail, "Detecting sentiment from bangla text using machine learning technique and feature analysis," *International Journal of Computer Applications*, vol. 153, no. 11, pp. 28–34, 2016.

[77] R. C. Dey and O. Sarker, "Sentiment analysis on bengali text using lexicon based approach," in *2019 22nd International Conference on Computer and Information Technology (ICCIT)*, 2019, pp. 1–5.

[78] S. Akter and M. T. Aziz, "Sentiment analysis on facebook group using lexicon based approach," in *2016 3rd International Conference on Electrical Engineering and Information Communication Technology (ICEEICT)*, 2016, pp. 1–4.

[79] F. Rahman and et al., "An annotated bangla sentiment analysis corpus," in *2019 International Conference on Bangla Speech and Language Processing (ICBSLP)*, 2019, pp. 1–5.

[80] H. Ali, M. F. Hossain, S. B. Shuvo, and A. A. Marouf, "Banglasenti: A dataset of bangla words for sentiment analysis," in *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 2020, pp. 1–4.

[81] F. Hossain, "Banglasenti: A dataset of bangla words for sentiment analysis," https://github.com/fahad35/BanglaSenti-A-Dataset-of-Bangla-Words-for-Sentiment-Analysis, December 2023, accessed: 2023-12-06.

[82] M. A. Iqbal, A. Das, O. Sharif, M. M. Hoque, and I. H. Sarker, "Bemoc: A corpus for identifying emotion in bengali texts," *SN Computer Science*, vol. 3, no. 2, 2022.

[83] M. E. Khatun and T. Rabeya, "A machine learning approach for sentiment analysis of book reviews in bangla language," in *2022 6th International Conference on Trends in Electronics and Informatics (ICOEI)*, 2022, pp. 1178–1182.

[84] S. Sazzed, "Bengsentilex and bengswearlex: creating lexicons for sentiment analysis and profanity detection in low-resource bengali language," p. e681, Nov. 2021. [Online]. Available: http://dx.doi.org/10.7717/peerj-cs.681

[85] K. I. Islam, S. Kar, M. S. Islam, and M. R. Amin, "Sentnob: A dataset for analysing sentiment on noisy bangla texts," in *Findings of the Association for Computational Linguistics: EMNLP 2021*. Association for Computational Linguistics, 2021.

[86] A. Hassan, M. R. Amin, A. K. A. Azad, and N. Mohammed, "Sentiment analysis on bangla and romanized bangla text using deep recurrent models," in *2016 International Workshop on Computational Intelligence (IWCI)*, 2016, pp. 51–56.

[87] S. A. Mahtab, N. Islam, and M. M. Rahaman, "Sentiment analysis on bangladesh cricket with support vector machine," in *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*. IEEE, Sep. 2018.

[88] R. A. Tuhin, B. K. Paul, F. Nawrine, M. Akter, and A. K. Das, "An automated system of sentiment analysis from bangla text using supervised learning techniques," in *2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS)*, 2019, pp. 360–364.

[89] M. T. Akter, M. Begum, and R. Mustafa, "Bengali sentiment analysis of e-commerce product reviews using k-nearest neighbors," in *2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD)*, 2021, pp. 40–44.

[90] S. A. Kaiser, S. Mandal, A. K. Abid, E. Hossain, F. B. Ali, and I. T. Naheen, "Social media opinion mining based on bangla public post of facebook," in *2021 24th International Conference on Computer and Information Technology (ICCIT)*, 2021, pp. 1–6.

[91] M. R. H. K. Rahib, A. H. Tamim, M. Z. Tahmeed *et al.*, "Emotion detection based on bangladeshi people's social media response on covid-19," *SN COMPUT. SCI.*, vol. 3, p. 180, 2022.

[92] N. Banik, S. Chakraborty, H. Seddiqui, M. A. Azim, and M. H. H. Rahman, "Survey on text-based sentiment analysis of bengali language," 2019.

[93] M. R. Karim, B. R. Chakravarthi, J. P. McCrae, and M. Cochez, "Classification benchmarks for under-resourced bengali language based on multichannel convolutional-lstm network," *arXiv*, 2020.

[94] M. Hoq, P. Haque, and M. N. Uddin, "Sentiment analysis of bangla language using deep learning approaches," in *Communications in Computer and Information Science*. Springer International Publishing, 2021, pp. 140–151.

[95] M. K. Bashar, "A hybrid approach to explore public sentiments on covid-19," Apr. 2022. [Online]. Available: http://dx.doi.org/10.1007/s42979-022-01112-1

[96] N. Tabassum and M. I. Khan, "Design an empirical framework for sentiment analysis from bangla text using machine learning," in *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, 2019, pp. 1–5.

[97] M. A. Hasan, J. Tajrin, S. A. Chowdhury, and F. Alam, "Sentiment classification in bangla textual content: A comparative study," in *2020 23rd International Conference on Computer and Information Technology (ICCIT)*, 2020, pp. 1–6.

[98] J. Ding, H. Sun, X. Wang, and X. Liu, "Entity-level sentiment analysis of issue comments," in *2018 IEEE/ACM 3rd International Workshop on Emotion Awareness in Software Engineering (SEmotion)*, 2018, pp. 7–13.

[99] M. Luo and X. Mu, "Entity sentiment analysis in the news: A case study based on negative sentiment smoothing model (nssm)," *International Journal of Information Management Data Insights*, vol. 2, no. 1, p. 100060, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2667096822000040

[100] Z. Huang and Z. Fang, "An entity-level sentiment analysis of financial text based on pre-trained language model," in *2020 IEEE 18th International Conference on Industrial Informatics (INDIN)*, vol. 1, 2020, pp. 391–396.

[101] B. Jehangir, S. Radhakrishnan, and R. Agarwal, "A survey on named entity recognition — datasets, tools, and methodologies," *Natural Language Processing Journal*, vol. 3, p. 100017, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2949719123000146

[102] B. VeeraSekharReddy, K. S. Rao, and N. Koppula, "Named entity recognition using crf with active learning algorithm in english texts," in *2022 6th International Conference on Electronics, Communication and Aerospace Technology*, 2022, pp. 1041–1044.

[103] M. S. Ullah Miah, J. Sulaiman, T. B. Sarwar, S. S. Islam, M. Rahman, and M. S. Haque, "Medical named entity recognition (medner): A deep learning model for recognizing medical entities (drug, disease) from scientific texts," in *IEEE EUROCON 2023 - 20th International Conference on Smart Technologies*, 2023, pp. 158–162.

[104] A. Das, O. Sharif, M. M. Hoque, and I. H. Sarker, "Emotion classification in a resource constrained language using transformer-based approach," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*. Online: Association for Computational Linguistics, Jun. 2021, pp. 150–158. [Online]. Available: https://aclanthology.org/2021.naacl-srw.19

[105] P. G. Hoang, L. Thanh, and H.-L. Trieu, "VBD_NLP at SemEval-2023 task 2: Named entity recognition systems enhanced by BabelNet and Wikipedia," in *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, A. K. Ojha, A. S. Doğruöz, G. Da San Martino,

H. Tayyar Madabushi, R. Kumar, and E. Sartori, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 1833–1843. [Online]. Available: https://aclanthology.org/2023.semeval-1.253

[106] M. Z. Haque, S. Zaman, J. R. Saurav, S. Haque, M. S. Islam, and M. R. Amin, "B-ner: A novel bangla named entity recognition dataset with largest entities and its baseline evaluation," p. 45194–45205, 2023. [Online]. Available: http://dx.doi.org/10.1109/ACCESS.2023.3267746

[107] S. Mukherjee, M. Ghosh, Girish, and P. Basuchowdhuri, "MLlab4CS at SemEval-2023 task 2: Named entity recognition in low-resource language Bangla using multilingual language models," in *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, A. K. Ojha, A. S. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar, and E. Sartori, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 1388–1394. [Online]. Available: https://aclanthology.org/2023.semeval-1.192

[108] K. I. Islam, M. S. Islam, and M. R. Amin, "Sentiment analysis in bengali via transfer learning using multi-lingual bert," in *2020 23rd International Conference on Computer and Information Technology (ICCIT)*, 2020, pp. 1–5.

[109] A. Bhattacharjee, T. Hasan, W. Ahmad, K. S. Mubasshir, M. S. Islam, A. Iqbal, M. S. Rahman, and R. Shahriyar, "BanglaBERT: Language model pretraining and benchmarks for low-resource language understanding evaluation in Bangla," in *Findings of the Association for Computational Linguistics: NAACL 2022*, M. Carpuat, M.-C. de Marneffe, and I. V. Meza Ruiz, Eds. Seattle, United States: Association for Computational Linguistics, Jul. 2022, pp. 1318–1327. [Online]. Available: https://aclanthology.org/2022.findings-naacl.98

[110] N. J. Prottasha, A. A. Sami, M. Kowsher, S. A. Murad, A. K. Bairagi, M. Masud, and M. Baz, "Transfer learning for sentiment analysis using bert based supervised fine-tuning," *Sensors*, vol. 22, no. 11, p. 4157, 2022.

[111] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[112] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "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), J. Burstein, C. Doran, and T. Solorio, Eds. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[113] S. Sarker, "Banglabert: Bengali mask language model for bengali language understanding," 2020. [Online]. Available: https://github.com/sagorbrur/bangla-bert

[114] M. Tubishat, F. Al-Obeidat, and A. Shuhaiber, "Sentiment analysis of using chatgpt in education," in *2023 International Conference on Smart Applications, Communications and Networking (SmartNets)*, 2023, pp. 1–7.

[115] K. Kheiri and H. Karimi, "Sentimentgpt: Exploiting gpt for advanced sentiment analysis and its departure from current machine learning," 2023. [Online]. Available: https://arxiv.org/abs/2307.10234

[116] C. Dhivyaa, K. Nithya, G. Sendooran, R. Sudhakar, K. Kumar, and S. Kumar, "Xlnet transfer learning model for sentimental analysis," in *2023 International Conference on Sustainable Computing and Smart Systems (ICSCSS)*, 2023, pp. 76–84.

[117] N. Azhar and S. Latif, "Roman urdu sentiment analysis using pre-trained distilbert and xlnet," in *2022 Fifth International Conference of Women in Data Science at Prince Sultan University (WiDS PSU)*, 2022, pp. 75–78.

[118] A. K. Singh and A. Verma, "An efficient method for aspect based sentiment analysis using spacy and vader," in *2021 10th IEEE International Conference on Communication Systems and Network Technologies (CSNT)*, 2021, pp. 130–135.

[119] S. M. Yimam, H. M. Alemayehu, A. Ayele, and C. Biemann, "Exploring Amharic sentiment analysis from social media texts: Building annotation tools and classification models," in *Proceedings of the 28th International Conference on Computational Linguistics*, D. Scott, N. Bel, and C. Zong, Eds. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 1048–1060. [Online]. Available: https://aclanthology.org/2020.coling-main.91

[120] M. L. McHugh, "Interrater reliability: the kappa statistic," *Biochem. Med. (Zagreb)*, vol. 22, no. 3, pp. 276–282, 2012.

[121] S. Sarker, "mbert bengali ner," https://huggingface.co/sagorsarker/mbert-bengali-ner, 2023.

[122] M. Bilal and A. A. Almazroi, "Effectiveness of fine-tuned bert model in classification of helpful and unhelpful online customer reviews," p. 2737–2757, Apr. 2022. [Online]. Available: http://dx.doi.org/10.1007/s10660-022-09560-w

[123] A. S. Talaat, "Sentiment analysis classification system using hybrid bert models," Jun. 2023. [Online]. Available: http://dx.doi.org/10.1186/s40537-023-00781-w

[124] R. K. Kaliyar, A. Goswami, and P. Narang, "Fakebert: Fake news detection in social media with a bert-based deep learning approach," p. 11765–11788, Jan. 2021. [Online]. Available: http://dx.doi.org/10.1007/s11042-020-10183-2

[125] A. Agrawal, S. Tripathi, M. Vardhan, V. Sihag, G. Choudhary, and N. Dragoni, "Bert-based transfer-learning approach for nested named-entity recognition using joint labeling," *Applied Sciences*, vol. 12, no. 3, 2022. [Online]. Available: https://www.mdpi.com/2076-3417/12/3/976

[126] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. [Online]. Available: https://openreview.net/forum?id=Bkg6RiCqY7

[127] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *3rd International Conference on Learning Representations (ICLR)*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1412.6980

[128] Y. Sun, A. K. Wong, and M. S. Kamel, "Classification of Imbalanced Data: A Review," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 23, no. 04, pp. 687–719, 2009. [Online]. Available: https://www.worldscientific.com/doi/abs/10.1142/S0218001409007326

[129] J. L. Leevy, T. M. Khoshgoftaar, R. A. Bauder, and N. Seliya, "A survey on addressing high-class imbalance in big data," *Journal of Big Data*, vol. 5, no. 1, pp. 1–30, 2018. [Online]. Available: https://link.springer.com/article/10.1186/s40537-018-0151-6