Adapting Contextual Embedding to Identify Sentiment of E-commerce Consumer Reviews with Addressing Class Imbalance Issues

Md. Abdur Rakib Mollah*, Mir Md. Jahangir Kabir†, Md. Sazid Reza‡, Monika Kabir§
Department of Computer Science and Engineering
Rajshahi University of Engineering and Technology, Rajshahi, Bangladesh
University of Technology Sydney, Sydney, Australia†
Murdoch University, Perth, Australia§
Email: *rakib1703115@gmail.com, †mmjahangir.kabir@gmail.com, ‡dihansazid@gmail.com, §monika@vu.edu.bd

***Abstract*—Understanding consumer attitudes toward specific products is crucial for boosting sales in the e-commerce industry. To effectively target customers with popular products based on reviews, the classification of consumer feedback becomes imperative. However, classifying product reviews can be challenging, particularly when dealing with imbalanced data labels, which often result in suboptimal classification performance. This study builds upon previous efforts that utilized the Amazon Fine Food Reviews dataset for classification tasks. While these prior attempts showed promise, they were hindered by either poor embeddings or the prevalent class imbalance issue. In response, this research tries to solve these problems by using word embeddings with RoBERTa, a pre-trained transformer-based language model, to classify reviews. Additionally, the XGBoost classifier was implemented, along with embeddings from the language model. Losses were first calculated with equal weights for all class labels, and a re-weighted loss was subsequently adopted to balance the impact of each class on the loss function during training. The incorporation of RoBERTa and XGBoost, along with the class label re-weighting, contributed to improved capturing of intricate word relationships within reviews. As a result, this approach achieves significantly improved accuracy in both binary and multiclass classifications compared to earlier endeavors. Notably, it attained an impressive accuracy of 83.84% in multiclass classification and 93.29% in binary classification tasks, marking a substantial advancement in the field of consumer review analysis.**

***Index Terms*—Natural Language Processing, Transformers, Reviews, RoBERTa, XGBoost, Class Imbalance**

I. Introduction

The advancement of e-commerce has a significant impact on consumer behavior, leading to increased regularity in product purchases. In general, consumers tend to conduct an investigation of user feedback and ratings found on product websites as a means of making informed decisions on their purchases [1].

Individuals frequently articulate their opinions on the internet using various platforms such as social media, blogs, or e-commerce websites. On a daily basis, a substantial number of new evaluations emerge, necessitating the laborious process of manually discerning favorable and negative feedback, which in turn demands a significant staff. There exists a necessity for the development of an effective approach to ascertain the sentiments and extract meaningful insights.

Sentiment analysis allows the identification and resolution of positive and negative feelings linked to a product, benefiting customers as well as retailers.

Sentiments encompass a range of affective states, including emotions, opinions, attitudes, and feelings. The opinions expressed in evaluations are typically categorized as either positive or negative [2], or assigned numerical ratings within the range of 1 to 5. In the rating system, a rating of 1 indicates that the consumer is entirely unhappy with the item, while a rating of 5 signifies that the buyer is very pleased with the product.

The accuracy of sentiment analysis in product evaluations, such as those obtained from Amazon, is generally higher compared to sentiment analysis of social media data [3]. This is mostly because social media data is often characterized by a higher level of noise and lacks the structured nature typically found in product reviews.

Sentiment analysis has evolved over the years. Starting with simple rule-based approaches and lexicon-based methods, it has gradually embraced machine learning, diving into supervised learning and, later, deep learning with the introduction of neural networks and transformer models. This advancement has not only increased accuracy but has also permitted aspect-based analysis and emotion identification. Sentiment analysis currently thrives in the arena of e-commerce, dealing with brief, informal text data, and has become a critical tool to advertise or recommend good products to consumers.

The application of analytics can be employed to examine consumer feedback and sentiment to acquire valuable insights about customer requirements and preferences. One possible use of text analytics is the utilization by manufacturers to evaluate consumer reviews of their products to discover potential areas for enhancement [4].

This study observes and acknowledges the shortcomings of previous attempts and implies an effective strategy. We used the Amazon Fine Food Reviews Dataset and addressed its class

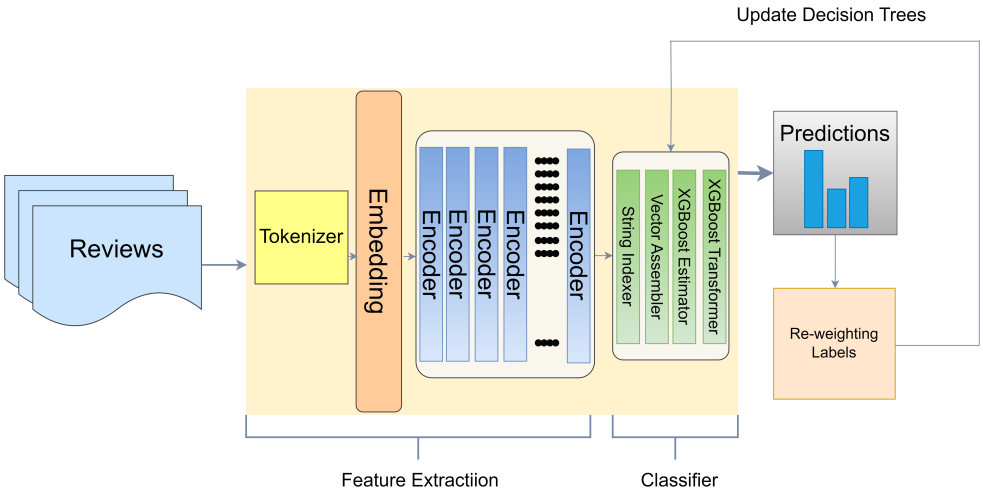

Fig. 1: Proposed RoBERTa+XGBoost Model with Re-weighting after Prediction

imbalance by re-weighting the class labels during training. We then used the RoBERTa classifier which was satisfactory but required more computational resources. So, for further improvement, the XGBoost classifier was introduced. As a result, this contribution outperformed earlier studies.

In Section II, we delve into a comprehensive review of related works in the field of sentiment analysis and product review classification. This section examines previous efforts, including those that utilized the Amazon Fine Food Reviews dataset, and identifies common challenges such as class imbalance issues and limitations related to embeddings. Subsequently, in Section III, we detail the methodology used in this study, which includes the use of RoBERTa-based word embeddings and the implementation of the XGBoost classifier. We also describe our approach to addressing class imbalance through re-weighted loss calculations. Following our methodology, we present a comprehensive analysis of our experiments in Section IV. We showcase the results obtained, including accuracy metrics for binary and multiclass classifications. In conclusion, the main outcomes of the study will be summarized, and the substantial progress made in evaluating consumer feedback using the approach employed will be emphasized.

II. Related Works

Over time, many studies have attempted to classify the Amazon Fine Food Reviews dataset by applying different techniques. A main aspect of sentiment analysis is word embeddings. Word embeddings are representations of words in a higher-dimensional vector space, where each word is mapped to a continuous-valued vector [5]. With the introduction of the transformer architecture, a way to derive contextual information between words is discovered. When producing embeddings for a certain word, it takes into account the words that surround it in a sentence. The ability of RoBERTa to comprehend context enables it to grasp subtleties and significance that might be overlooked by models such as FastText [6], which perceive words as separate entities.

The classifier then receives the embeddings to make predictions based on the reviews. In the past, a variety of classifiers have been used, from classic machine learning models like logistic regression, support vector machines, and random forests to more modern innovations like deep neural networks. The complexity of the sentiment analysis task, the quantity of data, the limitation of available hardware, and the desired level of accuracy- all influence the classifier that is selected. Each classifier brings its advantages and disadvantages to the table.

Ahmed et al. incorporated Linear SVC, Logistic Regression, and Naïve Bayes for sentiment analysis on the Amazon Fine Food Reviews dataset [7]. They used TF-IDF for feature extraction. Some attempts infused word negations and intensification in a BERT model for deriving the sentiment of reviews [8]. Zhao et al. used BERT for sentiment classification [9].

Thakkar et al. extended the baseline to modified RNN and GRU [10]. Yarkareddy et al. also used machine learning methods SVM, Random Forest, Naïve Bayes, and KNN. Iqbal et al. implemented deep learning-inspired long short-term memory and recurrent neural network-based models for sentiment classification and analysis [11].

Most of the previous works attempted either binary or multiclass classification but not both. And none of these works tried to resolve class imbalance issues properly which is beneficial in those datasets which could be biased towards one class. In the proposed method, re-weighting is integrated to put more emphasis on the negative reviews during training.

The loss function, also known as the cost or objective function, is crucial to model training, assessing how well predictions match targets [12]. The model's performance on mini-batches of data is measured by their combined loss during training. Each mini-batch instance should have equal weight,

although noisy or mislabeled instances possess higher loss values than clean ones in early training. This affects model accuracy since noisy cases outweigh clean ones due to their higher loss values.

This research acknowledges the limitations of previous studies and tries to resolve them by integrating contextual embedding, loss re-weighting for class imbalance, and XGBoost for classification. As a result, the proposed approach surpasses previous attempts.

III. Methodology

A. Dataset Description

Amazon Fine Food Reviews dataset is utilized for our research. This open-source dataset was collected from Kaggle [15]. The dataset comprises reviews of food products sourced from Amazon. The dataset has a time frame exceeding a decade, concluding in October 2012. Reviews encompass several components such as details on the product and user, evaluations in the form of ratings, and a textual review with the summary. Additionally, it encompasses evaluations from several additional categories available on the Amazon platform. The dataset consists of:

(a) Collection of reviews from Oct 1999 - Oct 2012
(b) 568,454 total reviews
(c) 256,059 individual users
(d) 74,258 distinct products
(e) Sentiments of the reviews in a range of 1 to 5

The dataset consisted of 10 columns in total, but only two columns were used: the review texts and the rating. Figure 2 shows the primary distribution of classes. There were 43 null values and after removing them, the final dataset included 5 classes, which were converted to positive and negative reviews. There are 486404 positive and 82007 negative reviews, which portray an imbalanced distribution of class labels.

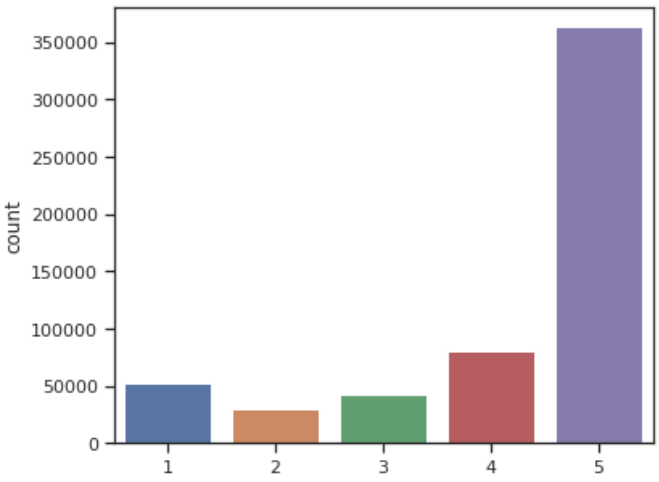

Fig. 2: Distribution of classes

B. Pre-trained Language Models

By utilizing their capacity to comprehend intricate contextual linkages in text, pre-trained language models like BERT, RoBERTa, and GPT have completely changed the way sentiment analysis is done [16]. Pretraining and fine-tuning are the two basic stages that these models go through. They are exposed to large text datasets during pretraining, which helps them learn the nuances of different languages. By fine-tuning, the model is made specifically for sentiment analysis using labeled data. The model takes a text input and transforms it into a high-dimensional vector representation or embedding that captures the subtle meaning. These pre-trained language models acquire their exceptional contextual knowledge through the attention mechanism, which allows them to weigh the value of various words and parts of the text when generating embeddings. The model is then given a classification layer on top to forecast emotion categories like positive, negative, or neutral. The process is facilitated by frameworks like Hugging Face's Transformers, which provide straightforward APIs for sentiment analysis using a variety of pre-trained models. However, performance is highly impacted by the model selection and dataset quality. For this research, RoBERTa has been utilized for contextual embedding.

C. Classification using RoBERTa

The RoBERTa (A Robustly Optimized BERT Pretraining Approach) model for text classification is a development over the earlier BERT language model [17]. The model improves the methods for pretraining by excluding BERT's next-sentence prediction and concentrating solely on the masked language model task [18]. Additionally a pioneer in the use of bigger batch sizes and extended training sequences, it effectively increases its exposure to various linguistic patterns. In terms of classification, there are clear similarities between BERT and RoBERTa. Both involve pretraining a model, adding a task-specific classification layer on top of it and then fine-tuning labeled data. Similar to previously developed language models, RoBERTa's classification layer adds fully connected neural network layers on top of the underlying model. It is modified to specific classification tasks by these additional layers, which transform its embeddings into predictions. Depending on the difficulty of the job, the architecture may vary; for simplicity, it may consist of a single dense layer with an activation function; for improved performance, it may consist of many dense layers, dropout, and attention mechanisms. A neuron with sigmoid activation for binary tasks or neurons matching classes with softmax for multiclass, producing probability scores, is the layer's ultimate configuration, which is in line with the class count. Task-specific architecture decisions are made in an effort to fully utilize RoBERTa's embeddings for precise prediction. But it also requires a large computational resource. Figure 3 and Figure 4 show the accuracy and loss respectively in both training and validation. It is notable that the model ran for only 3 epochs due to resource limitations.

TABLE I: Performance comparison of our method and earlier methods

Work Reference	No. of Reviews	Binary Accuracy	Multiclass Accuracy
Zhao et al. [9]	75,000	-	79.82%
Ahmed et al. [7]	1,64,015	87.00%	-
1qbal et al. [13]	25,000	87.00%	-
Agarwal et al. [14]	5,68,411	82.00%	-
RoBERTa	**5,68,411**	86.57%	-
	75,000	-	81.64%
RoBERTa+XGBoost	**5,68,411**	**93.29%**	-
	75,000	-	**83.84%**

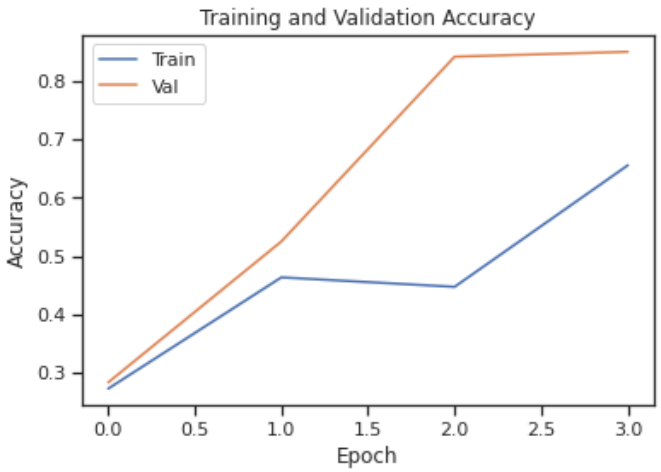

Fig. 3: Training accuracy and validation accuracy of RoBERTa classifier

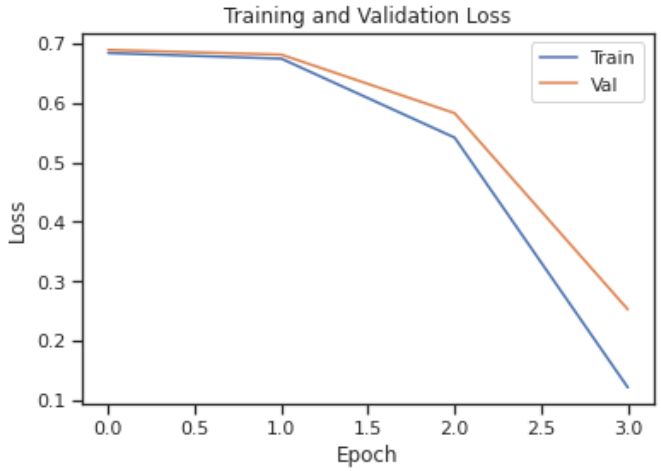

Fig. 4: Training loss and validation loss of RoBERTa classifier

D. XGBoost

XGBoost (Extreme Gradient Boosting) is an appealing option thanks to a number of its benefits [19]. It excels on smaller to medium-sized datasets, handling effectively without requiring a large volume of training data. There is a String Indexer component that converts categorical string data into numerical indices. Vector Assembler combines multiple feature columns into a single feature vector. XGBoost Estimator is the main algorithm used for training. The XGBoost Transformer applies a trained XGBoost model to generate predictions on new data. The integrated feature importance analysis assists in the selection of features and the comprehension of predictive aspects. The capacity of XGBoost to capture complicated non-linear interactions is well suited to instances where the data contains additive relationships. Furthermore, compared to deep neural networks, its computational efficiency enables faster training, making it a good choice for issues with constrained computational resources.

Figure 1 depicts how the XGBoost gets the embeddings or extracts features from words through RoBERTa and before updating its decision trees the re-weighting is done.

The RoBERTa classifier would require dense layers, hence to reduce the need for large computational resources, XGBoost is the alternative. Overall, it is a strong tool for a variety of machine learning applications due to its mix of interpretability, feature importance analysis, and efficiency.

E. Re-weighting for class imbalance

Machine learning uses re-weighting for class imbalance when one class in a classification issue has far less data than another. Biased models can hurt minorities due to class inequality[20]. To reduce this, re-weighting during model training gives the minority class a higher weight and the majority class a lower weight. The model now prioritizes the underrepresented class, helping it learn. To compute class weights using the `compute_class_weight()` function from the scikit-learn library, where the following formula is employed:

$$w_{\text{class}} = \frac{N}{C \times N_c}$$

where:

N : Total number of samples in the dataset
C : Number of classes in the classification problem
N_c : Number of samples in a specific class
w_{class} : Class weight assigned to the class

1) In RoBERTa: Tackling class imbalance involves adjusting the model's training process to give more weight to the minority classes. This is particularly important to ensure the classifier doesn't favor the majority class due to its higher frequency. By assigning higher weights to the minority class during the training phase, the model focuses on minimizing errors in predicting the outnumbered class.

It was achieved by modifying the loss function used for training. The adjusted loss function takes into account the class weights, emphasizing the importance of correct predictions for the minority class. As a result, the model becomes more adept at capturing patterns from both classes, ultimately leading to improved classification performance, particularly in the minority class. This technique enhances the classifier's ability to generalize on imbalanced datasets effectively.

2) In XGBoost: Re-weighting in algorithms like XGBoost involves changing sample weights in the training dataset. If the minority class has fewer samples, their weights can be higher than the majority class's. During optimization, the model concentrates more on correctly forecasting the minority class, where errors have a greater impact due to their increased weight.

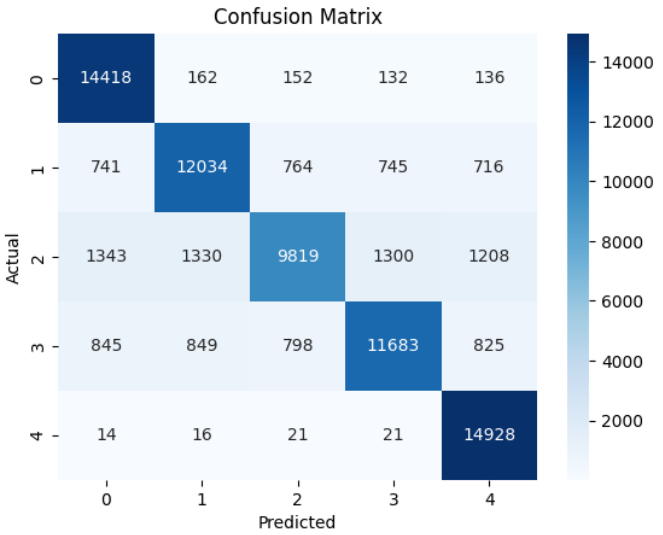

Fig. 5: Confusion Matrix of XGBoost for multiclass classification

IV. Experimental Analysis

A. Preprocessing

The input to the classifier would be the embeddings we got from the review texts of the dataset. But before that, the HTML links and stopwords were removed from the texts. We used RoBERTa Tokenizer for converting the texts. When the tokenizer is used, it first splits the texts into words or sequences and then maps words, phrases, or entire sentences to their corresponding dense vector representations. RoBERTa employs a Byte-Pair Encoding (BPE) tokenizer, a subword tokenization strategy. Padding tokens are added to make all tokenized sequences the same length. Following the processes of tokenization and padding, it creates an attention mask, which is basically a binary vector, that holds a length equivalent to that of the tokenized sequence. During both the training and inference stages, the employed attention mask identifies the relevant segments of the input sequence that require attention, while ignoring the irrelevant segments. The final output of the RoBERTa tokenizer consists of token IDs and an attention mask, and these components serve as the input features to the classifier model for further downstream tasks.

B. Experimental Design

This study was conducted on the Kaggle Cloud platform with a Tesla P-100 GPU and Python version 3.7. A batch size of 32 was used for training the RoBERTa classifier, which was the maximum size that could be operated due to hardware limitations. The model has 12 hidden layers with a parameter size of 124,647,939. Only 3 epochs were iterated for model execution because of hardware limitations. A dropout rate of around 0.01 was employed to prevent the model from relying too heavily on specific neurons. Both binary cross-entropy and categorical cross-entropy were applied, respectively, for the loss function. The weight decay hyperparameter for L2 regularization was set to 1e-5. A moderate threshold of 1.0 was used for gradient clipping. For binary classification, a layer with a sigmoid activation function was utilized. However, for multiclass classification, a softmax layer with as many output units as classes were employed. The learning rate was set at 0.000001, and it was optimized using an Adam optimizer. Momentum and RMSprop optimizations were excluded because both of their components are utilized by the Adam optimizer.

For XGBoost, K-Fold cross-validation was implemented by dividing the dataset into 20 subsets. In each iteration, one of the K folds was used as the test set, while the remaining K-1 folds were employed as the training set. With both classifiers, the weights for binary classification were 3.28 and 0.59 for the negative and positive classes, respectively. For multiclass classification, the weights were 2.17, 3.82, 2.64, 1.41, and 0.31.

C. Result Analysis

In most of the previous works, a smaller portion of the dataset was used. However, in this research, the models were managed to be run on the whole dataset for binary classification. The dataset was split into an 80:20 ratio, with the test set being completely autonomous and having no bearing on training. 20% of the train set was used for validation. After the split, the embedded training data was inserted into RoBERTa classifier, which gained an accuracy of 86.57% on binary classification, and 81.64% accuracy on multiclass. The classifier had dense layers for which it could be only run for 3 epochs. The performance might have been better if there could have been a few more epochs.

Hence, the alternative XGBoost is used which proved to be very useful. XGBoost was lightweight, and that is why once

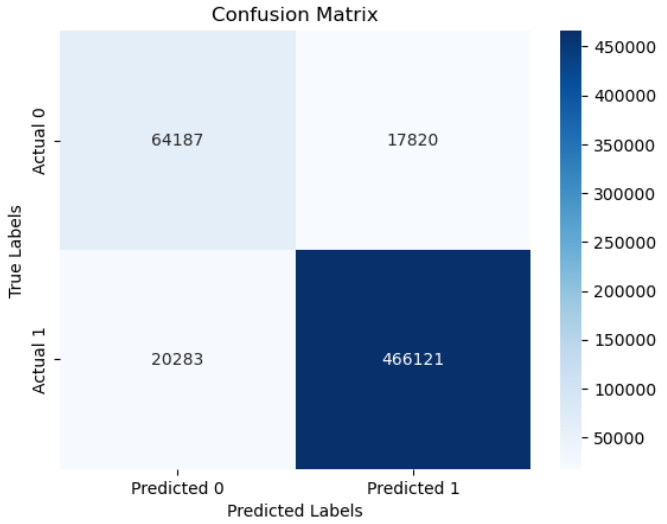

Fig. 6: Confusion Matrix of XGBoost for binary classification

the texts were transformed into embedding representations, they were fed into XGBoost. Thus the proposed approach took advantage of the language model for contextual embedding and the lightweight classifier for better performance. TABLE I compares the overall accuracy of our suggested model to past attempts. By a significant margin, this approach exceeded the performance of earlier works. It achieved 83.84% on multiclass and 93.29% on binary classification.

V. Conclusion

In this research, the entire Amazon Fine Food Reviews dataset from Kaggle was used for binary classification. RoBERTa was employed to encode the review texts. Both the dense layers of the language model and XGBoost were implemented as the classifier, where the performance of the latter was more effective. Weights were also assigned to minor classes to address class imbalance issues. The overall method yielded an accuracy of 83.84% for multiclass and 93.29% for binary classification. When the results were compared, it was evident that the strategy had significantly outperformed the earlier research. Additionally, the complete dataset was analyzed, as opposed to a subset, as was done in previous research. In the future, the aim is to ensemble multiple classifiers to further enhance accuracy.

References

[1] J. M. Kim, K. K.-c. Park, and M. M. Mariani, "Do online review readers react differently when exposed to credible versus fake online reviews?" *Journal of Business Research*, vol. 154, p. 113377, 2023.

[2] J. Piskorski, N. Stefanovitch, G. Da San Martino, and P. Nakov, "Semeval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multi-lingual setup," in *Proceedings of the the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, 2023, pp. 2343–2361.

[3] J. Hartmann, M. Heitmann, C. Siebert, and C. Schamp, "More than a feeling: Accuracy and application of sentiment analysis," *International Journal of Research in Marketing*, vol. 40, no. 1, pp. 75–87, 2023.

[4] S. S. Bawa, "Implementing text analytics with enterprise resource planning," *International Journal of Simulation–Systems, Science & Technology*, vol. 24, no. 1, 2023.

[5] R. A. Stein, P. A. Jaques, and J. F. Valiati, "An analysis of hierarchical text classification using word embeddings," *Information Sciences*, vol. 471, pp. 216–232, 2019.

[6] I. Santos, N. Nedjah, and L. de Macedo Mourelle, "Sentiment analysis using convolutional neural network with fasttext embeddings," in *2017 IEEE Latin American conference on computational intelligence (LA-CCI)*. IEEE, 2017, pp. 1–5.

[7] H. M. Ahmed, M. Javed Awan, N. S. Khan, A. Yasin, and H. M. Faisal Shehzad, "Sentiment analysis of online food reviews using big data analytics," *Hafiz Muhammad Ahmed, Mazhar Javed Awan, Nabeel Sabir Khan, Awais Yasin, Hafiz Muhammad Faisal Shehzad (2021) Sentiment Analysis of Online Food Reviews using Big Data Analytics. Elementary Education Online*, vol. 20, no. 2, pp. 827–836, 2021.

[8] B. Selvakumar and B. Lakshmanan, "Sentimental analysis on user's reviews using bert," *Materials Today: Proceedings*, vol. 62, pp. 4931–4935, 2022.

[9] X. Zhao and Y. Sun, "Amazon fine food reviews with bert model," *Procedia Computer Science*, vol. 208, pp. 401–406, 2022.

[10] K. Thakkar, S. Sharma, U. Chhabra, and M. C. Gupta, "Sentimental analysis on amazon fine food reviews," *International Journal of Scientific Research & Engineering Trends*, vol. 6, no. 1, pp. 318–324, 2020.

[11] S. Yarkareddy, T. Sasikala, and S. Santhanalakshmi, "Sentiment analysis of amazon fine food reviews," in *2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT)*. IEEE, 2022, pp. 1242–1247.

[12] G. Cui, Z. Liu, Y. Lin, and M. Sun, "Robust representation learning," in *Representation Learning for Natural Language Processing*. Springer, 2023, pp. 241–272.

[13] A. Iqbal, R. Amin, J. Iqbal, R. Alroobaea, A. Binmahfoudh, and M. Hussain, "Sentiment analysis of consumer reviews using deep learning," *Sustainability*, vol. 14, no. 17, p. 10844, 2022.

[14] A. Agarwal and S. Meena, "A comparative study of deep learning and machine learning algorithm for sentiment analysis," in *2022 3rd International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT)*. IEEE, 2022, pp. 1–7.

[15] J. J. McAuley and J. Leskovec, "From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews," in *Proceedings of the 22nd international conference on World Wide Web*, 2013, pp. 897–908.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[17] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[19] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[20] E. El-kenawy and E. SM, "A machine learning model for hemoglobin estimation and anemia classification," *International Journal of Computer Science and Information Security (IJCSIS)*, vol. 17, no. 2, pp. 100–108, 2019.