Classifying and Predicting The Rating Sentiment of Women's E-commerce Clothing Reviews: A Comparative Study Using SVM, ANN, and BERT Models

Immanuel Yabes
Computer Science Department,
BINUS Undergraduate Program,
Bina Nusantara University
Tangerang, Indonesia 15143
immanuel.yabes@binus.ac.id

Novita Aryanti
Computer Science Department,
BINUS Undergraduate Program,
Bina Nusantara University
Tangerang, Indonesia 15143
novita.aryanti001@binus.ac.id

Rayes Jordan Pradana
Computer Science Department,
BINUS Undergraduate Program,
Bina Nusantara University
Tangerang, Indonesia 15143
rayes.pradana@binus.ac.id

Karli Eka Setiawan
Computer Science Department, School
Of Computer Science
Bina Nusantara University
Jakarta, Indonesia
karli.setiawan@binus.ac.id

Muhammad Fikri Hasani
Computer Science Departmenet, School
Of Computer Science
Bina Nusantara University
Jakarta, Indonesia
Muhammad.fikri003@binus.ac.id

Abstract— **Product reviews are crucial for both customers and sellers in determining the quality of certain products. Alongside the review, the rating is also provided to give insight into the overall point of the product. As the review often becomes a consideration for a customer when buying a product online, the review needs to give an objective viewpoint. Therefore, it is important to determine whether the review is reliable. This could be seen both from the rating given by customers and the sentiment of the review. Machine learning and deep learning methods can be applied to classify the review sentiment effectively. The main objective was to check whether the review was considered positive or negative. The review was provided from women's clothing e-commerce by Nicapotato, which is available on Kaggle. This paper used both SVM and ANN for machine learning methods and BERT for deep learning method to determine the best method for the given dataset by identifying the most reliable and accurate model for classifying reviews. For the text vectorization, SVM and ANN models used TF-IDF technique while BERT model used sentenceBERT. The best method was determined quantitatively by comparing each model's accuracy and F1 Score results. Based on the accuracy and F1 Score result, it turned out that using TF-IDF for text vectorization with ANN gave the best performance compared to other models.**

Keywords—product reviews, rating, sentiment, SVM, ANN, BERT

I. Introduction

As internet use increases, various online media such as applications are also progressively becoming popular [1]. E-commerce applications are popular as customers can purchase items without having to go to the offline store. Moreover, most people prefer to shop online as they offer conveniences and various choices these days. A study stated that some people choose to buy items online because of the urge to save time from queuing [2]. The same study also stated that some people like to look for discounts from online stores as e-commerce offers more and various discounts than physical stores. This behavior lasts even after the pandemic [3]. Some e-commerce also focuses on selling certain categories. For people who are interested in fashion, there are several e-commerce fashion applications available. In fact, these applications are also used by people who are looking for clothes but are too busy to go to offline stores [4].

However, consumers often experience difficulties when shopping online. For fashion e-commerce, this happens when they need help to try on the clothes they are interested in directly [5]. This often leads the customer to return the purchased product if the clothes are not fit for them. Therefore, product reviews often become an alternative to help them determine whether the clothes suit their preferences.

Although most e-commerce applications are equipped with a rating system, the ratings given by consumers are sometimes not accurate with their comments [6]. There are times when a comment contains positive words, but the rating given alongside this comment by the customer is only 3 out of 5. Therefore, review rating prediction is needed. The predictions are expected to assist consumers and sellers in determining the quality of products. In addition, if the rating of products is accurate, the consumers could easily determine the right product to buy. Besides helping whether the product needs improvement, these predictions also could help sellers in making decisions for their next product development [6].

The experiment's focus was to apply the dataset to machine learning and deep learning model available in the field to help predict product review ratings. Furthermore, the models were also expected to help in classifying the sentiment of the comments. The results of each model were expected to help in determining which machine learning model could be considered most effective in predicting the rating and sentiment of each comment.

This paper aimed to contribute to sentiment analysis by comparing the performance of the proposed models. Using the available review dataset from clothing e-commerce, the models were expected to provide the best accuracy and F1 Score results. This paper could also have a positive impact by producing an accurate model in sentiment analysis.

The experiments from this paper were done by using several comparisons of existing machine learning models and deep *learning* models and comparing them all, the models proposed to be compared are SVM, ANN, and BERT. The models trained on the Women's E-Commerce Clothing review dataset released 5 years ago by Nicapotato on the Kaggle website to get output in the form of accuracy and F1 Score comparison. The dataset had 23486 rows of data with 10 feature variables. From the dataset, the pre-processing stage was carried out first and entered the feature selection stage, producing the 4 best features: Review Text, Rating, Recommended IND, and Positive Feedback Count.

In addition to the machine learning models, we compared two text vectorization techniques, which are tf-idf [7] and sentenceBert [8]. Afterwards, the dataset was tested on each machine learning model to get the model with the best accuracy and F1 score. This paper focused on the process of classifying labelled datasets. In this case, the label used as a reference was the user rating.

II. RELATED WORKS

Sentiment analysis is widely used to analyze and classify various kinds of reviews on the internet. The binary class sentiment, such as positive and negative reviews given by users regarding certain products, can be determined using machine learning models. For example, using a dataset of 25,000 film reviews from the internet, the results showed that Logistic Regression had the highest accuracy compared to using the NB and Multinomial NB [9]. Besides, using a dataset of reviews obtained from the travel website TripAdvisor.com, DT and C 4.5 had the best performance with large feature sets compared to other proposed models such as SVM, ANN, NB, and K-NN models [10]. Looking at another experiment, the results of using a reviews dataset from the e-commerce platform daraz.com.bd with Multinomial Naïve Bayes, Logistic Regression, SVM, RF, K-NN, and DT models show that the model can provide different results based on the language of the review [11]. Machine learning models could also work to determine the sentiment from multiclass classification problems. Logistic Regression, Multinomial Naïve Bayes, Linear SVC, Logistic Regression (Tuned), and Linear SVC (Tuned) models could be used on a product review dataset to determine whether the review has positive, neutral, or negative sentiment [12]. The experimental results mentioned show that the accuracy and performance of the model were affected by the dataset used. A research using E-commerce Customer Reviews tried on figuring sentiment analysis using several well-known machine learning such as XGBoost,, Random Forest, SVM, and LightGBM found that LightGBM gave best performance compared to other models [13].

A model could be combined with another model (hybrid) to do sentiment analysis. An experiment showed that the proposed method called Hybrid Sequential Binary Classification (HSBC) performed better on 12,000 Kindle e-book reviews obtained by Amazon.com compared with several models such as Naïve Bayes, Logistic Regression, SVM, CNN, LSTM using as the dataset [14]. In addition, there is also a method called Global Optimization-based Neural Network (GONN) where the performance of RF, Naïve Bayes, and SVM uses 5,269 product reviews from the beauty category on Amazon.com [15]. This proves that the hybrid model can be a solution and could be further developed.

In addition to classifying and analyzing sentiment from reviews, machine learning models can also make rating predictions. Using the Bagging, RF, J48, IBK, and Naïve Bayes models with a dataset in the form of a list of Hollywood films in 2018 and their ratings at IMBd, it was found that the RF model gave the best results [16]. In another case, using the RF, XGBOOST, and Logistic Regression models on a dataset in the form of 71,045 reviews of 1,000 assorted products, the result was that the RF model gave the best results [17].

Deep learning is also applied in further applications to help analyze sentiment from reviews, such as using RNN, GRU [18], CNN, LSTM [19], and BERT [20]. Deep learning can also improve performance in providing rating predictions from reviews compared to only machine learning models [21]. Deep learning is also often combined to obtain even higher performance [19]. In a study that compares tf-idf and word2vec it is shown that tf-idf performs better than word2vec by a small margin [7]. Comparing Word2Vec, TF-IDF, and BERT word embeddings for subjectivity analysis using corpora from Shopee Product Reviews and Wikipedia, word embeddings from BERT worked well [8].

III. RESEARCH METHODOLOGY

A. Datasets

The datasets used in this research was Women's E-Commerce Clothing Reviews which was obtained from Kaggle. This dataset, which has 23,486 rows and 10 feature, contains approximately 23,000 Customer Reviews and Ratings. The features Clothing ID, Age, Title, Review Text, Rating, Recommended IND, Positive Feedback Count, Division Name, Department Name, and Class Name are all included in each row, which represents a customer review. This study mainly concentrated on Review Text, Rating, and Recommended IND variables.

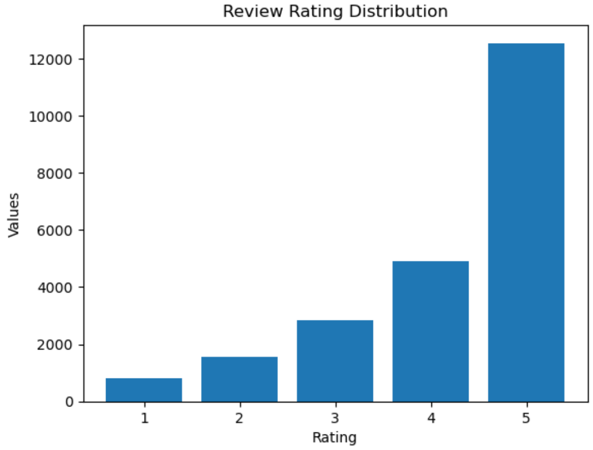

Fig. 1. Rating Count in the Dataset

B. Text Vectorization

a) TF-IDF

In tasks involving natural language processing and information retrieval, the text vectorization method known as TF-IDF is frequently used [7]. It is intended to convey the significance of a term both within and across a set of documents. Terms in a document are given weights according to their frequency (TF) and rarity (IDF) in the corpus using the TF-IDF algorithm. The TF component of the TF-IDF counts how often a term appears in a document. It is determined by tallying up the instances of each term in a document and dividing the result by the total number of terms in the document. According to the theory behind TF, terms that are used more frequently within a document are probably more significant or indicative of its content.

The IDF element of TF-IDF quantifies a term's rarity within the corpus. It is beneficial to give terms with greater discrimination across the entire corpus of documents a higher weight. The ratio between the total number of documents in the corpus and the number of documents containing the term is taken as the logarithm to determine the IDF. IDF avoids giving extremely high weights to terms that appear in only a few documents by using the logarithm instead.

A term's TF and IDF weights are calculated by multiplying their respective values. As a result, terms that occur frequently in a document but infrequently across the corpus will have higher TF-IDF weights, indicating that they are significant terms in that particular document. When text is vectorized using TF-IDF, each document is represented as a vector with each dimension standing for a different term in the corpus. The value in each dimension corresponds to the term's TF-IDF weight in the document. After that, different machine learning algorithms can use the resulting vectors as input.

The TF-IDF approach is particularly beneficial for tasks like text mining, information retrieval, and document classification. The TF-IDF method makes it possible to compare and analyze textual data effectively by giving higher weights to terms that are both common within a document and uncommon across the corpus. It is crucial to remember that TF-IDF is just one of many text vectorization techniques available, and that the task and dataset specificity can affect how effective it is. Word embeddings (such as Word2Vec, GloVe) and neural network-based techniques have also grown in popularity recently.

b) sentenceBERT

Sentence-BERT is a text vectorization method that specializes in producing superior sentence embeddings. Sentence embeddings are detailed numerical models that capture the semantic content of sentences and are used for a variety of natural language processing (NLP) tasks, including classification, clustering, and sentence similarity. Sentence-BERT works at the sentence level, taking into account the context and meaning of the entire sentence, as opposed to conventional word-level embeddings like Word2Vec or GloVe, which represent individual words. To produce sentence embeddings, it makes use of a pre-trained BERT model, a cutting-edge transformer-based neural network architecture.

Sentence-BERT's main goal is to harness the power of contextualized word representations from BERT and modify them for tasks at the sentence level. By anticipating masked words in sentences, BERT is initially trained on a large corpus to learn contextualized word embeddings. Sentence-BERT enhances the pre-trained BERT model by focusing on particular sentence-level tasks, such as semantic similarity or sentence classification. Sentence-BERT modifies the BERT architecture during fine-tuning to create sentence embeddings. To create a fixed-size sentence representation, a pooling layer is frequently added to the top of the BERT model. Simple mean pooling, maximum pooling, or a more sophisticated pooling mechanism like the Sentence Transformer pooling can all be used as this layer's pooling technique. The outcome is a fixed-length vector representation of the input sentence that captures its semantic meaning. To create sentence embeddings for new sentences, the Sentence-BERT model must first be refined. The model takes a sentence as input, runs it through a modified version of the BERT architecture, applies the pooling layer, and outputs the corresponding sentence embedding. After that, these embeddings can be applied to numerous downstream NLP tasks [8].

To produce sentence embeddings, Sentence-BERT modifies the BERT architecture during fine-tuning. The BERT model is frequently topped with a pooling layer to produce a fixed-size sentence representation. This layer's pooling method can be any of the following: simple mean pooling, maximum pooling, or a more complex pooling mechanism like the Sentence Transformer pooling. The result is a sentence that has a fixed-length vector representation that accurately captures its semantic meaning. The Sentence-BERT model must first be improved before new sentence embeddings can be produced. A sentence is input into the model, which then processes it using a modified version of the BERT architecture, applies the pooling layer, and outputs the corresponding sentence embedding. These embeddings can then be used for a variety of downstream NLP tasks [20].

C. *Model Architectures*

a) Support Vector Machine (SVM)

SVM is a supervised learning method used for classification where the model seeks the best possible surface to separate the samples from the predicted classes [10]. In this case, the predicted classes were positive and negative rating reviews. This model portrays the data from the dataset, each having “n” number of features plotted as points in a n-dimensional space separated into categories by a clear margin broadest possible, called a hyperplane. Afterwards, the data items are labelled with the class based on which side of the hyperplane they fall [22][22]. Choosing a decision boundary with a maximum margin between points from both classes is necessary for selecting the best class. The kernel used in this research was Radial Basis Function (RBF) kernel which the equation:

$$K(X, X') = \exp\left(-\frac{\|X - X'\|^2}{2\sigma^2}\right) \qquad (1)$$

Where X, X' represents the data training set which shows the dataset’s feature vectors. The squared Euclidean difference between the two feature inputs is represented by $\|X - X'\|^2$, and σ is a free parameter [23].

To implement the SVM model, this research used the SVM module from the sklearn library to implement the rating classification for the training.

b) Artificial Neural Network (ANN)

ANN is a mathematical or computational model that takes after the composition and operation of biological neural networks. Due to information passing through the network throughout the learning phase, whether internal or external, it is an adaptable system that adjusts its structure [10]. An artificial neural network is a feed-forward neural network if the connections between the units do not form a directed loop. Information moves unidirectionally across this network, possibly traveling through hidden nodes as it moves straight from input nodes to output nodes without any loops or cycles. A supervised learning technique called the backpropagation algorithm involves propagation and weight update. Backpropagation algorithms compare the output values to the proper response and then feed the error back across the network. Until the network performance is adequate, the two steps are repeated. After applying this method for a sufficient number of training cycles, the network is often going to converge to a state where the error in computation is small.

This research used feed-forward neural network as the network structure with sigmoid as the activation function. The model used 10 input layers, 3 hidden layers with 15 nodes of each layer, and 5 output layers to represent 1-star, 2-star, 3-star, 4-star, and 5-star rating labels.

c) BERT

BERT is a deep learning-based pretraining language model. The model architecture of the BERT model is composed of a multi-layer bidirectional Transformer encoder. As a neural network language model, this model can directly train a large number of untagged texts, which is applicable for text classification [24].

Two main phases of the BERT training process are pre-training and fine-tuning [25]. BERT is trained on a huge corpus of unlabeled text during pre-training by predicting masked words in a sentence and detecting if two sentences naturally follow one another. BERT learns contextualized representations of words that capture detailed semantic and syntactic information due to this unsupervised pre-training. Following pre-training, BERT is refined for a selection of downstream tasks. Using labelled data specific to the target task, BERT is merged with task-specific layers, such as additional neural network layers or a linear classifier, and then optimized. BERT can customize and specialize for the particular NLP work at hand due to the fine-tuning stage.

One of BERT's main advantage is the capability to properly capture contextual information, even for words with numerous meanings or words that rely on the context for disambiguation. BERT is extremely capable of handling numerous NLP tasks with remarkable performance because it can develop an in-depth comprehension of language structures and details by utilizing large-scale pre-training. For the text classification task such as this goal of the research, the BERT model enables further training of the fundamental model for a particular job with just one additional layer of neurons [26]. After more training using a specific methodology and BERT, obtaining models with the current top performance in text classification tasks is possible.

D. Performance Metrics

a) Accuracy

Performance metrics is used to determine how well the model did on the test data based on the training data it was trained on. To calculate the performance metrics that were used in this models, we need to know what a confusion matrix is. Confusion matrix is used to represent each outcome of the model [27], it has 4 components:

- True Positives (TP) expresses the quantity of samples that were accurately classified as "positive."
- False Positives (FP) expresses the quantity of samples that were incorrectly classified as "positive."
- True Negatives (TN) expresses the quantity of samples that were accurately classified as "negative."
- False Negatives (FN) expresses the quantity of samples that were incorrectly classified as "negative."

b) F1 Score

F1 score serves as a performance metric that functions as a harmonic mean between accuracy and recall [27]. It can be considered the optimal choice among the three methods, as increased precision comes at the expense of recall, and vice versa. Therefore, maximizing the F1 Score implies the simultaneous maximization of accuracy and recall scores. Mathematically, it is defined as follows:

$$F1\,Score = \frac{2}{\frac{1}{precision} + \frac{1}{recall}} \tag{2}$$

or

$$F1\,Score = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \tag{3}$$

IV. Result & Discussion

Several experiments were done on the dataset to determine the rating from the review. After the pre-processing and text vectorization steps, the SVM, ANN, and BERT models were implemented. Both the SVM and ANN models used TF-IDF for text vectorization while the BERT model used sentenceBERT.

The accuracy of the BERT model using 19 epochs is shown in Fig. 2.

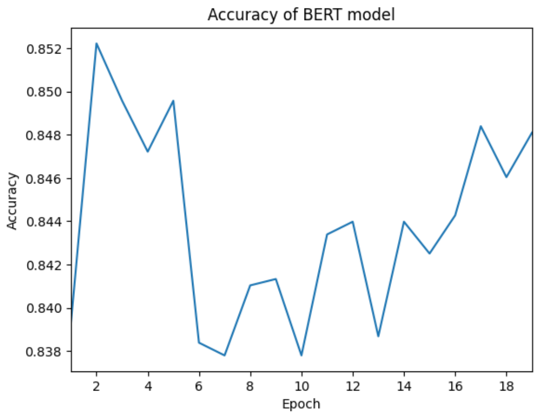

Fig. 2. Accuracy of BERT Model

The F1 score of the BERT model using 19 epochs is shown in the Fig. 3.

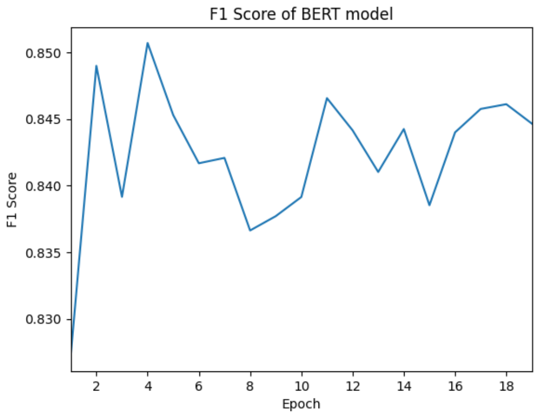

Fig. 3. F1 Score of BERT Model

From both Fig. 2 and Fig. 3, it is shown that the BERT model's performance varies between training iterations because the accuracy and the F1 score value fluctuated. The highest accuracy reached 0.852, while the highest F1 score reached 0.851. To improve the performance, the parameter must be adjusted to obtain different accuracy and F1 score results.

For the SVM and ANN models, the sentiment of the review was used rather than the rating as the labels. A review with a rating under 3-star was regarded as a review with negative sentiment. For a review with a rating over 3-star, it was considered a review with positive sentiment.

The accuracy and F1 score from the SVM, ANN, and BERT models are shown in Table. I.

TABLE I. ACCURACY AND F1 SCORE COMPARISON OF MODELS

Model	**Performance Metrics**	
	Accuracy	**F1 Score**
TF-IDF - SVM	0.923	0.961
TF-IDF - ANN	0.933	0.963
sentenceBERT - BERT	0.844	0.842

From Table. I, it is shown that ANN model with TF-IDF as the text vectorization obtained both highest accuracy and F1 score, followed by SVM model with TF-IDF. Lastly, BERT model obtained lowest accuracy and F1 score compared to the other two models. This could happen considering the amount of data in used dataset was relatively small.

V. CONCLUSION

This paper uses two well-known machine learning models and one deep learning model to classify whether the comment was categorized as positive or negative from analyzing the sentiment of the comment. The two machine learning models were SVM and ANN, and the deep learning model was Bidirectional Encoder Representations from Transformers (BERT). The models ANN and BERT were proposed in order to gain better performance compared to popular machine learning models such as SVM. Prior to training the models on the comments, the comments underwent several stages, including pre-processing and feature extraction conducted by the dataset's creator. Additionally, text vectorization was performed using TF-IDF and sentenceBERT. Subsequently, each of the three models underwent training based on the comments. The model with the highest F1 score was TF-IDF with ANN, achieving a score of 0.963, followed by TF-IDF with SVM with a score of 0.961, and finally the BERT model with a score of 0.842. Regarding the accuracy of these models, the highest one was TF-IDF with ANN with a score of 0.933, followed by TF-IDF with SVM with a score of 0.923, and finally the BERT model with a score of 0.844. Therefore, overall, TF-IDF text vectorization with ANN model demonstrated the best performance in terms of both F1 score and accuracy. This result demonstrates its usefulness for binary sentiment classification. The studies also had some limitations too such as limited model diversity, dataset dependency, and scaling issues with BERT. Future study should look at model ensembles, fine-tuning BERT, cross-domain assessments, and feature analysis to develop complete sentiment analysis guidelines that are applicable to a wide range of applications and domains.

REFERENCES

[1] Statista, "Global digital population as of January 2020," *2020*. https://www. statista.com/statistics/617136/digital-population-worldwide/ (accessed May 03, 2023).

[2] B. Haralayya, "CHANGE IN CONSUMER BUYING BEHAVIOR:INCLINATION TOWARD E-COMMERCE COMPANIES," *Article in International Journal of Early Childhood Special Education*, 2022, doi: 10.9756/INTJECSE/V14I5.742.

[3] H. Linh DANG, N. Van BAO, and Y. Cho, "Consumer Behavior towards E-Commerce in the Post-COVID-19 Pandemic: Implications for Relationship Marketing and Environment," *Asian Journal of Business Environment*, vol. 13, no. 1, pp. 9–19, 2023, doi: 10.13106/ajbe.2023.vol13.no1.9.

[4] S. Farag, T. Schwanen, M. Dijst, and J. Faber, "Shopping online and/or in-store? A structural equation model of the relationships between e-shopping and in-store shopping," *Transp Res Part A Policy Pract*, vol. 41, no. 2, pp. 125–141, 2007, doi: 10.1016/j.tra.2006.02.003.

[5] A. Nestler, N. Karessli, K. Hajjar, R. Weffer, and R. Shirvany, "SizeFlags: Reducing Size and Fit Related Returns in Fashion E-Commerce," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, Aug. 2021, pp. 3432–3440. doi: 10.1145/3447548.3467160.

[6] OOECD, "Understanding online consumer ratings and reviews," Paris, 2019. doi: https://doi.org/10.1787/eb018587.

[7] D. E. Cahyani and I. Patasik, "Performance comparison of tf-idf and word2vec models for emotion text classification," *Bulletin of*

Electrical Engineering and Informatics, vol. 10, no. 5, pp. 2780–2788, Oct. 2021, doi: 10.11591/eei.v10i5.3157.

[8] X. J. Lee, T. T. V. Yap, H. Ng, and V. T. Goh, "Comparison of Word Embeddings for Sentiment Classification with Preconceived Subjectivity," in *Proceedings of the International Conference on Computer, Information Technology and Intelligent Computing (CITIC 2022)*, Atlantis Press International BV, 2022, pp. 488–502. doi: 10.2991/978-94-6463-094-7_39.

[9] S. Chirgaiya, D. Sukheja, N. Shrivastava, and R. Rawat, "Analysis of sentiment based movie reviews using machine learning techniques," *Journal of Intelligent and Fuzzy Systems*, vol. 41, no. 5. IOS Press BV, pp. 5449–5456, 2021. doi: 10.3233/JIFS-189866.

[10] B. Noori, "Classification of Customer Reviews Using Machine Learning Algorithms," *Applied Artificial Intelligence*, vol. 35, no. 8, pp. 567–588, 2021, doi: 10.1080/08839514.2021.1922843.

[11] M. J. Hossain, D. Das Joy, S. Das, and R. Mustafa, "Sentiment Analysis on Reviews of E-commerce Sites Using Machine Learning Algorithms," in *2022 International Conference on Innovations in Science, Engineering and Technology, ICISET 2022*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 522–527. doi: 10.1109/ICISET54810.2022.9775846.

[12] L. H. Fung and S. L. M. Belaidan, "Sentiment Analysis in Online Products Reviews Using Machine Learning," *Webology*, vol. 18, no. Special Issue, pp. 914–928, 2021, doi: 10.14704/WEB/V18SI05/WEB18271.

[13] X. Lin, "Sentiment Analysis of E-commerce Customer Reviews Based on Natural Language Processing," in *ACM International Conference Proceeding Series*, Association for Computing Machinery, Apr. 2020, pp. 32–36. doi: 10.1145/3436286.3436293.

[14] H. Zhang, "Model Comparison in Sentiment Analysis: A Case Study of Amazon Product Reviews," 2022. doi: https://doi.org/10.54097/hset.v16i.2224.

[15] K. Jindal and R. Aron, "A Hybrid Machine Learning Approach for Sentiment Analysis of Beauty Products Reviews," *Journal of Information Systems and Telecommunication*, vol. 10, no. 37, pp. 1–10, Dec. 2022, doi: 10.52547/jist.15586.10.37.1.

[16] W. R. Bristi, Z. Zaman, and N. Sultana, "Predicting IMDb Rating of Movies by Machine Learning Techniques," in *2019 10th International Conference on Computing, Communication and Networking Technologies, ICCCNT 2019*, Institute of Electrical and Electronics Engineers Inc., Jul. 2019. doi: 10.1109/ICCCNT45670.2019.8944604.

[17] M. I. Hossain, M. Rahman, T. Ahmed, and A. Z. M. Touhidul Islam, "Forecast the Rating of Online Products from Customer Text Review based on Machine Learning Algorithms," in *2021 International Conference on Information and Communication Technology for Sustainable Development, ICICT4SD 2021 - Proceedings*, Institute of Electrical and Electronics Engineers Inc., Feb. 2021, pp. 6–10. doi: 10.1109/ICICT4SD50815.2021.9396822.

[18] N. Shrestha and F. Nasoz, "Deep Learning Sentiment Analysis of Amazon.Com Reviews and Ratings," *International Journal on Soft Computing, Artificial Intelligence and Applications*, vol. 8, no. 1, pp. 01–15, Feb. 2019, doi: 10.5121/ijscai.2019.8101.

[19] A. U. Rehman, A. K. Malik, B. Raza, and W. Ali, "A Hybrid CNN-LSTM Model for Improving Accuracy of Movie Reviews Sentiment Analysis," *Multimed Tools Appl*, vol. 78, no. 18, pp. 26597–26613, Sep. 2019, doi: 10.1007/s11042-019-07788-7.

[20] M. Li, L. Chen, J. Zhao, and Q. Li, "Sentiment analysis of Chinese stock reviews based on BERT model," *Applied Intelligence*, vol. 51, no. 7, pp. 5016–5024, Jul. 2021, doi: 10.1007/s10489-020-02101-8.

[21] K. Puh and M. Bagić Babac, "Predicting sentiment and rating of tourist reviews using machine learning," *Journal of Hospitality and Tourism Insights*, 2022, doi: 10.1108/JHTI-02-2022-0078.

[22] D. A. Pisner and D. M. Schnyer, "Support vector machine," in *Machine Learning: Methods and Applications to Brain Disorders*, Elsevier, 2019, pp. 101–121. doi: 10.1016/B978-0-12-815739-8.00006-7.

[23] S. N. Alsubari *et al.*, "Data analytics for the identification of fake reviews using supervised learning," *Computers, Materials and Continua*, vol. 70, no. 2, pp. 3189–3204, 2022, doi: 10.32604/cmc.2022.019625.

[24] Y. Wen, Y. Liang, and X. Zhu, "Sentiment analysis of hotel online reviews using the BERT model and ERNIE model—Data from China," *PLoS One*, vol. 18, no. 3 March, Mar. 2023, doi: 10.1371/journal.pone.0275382.

[25] J. Devlin, M.-W. Chang, K. Lee, and T. Kristina, "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)*, Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. doi: http://dx.doi.org/10.18653/v1/N19-1423.

[26] Koroteev MV, "BERT: A Review of Applications in Natural Language Processing and Understanding."

[27] C. Goutte and E. Gaussier, "A Probabilistic Interpretation of Precision, Recall and Extreme Classification View project Modeling dependencies in latent topic models View project A Probabilistic Interpretation of Precision, Recall and F-score, with Implication for Evaluation," in *Advances in Information Retrieval: 27th European Conference on IR Research, ECIR 2005*, Santiago de Compostela, Spain: Springer Berlin Heidelberg, 2005. doi: https://doi.org/10.1007/978-3-540-31865-1_25.