Original Article

AI-DRIVEN SENTIMENT ANALYSIS OF REVIEWS IN THE FOOD AND BEVERAGE INDUSTRY

Ahmad Faizal Bin Rahman

Institute of Sustainable Energy, National Energy University, Selangor, Malaysia

DOI: https://doi.org/10.5281/zenodo.16407981

ABSTRACT Nowadays, social media plays an important role in receiving all kinds of information, including customer reviews or feedback for products or services. Data generated from social media may give significant input to a company, including, customer satisfaction. This study is conducted to assist in selecting a suitable sentiment analysis model focusing on Malay language and social media data type in the Food and Beverages (F&B) industry. Data were retrieved from online review platforms, using Python as the web-scraping technique. A standard text-processing approach was adapted to clean the textual data for succeeding analysis. Eight types of the Transformer model, namely BERT, Tiny-BERT, ALBERT, Tiny-ALBERT, XLNet, ALXLNet, Fast former, and Tiny-Fast former that have been pre-trained in Malaya Documentation were used and the sentiment class is grouped into three, namely positive, neutral and negative. Based on standard classification performance metrics, XLNet outperforms other models with 75.96% accuracy and 78.91% AUC value. This shows that, although the Malaya Documentation claimed that the Fast former model has the highest accuracy for the general media social dataset. Ultimately, XLNet is presented as the suitable model for the F&B dataset.

KEYWORDS Sentiment analysis, machine learning approach, customer reviews, and social media analytics, XLNet, Food and Beverages (F&B) reviews.

I. INTRODUCTION third-party website, such as a forum or community. On use to enhancements in technology, social media data has websites, consumers may post reviews of items using influenced society, employment, and beliefs due to the numerical star ratings (often using a scale of 1 to 5) and their Using technology that involves sentiment analysis and predictive analytics allows businesses to anticipate client wants and change their plans dynamically [6]. Massive amounts of user-generated data collected daily [1]. For example, it is possible to gain significant business knowledge and expertise by analyzing customer input on a product through social media to identify problems and improve product quality. According to [2], customer feedback should be seen as a chance for businesses to learn something new about their products or services. Additionally, they define it as a statement about unmet expectations and a chance to please consumers by enhancing the product or service [2]. Moreover, a customer has the right to complain at any time, and taking the time to give a point of view on the product or service signifies the consumer's commitment to the company. Research shows that consumers often do product research online and compare competing products based on user comments [3]. They define online customer reviews as consensus product evaluations published on a business or a thoughts using text. User evaluations significantly impact consumers' purchase choices and the price they are willing to pay for a product or service [3]. In other words, shoppers are more likely to purchase from a website that includes other user evaluations compared to the ones that do not have reviews. If reviews are visible on social media, they are more likely to make people feel more confident, leading to a higher conversion rate [4]. Accordingly, companies must plan to analyze the vast volumes of data shared on social media to gain deeper insights into how customers perceive their competitors. At this time, the use of technology is compulsory to ensure the data is analyzed intelligently. Advanced analytics technologies, such as big data platforms and artificial intelligence algorithms, can rapidly handle and understand this data, delivering significant insights into industry trends and customer preferences. Furthermore, For example, one of the NLP approaches, sentiment analysis, can offer a systematic method for reading and analyzing client comments in a relatively short period. Sentiment analysis finds, extracts, and measures emotions from customer-written evaluations [7]. The firm may use sentiment analysis to determine the general perspective of the reviews, such as whether they are more favorable or negative, without having to read each review [8, 9]. Sentiment analysis employs two basic techniques: machine learning and lexicon-based approaches [10]. The former approach has garnered huge interest during the last decade, but the latter technique has just gotten some. The machine learning approach's effectiveness stems from its capacity to manage huge datasets and complicated patterns, which makes it perfect for sentiment analysis [11]. On the other hand, the lexical based method has grown in prominence due to its simplicity and ease of execution [12]. Machine learning approaches, such as deep learning and neural networks, demonstrate great potential in analyzing big datasets and generating accurate sentiment predictions [11]. However, the lexical-based method, which employs preset dictionaries of sentiment-laden terms, is more straightforward and simpler to deploy, although it may lack the nuance obtained by machine learning models [13]. Recent advances in blended approaches, which integrate both techniques, have led to enhanced sentiment analysis efficiency [14]. Even though the former technique has received attention for more than 12 years, the study of Malay words is still lacking, and only a few models based on machine learning approaches have been conducted for Malay words. Hence, this research studies the performance of the existing machine learning models that are suitable for Malay text, using the same dataset in the food and beverage industry.

II. MACHINE LEARNING MODELS

In this study, the performance of eight machine learning models that employ a deep learning approach, to be specific, the variations of the Transformer model, will be evaluated. The transformer was proposed by Vaswani et al. [15], and since then, other available models have been developed inspired by Transformer. To name a few, there are BERT, Tiny-BERT, ALBERT, Tiny-ALBERT, XLNet, ALXLNet, Fastformer and Tiny-Fastformer. Transformers and their variants have achieved great success in many fields [16].

A. BIDIRECTIONAL ENCODER REPRESENTATIONS FROM TRANSFORMERS, BERT

Bidirectional Encoder Representations from Transformers (BERT), is proposed by Devlin et al. [17]. The name BERT reflects its foundation on the Transformer model. BERT is engineered to pre-train profound bidirectional representations from unlabeled text by considering the left and right context across all layers simultaneously. Initially, the model is trained on various unlabeled data and pre-training tasks. Initial tweaking of the BERT model utilized pre-trained parameter values. The labeled data from succeeding functions is then utilized to fine-tune all parameters. Although they all began with the same pre-trained parameters, each subsequent assignment has its customized models. BERT is unique because it employs a single architecture for all tasks. It has demonstrated superior performance across various natural language processing applications, such as answering questions and language comprehension [5]. Furthermore, BERT's framework enables it to record context with greater accuracy than earlier models, greatly improving its understanding of ambiguous language [6].

B. TINY-BERT

Tiny-BERT is one of the enhanced algorithms generated from BERT. It is designed particularly for knowledge distillation (KD) of transformer-based models. An innovative technique that promises to speed up speculation, decrease model size, and maintain accuracy makes the KD method easier to use. The KD approach makes it simple to transfer the vast quantity of information contained in a huge 'teacher' like BERT to a smaller 'student' like Tiny-BERT. [20] Presented a new TinyBERT learning architecture with two phases based on this notion. Transformer distillation was used throughout the first pre-training and final task-specific learning stages.

This method assures that Tiny-BERT may supplement BERT with task-specific data and public-domain learning. Furthermore, Tiny-BERT has shown superior results in different natural language processing tasks while drastically lowering computing costs [21]. In addition, the model's efficiency makes it appropriate for deployment on resource constrained devices, increasing its use in real applications [22].

C. ALBERT AND TINY-ALBERT

Lan et al. developed another upgraded model that evolved from

BERT, known as A Lite BERT (ALBERT). ALBERT is improved in terms of variable count since it uses fewer parameters than classic BERT while maintaining performance. Lowering the amount of parameters helps to conserve memory and improve training efficiency. Tiny-ALBERT is generated from ALBERT in the same way as Tiny-BERT is derived, which is through the Transformer distillation process. This decrease is accomplished by parameter-sharing approaches and factorized integrating parameterization, which preserves model capability while consuming fewer resources [23]. Furthermore, research has shown that ALBERT performs similarly or even better on several NLP tasks than BERT, demonstrating its efficiency and efficacy [24, 25].

D. XLNET AND ALXLNET

XLNet is one of the improved Transformer types. XLNet, driven by the most recent advances in auto-regression language modeling, solves BERT's shortcomings. During the pretraining phase, XLNet combines the most favorable features of augmented reality and artificial intelligence. The concept stems from a generalization of the auto-regressive pre-training approach, which allows for learning bidirectional contexts by maximizing the predicted probability over all potential factorization permutation orders [26]. This method overcomes the limitations of masked language models by combining permutation-based language modeling, allowing the model to capture a more thorough grasp of context [27]. Furthermore, this method makes use of the Transformer-XL architecture to efficiently handle longer sequences and dependencies [28]. ALXLNet, like Tiny-BERT and Tiny-ALBERT, enhanced the XLNet through the Transformer distillation process.

FASTFORMER AND TINY-FASTFORMER

Original Article

Wu et al. suggested Fastformer, another upgraded Transformer model. Fastformer models' global contexts with an additive attention mechanism rather than pair-wise interactions between tokens. This method allows the model to efficiently aggregate information from all tokens, better capturing the broad significance of the text [30]. Furthermore, by reducing attention computation, Fastformer decreases computational complexity, making it more appropriate for extended text sequences [31]. Before changing each token's, representation based on its interaction with global context representations, the Fastformer benefited and performed better in lengthy text modeling [29]. Tiny-Fast former, like Tiny-BERT, Tiny-ALBERT, and ALXLNet, enhanced the Fastformer through the Transformer extraction process.

III. METHODOLOGY

A. DATA COLLECTION AND DATA PROCESSING

This section describes our approach to sentiment analysis for online reviews in Malaysia's Food and Beverage (F&B) business. To evaluate the performance of each model, we analyzed evaluations from several ice cream-specific F&B shops, such as Llaollao, Bubblebee, InsideScoop, Mokti, Gulacakery, and Appetit Village. Web scraping techniques were used to extract data from online review sites such as Facebook, Google Reviews, Instagram, Shoppe, TikTok, and YouTube. This thorough methodology guarantees that the dataset is broad and representative, allowing for a detailed examination of sentiment analysis methodologies across multiple platforms and brands. Web scraping is a technique for turning unstructured HTML data into spreadsheet format. Python is used in this study for online scraping because of its robust, effective, and customizable data extraction capabilities. BeautifulSoup, a Python module that interfaces with the Natural Language Toolkit (NLTK), parses HTML and XML files to get textual reviews, reviewer IDs, user locations, and ratings. This approach resulted in 2,338 review records from the selected channels and platforms. This method produces a thorough and organized dataset, allowing in-depth study and definitions. The textual data was then cleaned using a typical textprocessing method in preparation for further analysis. This included deleting special characters, extending contractions, changing letter cases, removing stop words, and conducting tokenization. Such preparation is critical since review data frequently contains different noises, such as emojis, punctuation, and incorrect capitalization or characters. By resolving these concerns, the cleaning procedure improves data quality, resulting in more accurate and dependable analytical outputs.

B. SENTIMENT ANALYSIS

After the textual material has been cleaned, sentiment analysis is performed. In this investigation, eight variants of the Transformer model are used, each pre-trained on the Malaya Dataset. The pre-training procedure included approximately 10,000 samples from the IIUM-Confession and Malaysian Parliament texts, including standard Malay, social media Malay, and Manglish, a Malay-English hybrid. The performance of these models is then measured against the same Food and Beverage (F&B) dataset. [32] Found comparable preprocessing procedures, which support the strategy utilized in this investigation. To assess the effectiveness of these models, we employ typical binary classification metrics as a guideline for multiclass classification. The measures used include precision, recall, F-measure, and accuracy. These metrics evaluate the binary classifier's four outcomes: true positive (TP), false positive (FP), true negative (TN), and false negative (FN). True positive (TP) refers to appropriately determined content as positive by the categorization technique. False negatives (FN) occur when a classification technique mistakenly forecasts good material as negative. True negative (TN) refers to content that has been accurately classified as negative, but false positive (FP) is when the approach predicts negative

Original Article

information as positive, as indicated in other recent publications [33]. Table 1 summarizes these four outcome parameters of the binary classifier.

Table 1. Outcome parameters from the binary classifier

		Predicted Class	
		Positive	Negative
Actual Class	Positive	TP	FN
	Negative	FP	TN

In this study, a multi-class classification technique is used to divide results into three categories: positive, neutral, and negative. The four outcome parameters (TP, TN, FP, and FN) are specified using their conventional definitions. Specifically, the TP ratio is calculated from cases in which the actual class matches the projected class. The TN value is calculated by summing values from all rows and columns excluding the respective class's row and column. The FP value is derived from the sum of values in the columns corresponding to other classes, excluding the TP value. The FN value is determined by summing values from rows other than the TP value. Tables 2, 3, and 4 illustrate the calculation of TP, TN, FP, and FN values for the positive, neutral, and negative cases, respectively. This methodology aligns with recent practices in multi-class sentiment analysis, as demonstrated by [34], who utilized similar preprocessing techniques to ensure comprehensive and accurate classification. After calculating the TP, TN, FP, and FN values, the next step is performance analysis using accuracy, recall, and Fmeasure measures. Equation (1) may be used to determine the accuracy metric, which is the fraction of precisely predicted contents compared to total anticipated contents. The precision metric is intended to assess the likelihood of proper sentiment categorization.

$$\text{Precision}=\frac{TP}{TP+FP} \quad (1)$$

The remembrance metric is the proportion of correct projected elements to total number of contents, which may be calculated using Equation (2). The retention metric is used to determine the accuracy of predicted material.

$$\text{Recall}=\frac{TP}{TP+FN} \quad (2)$$

Table 2. TP, TN, FP, and FN values for the positive actual content

Positive Case		Predicted Class		
		Positive	Neutral	Negative
Actual Class	Positive	TP	FN	FN
	Neutral	FP	TN	TN
	Negative	FP	TN	TN

Table 3. TP, TN, FP and FN values for the neutral actual content

Neutral Case	Predicted Class		

Original Article

		Positive	Neutral	Negative
Actual Class	Positive	TN	FP	TN
	Neutral	FN	TP	FN
	Negative	TN	FP	TN

Table 4. TP, TN, FP and FN values for the negative actual content

Negative Case		Predicted Class		
		Positive	Neutral	Negative
Actual Class	Positive	TN	TN	FP
	Neutral	TN	TN	FP
	Negative	FN	FN	TP

Performance is measured using the F-measure. The Fmeasure indicates the best balance of accuracy and recall. The average F-measure score assesses an individual's overall performance on a scale of 0.0 to 1.0, with 0.0 indicating the lowest possible performance and 1.0 reflecting the highest potential performance. F-measure is calculated based on Equation 3 [37]. To determine the accuracy of the output, an accuracy metric is used, and it can be calculated using Equation 4 [38]. The higher value of accuracy determines that the sentiment classification predicts the actual polarity.

$$F\text{-measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

Other than performance metrics, the Receiver Operator Characteristics curve is also used to quantify the accuracy of a model in making distinctions between classes. In this study the Area under the Curve (AUC) is used as a mathematical representation of ROC. The higher value of AUC means the classification model can dependably distinguish between true positive and false negative [39]. This is shown from the axes, where the x-axis represents the FP rate while y-axis represents TP rate. The rate of false positives is determined by the proportion of negative events that are incorrectly identified as positive. Thus, the higher value in the x-axis shows a greater number of false positives compared to true negatives. Oppositely, true positive rate is an outcome where the model correctly predicts the positive class, hence the higher value in y-axis shows a greater proportion of true positive as compared to a false negative. The respective model achieves optimal performance when the AUC value is 1. On the contrary, when the value of AUC is 0, this means that the model treats opposite values for each actual and predicted outcome, in a simpler explanation, the model accurately predicts the opposite output for each data.

IV. RESULT AND DISCUSSION

The same dataset is tested in eight types of deep learning models, namely BERT, Tiny-BERT, ALBERT, TinyALBERT, XLNet, ALXLNet, Fastformer, and Tiny- Fastformer. Table 5 shows the summary of the performance metric for all models. As shown in Table 5, three models give a competitive percentage of accuracy,

Original Article

that is, Tiny-BERT gave 75.74%, Tiny-ALBERT gave 74.15% and XLNet gave 75.96%. From these three models, XLNet gives the higher average percentage for all precision metrics, recall metrics and F-measure metrics, as well as accuracy, as shown in Table 6.

In terms of AUC value, Fig. 1 summarizes the AUC graph for all models. Based on Fig. 1, XLNet reported the highest score, that is 78.91%, while the other two competitive models, that is, Tiny-BERT gives 59.58% AUC value and TinyALBERT gives 64.31% AUC value. Two models give the value under 50%, that is, Fastformer, 44.73%, and TinyFastformer, 45.37%. Full AUC value for the rest of the models is shown in Table 7.

Table 5. Summary of performance metrics for all models

Model	Precision			Recall			F-measure			Accuracy
	Positive	Neutral	Negative	Positive	Neutral	Negative	Positive	Neutral	Negative	
BERT	72%	64%	51%	73%	76%	17%	73%	70%	26%	68.0%
Tiny-BERT	76%	75%	78%	86%	80%	16%	81%	77%	26%	75.74%
ALBERT	67%	82%	71%	95%	54%	13%	79%	65%	22%	70.71%
Tiny-ALBERT	73%	77%	59%	89%	71%	14%	80%	74%	23%	74.15%
XLNet	73%	87%	67%	95%	60%	43%	82%	71%	52%	75.96%
ALXLNet	73%	71%	69%	87%	71%	13%	79%	71%	21%	72.34%
Fastformer	65%	48%	8%	46%	69%	7%	54%	57%	7%	49.94%
Tiny-Fastformer	67%	47%	8%	51%	74%	2%	58%	57%	3%	53.76%

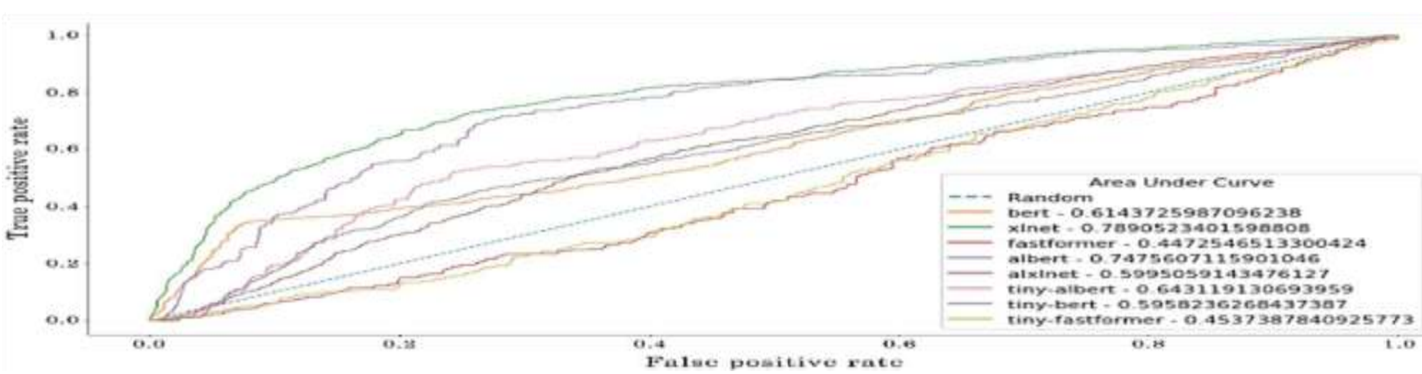

Figure 1. The area under the curve, AUC, graph for all models

Table 6. Summary of average performance metrics for top three models

Model	Precision	Recall	F-measure
Tiny-BERT	76%	61%	61%

Original Article

Tiny-ALBERT	70%	58%	59%
XLNet	76%	66%	68%

Table 7. The area under curve, AUC, value for all models

Model	AUC Value
BERT	61.44%
Tiny-BERT	59.58%
ALBERT	74.76%
Tiny-ALBERT	64.31%
XLNet	78.91%
ALXLNet	59.95%
Fastformer	44.73%
Tiny-Fastformer	45.37%

```
Model:   xlnet
True      1766
False      559
Name: xlnetout, dtype: int64
[[ 115   40   16]
 [  24  503   50]
 [ 130  299 1148]]
Accuracy: 0.7595698924731182

              precision    recall  f1-score   support

    negative       0.67      0.43      0.52       269
     neutral       0.87      0.60      0.71       842
    positive       0.73      0.95      0.82      1214

    accuracy                           0.76      2325
   macro avg       0.76      0.66      0.68      2325
weighted avg       0.77      0.76      0.75      2325
```

Figure 2. Report of performance metric for XLNet model

Based on the overall performance, XLNet outperforms other classifier models with 75.96% in accuracy. To be specific, the score of positive class for each precision, recall and F-measure were 73%, 95% and 82%, respectively. The score of neutral class for each precision, recall and F-measure were 87%, 60% and 71%, respectively, and the score of negative class for each precision, recall and F-measure were 67%, 43% and 52%, respectively. Details of these scores is shown in Fig. 2.

As achieved in performance metric, XLNet also reported the higher value for AUC that is 78.91%. As shown in Fig. 3, the AUC score for XLNet heading towards value of 1, in other words, towards the optimal value.

Original Article

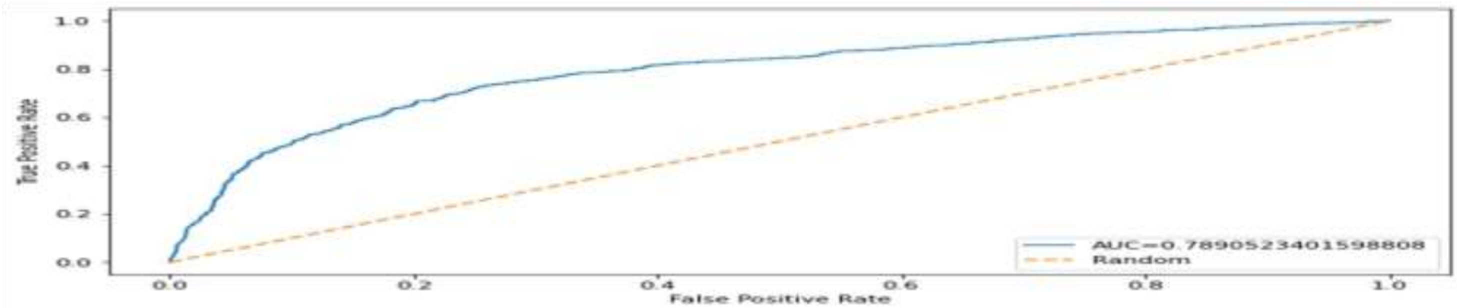

Figure 3. The area under the curve AUC, resulted from XLNet

VI. CONCLUSIONS

This study evaluates various deep learning models for sentiment analysis in the Food and Beverages (F&B) industry, with a focus on ice cream-related datasets. We assessed eight Transformer-based models: BERT, Tiny-BERT, ALBERT,

Tiny-ALBERT, XLNet, ALXLNet, Fastformer, and TinyFastformer. Our analysis shows that XLNet outperforms the other models, achieving an accuracy of 75.96% and an AUC of 78.91%. This indicates that XLNet excels in both precise classification and distinguishing between sentiment categories. XLNet's superior performance is attributed to its innovative use of autoregressive language modeling combined with autoencoding. This hybrid approach allows XLNet to effectively capture contextual information while addressing the limitations of previous models. In contrast to the other models, which either had lower accuracy or AUC scores, XLNet’s robustness in handling the complexities of sentiment analysis in social media reviews highlights its suitability for this task. Overall, XLNet demonstrates significant advantages in both classification accuracy and model reliability, making it the preferred choice for sentiment analysis in the F&B sector, especially for datasets related to the ice cream domain.

References

W. He, S. Zha, and L. Li, "Social media competitive analysis and text mining: A case study in the pizza industry," Int. J. Inf. Manag., vol. 33, no. 3, pp. 464-472, 2019.

X. Liu and R. A. Lopez, "Customer feedback analysis: Opportunities and challenges for modern businesses," J. Bus. Res., vol. 123, pp. 456-467, 2021.

A. Bauman, L. Tinguely, and C. Gros, "Product reviews and consumer behaviour," Electron. Commer. Res. Appl., vol. 23, pp. 56-67, 2017.

A. B. Rosario, F. Sotgiu, K. D. Valck, and T. H. A. Bijmolt, "The effect of electronic word of mouth on sales: A meta-analytic review of platform, product, and metric factors," J. Mark., vol. 84, no. 2, pp. 16-34, 2020.

D. Chaffey and F. Ellis-Chadwick, Digital Marketing: Strategy, Implementation and Practice, Pearson, 2022.

P. Mikalef, R. Van de Wetering, and J. Krogstie, "Building dynamic capabilities by leveraging big data analytics: The role of organizational inertia," Inf. Manag., vol. 57, no. 2, p. 103174, 2020.

K. Lipianina-Honcharenko, T. Lendiuk, A. Sachenko, O. Osolinskyi, D. Zahorodnia, and M. Komar, "An intelligent method for forming the advertising content of higher education institutions based on semantic analysis," in Int. Conf. Inf. Commun. Technol. Educ. Res. Ind. Appl., Cham: Springer International Publishing, 2021, pp. 169-182.

R. Gramyak, H. Lipyanina-Goncharenko, A. Sachenko, T. Lendyuk, and D. Zahorodnia, "Intelligent Method of a Competitive Product Choosing based on the Emotional Feedbacks Coloring," IntelITSIS, vol. 2853, pp. 246-257, 2021.

K. Lipianina-Honcharenko, C. Wolff, A. Sachenko, O. Desyatnyuk, S. Sachenko, and I. Kit, "Intelligent Information System for Product Promotion in Internet Market," Appl. Sci., vol. 13, no. 17, p. 9585, 2023.

S. Shamsudin, R. Hassan, and N. Zahari, "Sentiment analysis: Review of methods and applications," Int. J. Adv. Sci. Eng. Inf. Technol., vol. 5, no. 4, pp. 324-331, 2015.

B. Liu, M. Hu, and J. Cheng, "Opinion observer: Analyzing and comparing opinions on the web," in Proc. 14th Int. Conf. World Wide Web, 2012.