Journal of Theoretical and Applied Information Technology
31st August 2025. Vol.103. No.16

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

# FUZZYFAKEROBERTA: FAKE REVIEW IDENTIFICATION IN E-COMMERCE PLATFORM

**TEH NORANIS MOHD ARIS[1], KHAIRUNNISA ABDUL RAMAN[2], AZREE SHAHREL AHMAD NAZRI[3]**

[1, 2, 3] Department of Computer Science, Faculty of Computer Science and Information Technology,

Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia

E-mail: [1]nuranis@upm.edu.my (corresponding author), [2]khairunnisaraman@gmail.com, [3]azree@upm.edu.my

**ABSTRACT**

Fake reviews on e-commerce platforms pose a substantial risk to the integrity of online reviews and can significantly mislead consumers, leading to unfavorable buying choices. These fake reviews can distort consumer perceptions, potentially resulting in financial losses and decreased trust in online shopping. This study investigates the integration of fuzzy logic with the fine-tuned RoBERTa model, resulting in the "fuzzyfakeRoBERTa" model, designed to detect fake reviews on e-commerce platforms. The fuzzyfakeRoBERTa model enhances the accuracy and precision of fake review identification by effectively addressing the inherent imprecision and uncertainty often present in data. The research methodology included replicating the original fakeRoBERTa model, incorporating fuzzy logic to handle ambiguous data better, and thoroughly evaluating the model's performance using key metrics such as accuracy, precision, recall, and F1-score. The fuzzyfakeRoBERTa model achieved a notable accuracy rate of 97.59%, indicating a significant improvement over the original model's performance. This enhanced accuracy demonstrates the model's superior robustness and effectiveness in identifying fake reviews. The findings suggest that integrating fuzzy logic into deep learning models can substantially improve their performance in tasks that involve complex and nuanced data. This research enhances the reliability and credibility of e-commerce platforms by offering a more precise and effective tool for detecting fraudulent reviews, thereby helping maintain consumer trust and ensure fair competition in the digital marketplace.

**Keywords:** *Fake Reviews, E-Commerce, Fuzzy Logic, Fakeroberta, Machine Learning, Deep Learning, Text Classification*

## 1. INTRODUCTION

The proliferation of e-commerce platforms has significantly altered how consumers make purchasing decisions, with online reviews playing a crucial role in this process. However, the increasing presence of fake reviews, created either by human writers or by automated systems, threatens the credibility of these platforms. Fake reviews can mislead consumers, damage the reputation of businesses, and negatively impact overall consumer trust in the platform.

To address this growing concern, this study proposes a novel hybrid approach—fuzzyfakeRoBERTa—which integrates fuzzy logic with the transformer-based model fakeRoBERTa to enhance the detection of fake reviews. This model is designed to handle the uncertainty and imprecision often present in text-based data, which traditional machine learning models may struggle with.

COVID-19 affected consumer behaviors, especially internet buying. Online shopping has become more popular due to lockdowns and social isolation to prevent the virus's spread. A United Nations Conference on Trade and Development (UNCTAD) assessment found that the epidemic has transformed internet purchasing forever. Over half of survey respondents purchase online more often and use the Internet for news, health information, and entertainment [1]. The poll also found a 6–10% growth in internet purchasing of most product categories. After COVID-19 lockdowns, internet shopping dropped, although it remains intensive [2]. Online transaction numbers have not decreased despite a dip in retail transaction values. Since the epidemic, consumers have adapted to internet shopping's ease, affordability, and benefits.

Changes in customers' online purchasing habits highlight the necessity of e-commerce platforms in delivering secure and convenient shopping amid crises. E-commerce shoppers cannot personally check a product or service before buying, hence product reviews are essential. Reviews provide prospective buyers an idea of a product's quality, usefulness, and evaluation based on previous buyers' experiences. Businesses need real product reviews to develop consumer trust. Reviews provide openness so clients may choose based on others' experiences. Sales and reputation may also be affected by product reviews. Ahsan, A. found that user reviews may affect product sales and aid consumers make purchases [3].

The rise of fake reviews challenges this trust. Fake reviews might mislead people into buying or not buying things. Research by Cao, C. (2023) shown that fake reviews may change buyers' demand perception and increase buying desire[4]. Customers evaluate merchant reputation, and fake reviews strongly influence their purchase. Deceived customers dislike the seller and platform which then provide bad evaluations [5]. These are how fake reviews hurt sales [6,7].

Due to fraudsters' advanced methods and internet platforms' constant change, spotting fake reviews is difficult. Despite these obstacles, machine learning and deep learning can solve this issue. These methods train models on a dataset of 'real' and 'false' reviews. Learning patterns, the trained model can predict if a new review is authentic or false.

These are some of the research problems for this research,

a) As fraudsters use intricate strategies and online platforms evolve, the number of destructive false reviews harming shops and customers is rising annually. Current detection approaches generally fail, underscoring the need for more powerful machine learning and deep learning models. Joni Salminen et al. (2022) used transformer-based fakeRoBERTa to identify fake reviews [8].

b) Online platforms evolve, and consumers adjust their behaviors over time. Thus, patterns learnt by a model may not apply in the future. A fuzzy logic mathematical framework may describe and reason about data uncertainty and imprecision. Sharma, D. K., et al. (2023) utilized it to identify sarcasm. No applications for fuzzy logic and a transformer-based approach to identify fake reviews exist [9].

This research has the following objectives:

a) The goal is to duplicate the fakeRoBERTa model locally and enhance its accuracy in identifying both fake and real reviews, ensuring the model can correctly classify reviews from both categories.

b) Develop the improvement, potentially using Fuzzy Logic techniques, to identify these fake reviews.

This study introduces an enhanced model, fuzzyfakeRoBERTa, which integrates fuzzy logic with the fakeRoBERTa model to improve the detection of fake reviews. The investigation will use Salminen J. et al. (2022) dataset. The dataset comprises 20,000 fake reviews created by computer-generated and 20,000 actual product reviews (from Amazon reviews) [8]. Therefore, the scope of this research focuses on utilizing fuzzy logic techniques to identify fake reviews on e-commerce platforms, and utilizing Salminen J. et al. (2022) dataset. The generated reviews dataset consists of original reviews presumably human-created and authentic, which are computer-generated fake reviews. The study involves developing and testing the fuzzy logic model to identify fake reviews. Salminen J. et al. (2022) paper is used as benchmark work to be compared with the experiment findings. Accuracy, sensitivity, specificity, precision, and f-score will be used for assessment of performance, obtained from the confusion matrix.

This research aims to leverage the strengths of fuzzy logic in handling uncertainty and imprecision to enhance the accuracy and reliability of fake review identification. The novelty of this research work is the combination of fuzzy logic and transformer-based model [8].

This paper begins with a background study on the topic, reviews existing literature on fuzzy logic techniques, discusses standard methods. Then this paper will present the methodology for implementing the fuzzy logic technique with fakeRoBERTa [8]. Later part presents the results of implementing fuzzy logic with fakeRoBERTa for fake review identifications. The final part concludes with potential improvements for future work, identifies limitations, and provides recommendations for improvement.

## 2. LITERATURE REVIEW

This part provides an overview of e-commerce platforms and customer reviews, discussing current research on fake review detection methods, features, and datasets. It discusses techniques used in previous studies in machine learning, deep learning, and fuzzy logic, and analyzes related works in this field.

### 2.1 Online Customer Reviews

Before the Internet, people often relied on acquaintances for purchasing decisions [10, 11, 12]. However, customer reviews have gained popularity in recent years, with websites like Yelp and Facebook encouraging users to post feedback on products and services [13, 14]. These platforms encourage customers to write reviews, potentially attracting new customers. Online customer reviews, which provide first-hand knowledge, are becoming increasingly popular as people share their ideas online, allowing them to base their decisions on others' opinions.

Real reviews are honest, truthful, and sincere opinions from customers who have used a product or service. They provide valuable judgment and information, helping others make informed decisions [15,16]. In 2021, over 70% of online customers read reviews before purchasing, and authentic positive feedback can significantly boost conversion rates [17]. Identifying real reviews involves checking the review date, specific words, scene-setting, profile, spelling and grammar, being wary of black-and-white reasoning, and watching for customer jacking. Several journal references *discuss* the importance of real reviews and methods to differentiate between fake and real reviews [15,18,19,20].

Fake reviews are misleading assessments of products or services on e-commerce platforms, often created to increase the perceived value of a product or tarnish competitors' prestige [21]. They can be created by humans, corporations, or artificial intelligence and can significantly influence consumer behavior and e-commerce revenue. Fake reviews can be categorized into three types [22,23]: misleading reviews intentionally created to deceive readers, reviews focusing solely on brands, and non-reviews that lack personal perspectives.

Identifying fake reviews, also known as spam review detection, is a challenge. Machine learning and deep learning are popular techniques for detecting fake reviews [16,18,24,25,26,8,27,28,29]. Supervised learning is a standard method for detecting fake reviews, which uses labeled data to differentiate fake reviews from real ones based on specified attributes. However, distinguishing fake reviews from real ones is challenging when reading numerous reviews. These techniques can identify fake reviews by identifying hidden text patterns that humans cannot detect [30]. Consumers must be aware of the dangers of fake reviews and employ measures to identify and evade them.

### 2.2 Current Research on Fake Review Detection

Jindal and Liu's 2007 study on detrimental reviews or opinions identified three types: groundless positive reviews, malignant negative reviews, and non-malicious fake reviews [21]. The first study categorizes reviews as either groundless positive reviews or malignant negative reviews, while the authors also note non-malicious fake reviews, such as brand-focused reviews and non-reviews, which refer to advertisements for other items or reviews without any viewpoint [22,23].

Identifying fake reviews has been the main area of interest in multiple research endeavors, utilizing a range of machine learning and deep learning methods [8,16,18,19,20,31, 32, 33, 34]. Conventional machine learning methods, such as Support Vector Machines (SVM), Logistic Regression (LR), and Random Forest (RF), have been employed to categorize reviews using features collected from the text. These techniques depend on datasets that have been labeled, with reviews categorized as either fake or genuine. This enables the models to identify patterns that is characteristic of fraudulent content and learn from them. Table 1 displays the previous technique employed for identifying fake reviews:

*Table 1: Summary of Previous Methods Used to Detect Fake Reviews*

| Ref. | Dataset | Method | Result | Comments |
|---|---|---|---|---|
| [8] | Amazon Review Data (2018) dataset | NBSVM -OpenAI - fakeRoBERTa | Accuracy - 95.6%. - 82.8%. -96.64% | -Used GPT-2 to build the fake review dataset. -The dataset created 20k fake reviews and 20k real product reviews. |
| [16] | Review from Naver Shopping | -SVC -LGBM -RF | Accuracy - 85.13% -83.88% -83.75% | -The reviews were in Korean language. -The dataset obtains probably already classified. |
| [24] | Reviews of 20 Hotels in Chicago hotel dataset | -NB -SVM | Able to classify the reviews to fake or real | - There was no accuracy result, as the research may still in progress |
| [26] | Amazon-China dataset | -Isolation forest -ARIMA -LOF -SVM | Accuracy - 0.83 -0.79 -0.80 -0.77 | - The dataset is in Chinese language |
| [27] | Amazon product reviews database | FRD-LSTM | Accuracy 97.21% | Leveraging DCWR for feature extraction and PCA for dimensionality reduction to improve the |

| | | | | |
|---|---|---|---|---|
| | | | | accuracy |
| [28] | online reviews sourced from popular e-commerce | Combining CNN and APSO | Achieving a high accuracy rate | Reduces the time required for model training and testing |
| [29] | OpenWeb Text, Stories, and others, totaling approximately 160GB of text data | Created several significant modifications to the original BERT model such as employs dynamic masking | Accuracy - 83.2% | Eliminating the NSP task, adopting dynamic masking, and leveraging larger datasets and batch sizes, RoBERTa achieves substantial performance gains over the original BERT model |
| [32] | Wikimedia and Yelp dataset | Deep learning paradigm known as DenyBERT built on TinyBERT architecture and Knowledge Distillation (KD) techniques | Accuracy – 96.12 | The application only focuses on English language product and reviews, and the use of standalone software may cause difficulties for users |
| [33] | The Arabic Fake Reviews Detection (AFRD) dataset, consisting across hotels, restaurants, and product domains | Deep learning (DL) and Multiscale Cascaded domain based (MCDB) approach | Accuracy – DL+MCDB, Hotel : 100% - DL+MCDB, Restaurant : 100% - DL+MCDB, Product : 87.23% | - Can only detect fake reviews in Arabic in selected domains only (Hotel, Restaurant and Product) |
| [34] | Amazon customer review dataset | combination of novel machine learning and DL approach that benefits from the semantic textual knowledge embeddings of BERT, reproduces the complexity using bi-LSTM architecture | Accuracy – 97% | - The limitations of this research is that it only focuses on the negative impact of fake reviews and Amazon review dataset. |

The machine learning methods in this previous research, enhance the ability to detect fake reviews by leveraging their unique strengths in understanding and analysing text data [16], [24], [26], [27], [28], [29]. However, the accuracy produced by these methods can still be improved.

Deep learning models have significantly progressed by utilizing extensive datasets and intricate designs to enhance detection accuracy. Convolutional Neural Networks, or CNNs, and RNNs, or recurrent neural networks, have been used to detect complex patterns in review text, whereas transformer-based models such as BERT (Bidirectional Encoder Representations from Transformers) and its variations have established new standards in natural language processing (NLP) tasks [8], [32], [33], [34].

The fakeRoBERTa model [8], developed by Salminen et al. (2022), is a fine-tuned version of RoBERTa specifically tailored for fake review detection. It utilizes a transformer-based architecture to process and analyze review text, achieving high accuracy in classifying reviews as fake or genuine.

Fake reviews are considered as complex issues in real life. The fake review content is full of uncertain and vague information. Fuzzy logic method is the answer to handle the complex, uncertainty and vagueness in fake reviews. Fuzzy logic has the capability to provide more nuanced, interpretable and accurate results compared to the mentioned methods in the literature review.

### 2.3 fakeRoBERTa [8]

Salminen J. et al. (2022) used an Amazon e-commerce dataset to develop a fine-tuned model, fakeRoBERTa, inspired by OpenAI's concept of fine-tuning the RoBERTa model for specific tasks. The study used two baseline models: the Naïve Bayes and Support Vector Machine (NBSVM) algorithm and the OpenAI fake detection model. The NBSVM was a hybrid of classic baseline algorithms used in NLP tasks, such as fake review detection. The OpenAI model was specifically developed for fake review detection, fine-tuning a RoBERTa model for the specific task. The results showed that fakeRoBERTa had the highest accuracy at 96.64%, precision at 97.35%, recall at 96.17%, and F1-score at 0.97. The NBSVM had an accuracy of 95.82%, precision at 97.53%, recall at 94.53%, and F1-score at 0.95. The OpenAI model had an accuracy of 83.00%, precision at 73.02%, recall at 92.41%, and F1-score at 0.82. The study concluded that fakeRoBERTa provides a more effective solution than both the NBSVM model and

the OpenAI fake detection model, making it a leading technique for addressing fake review detection in e-commerce datasets.

To summarize, the techniques discussed face problems as follows:

a) Due to fraudsters' complex tactics and online platforms' ever-changing nature, the number of cases of malicious fake reviews hurting retailers and consumers keeps increasing yearly. Current detection methods often fail to meet these challenges, highlighting the need for more advanced machine learning and deep learning models; Joni Salminen et al. (2022) have contributed to identifying fake reviews using fakeRoBERTa, a transformer-based model [8].

b) Online platforms are constantly evolving, with users adapting their behaviour over time. Hence, the patterns that a model learned at one point may not necessarily apply in the future. A fuzzy logic mathematical framework can be utilized to model and reason about uncertainty and imprecision in data. It has been used in various applications, including sarcasm detection by Sharma, D. K., et al. (2023) [9]. However, there have not existed any applications for fuzzy logic combined with a transformer-based model used to detect fake reviews.

Based on the mentioned problem statement, the objectives of the research are:

a) To replicate fakeRoBERTa model in a local machine and improve the accuracy obtained from fakeRoBERTa.

b) Develop the improvement, potentially using fuzzy logic technique, to identify these fake reviews.

This improvement validates the hypothesis that fuzzy logic can effectively complement deep learning techniques due to the capability to cater uncertainty, vagueness, and inaccurate information, leading to better performance in complex tasks, which will be explained in section 4.

## 3. METHODOLOGY

The study employs a methodical procedure, commencing with duplicating the fakeRoBERTa model, subsequently improving it through the application of fuzzy logic, and ultimately assessing the effectiveness of the combined model. The methodology is specifically crafted to guarantee a thorough assessment of the efficacy of the proposed model in identifying counterfeit reviews. Figure 1 displays the flowchart employed in the study conducted for this paper.

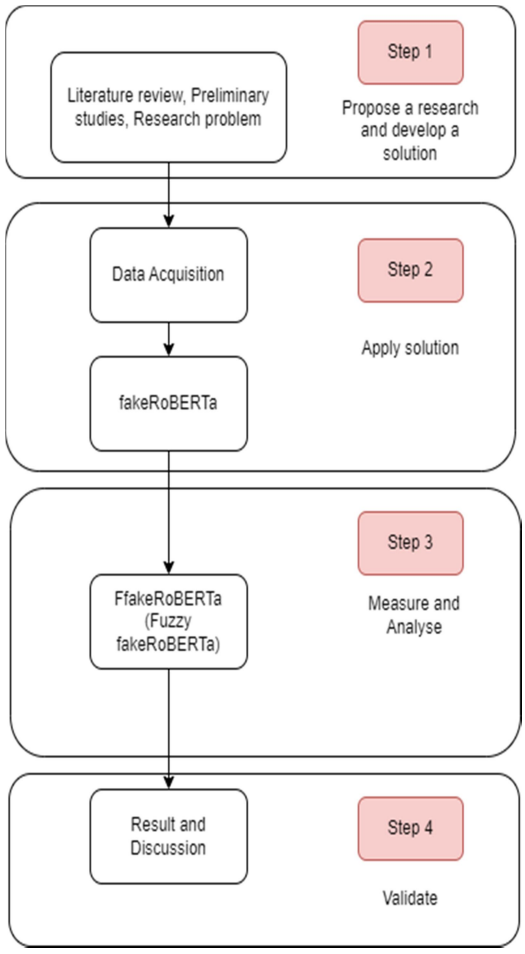


*Figure 1: Flow Chart for the research*

### 3.1 Dataset

The dataset used in this study was derived from the research by Salminen et al. (2022), which focused on creating and detecting fake reviews using machine learning models. The dataset consists of 40,000 product reviews, with an equal split between 20,000 genuine reviews and 20,000 fake reviews, ensuring a balanced dataset for the classification task. The genuine reviews were collected from the Amazon Review Dataset (2018), a widely used and reputable resource for e-commerce reviews.

#### 3.1.1 Dataset Generation

The fake reviews were generated using the GPT-2 language model, which was fine-tuned specifically for this task. GPT-2, a transformer-based model, was chosen for its superior text generation capabilities compared to other models

like ULMFiT. To generate fake reviews, the first five words of real reviews from the Amazon dataset were used as input, and GPT-2 completed the review, producing synthetic content that mimics the structure and style of human-written reviews. This method allowed the generation of a large volume of high-quality fake reviews, closely resembling real reviews in their linguistic features.

### 3.1.2 Dataset Composition

The dataset covers reviews from the top 10 most popular product categories on Amazon, including Beauty, Fashion, Automotive, Home and Kitchen, Electronics, and Sports. Each category contains a balanced number of fake and real reviews, which ensures that the model learns to classify reviews across various product types. Additionally, the reviews vary in length and sentiment, with ratings spanning from 1 to 5 stars, reflecting the diversity found in actual e-commerce review data.

### 3.1.3 Characteristics of Fake and Real Reviews

Fake reviews in the dataset tend to exhibit certain linguistic characteristics, such as more exaggerated sentiments, both positive and negative. In contrast, genuine reviews often present a more balanced tone and provide detailed product feedback. Fake reviews are generally shorter in length compared to genuine reviews, although longer reviews were also generated to simulate the range of review lengths found on e-commerce platforms.

This diverse and well-balanced dataset offers a robust foundation for training the fuzzyfakeRoBERTa model to detect fake reviews with high accuracy. The dataset used in this study includes 40,000 reviews from the Amazon dataset, equally split between fake and genuine reviews which obtain from [8]. This balanced dataset ensures a fair evaluation of the model's performance. The reviews cover various product categories, providing a diverse set of data for training and testing the model.

## 3.2 Implementatation of fakeRoBERTa


*Figure 2 : Flow Chart for fakeRoBERTa [6]*

The fakeRoBERTa [8] model, depicted in Figure 2, is fine-tuned on the review dataset to accurately categorize reviews as either fake or genuine, using the RoBERTa [29] model as a foundation. Text categorization and sentiment analysis are two of the numerous natural language processing applications where the transformer-based model RoBERTa has shown impressive results. The fakeRoBERTa model utilizes this architecture to efficiently process and analyze review text, accurately detecting patterns that suggest the presence of fake content.

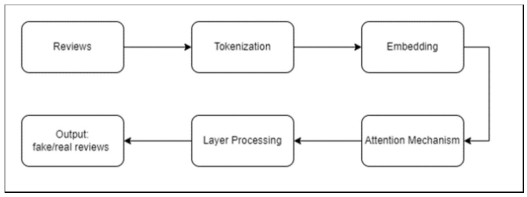


*Figure 3: The mechanism of fakeRoBERTa [8]*

fakeRoBERTa as shown in figure 3 is a deep learning model that uses a series of transformer layers to analyze reviews. It first breaks down the review into tokens, which are then converted into numerical representations. The model uses an attention mechanism to identify key phrases or patterns indicative of deceptive content, improving classification accuracy. The model then assigns different importance to each word based on the review's context. In the final layer, the model produces a probability score indicating the likelihood of the review being fake, with a score above a certain threshold indicating fake content.

## 3.3 Fuzzy logic implementation to fakeRoBERTa

Then Fuzzy logic is integrated with the fakeRoBERTa model to create the fuzzyfakeRoBERTa model. The fuzzy logic component processes the output probabilities of the fakeRoBERTa model, applying fuzzy rules to handle the uncertainty and imprecision in the data.

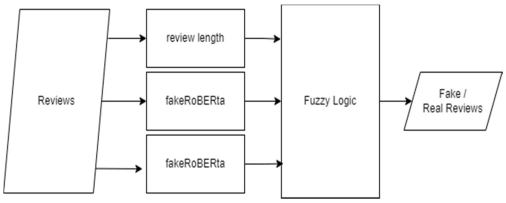


*Figure 4: Flow chart for Fuzzy Logic integration*

The flow chart that shows how fuzzy logic is incorporated into the fakeRoBERTa system is shown in Figure 4. The implementation of fuzzy logic for detecting fake reviews involves a few steps as follows:

i. identifying fuzzy variables such as sentiment, review length, model probability, and fake review.
ii. Fuzzification is the process of altering crisp input values into fuzzy values or sets, with triangular membership functions defined for each variable.
iii. The membership function is used for all variables, creating a triangle-shaped function where 0 is fully negative and 0.5 is not negative at all. If then fuzzy rules are defined, which combine the fuzzy input values to determine the fuzzy output values for 'fake_review'.
iv. Define the if…then fuzzy rules (inference). The inference engine applies fuzzy rules to the fuzzy inputs to derive fuzzy outputs. Below are the rules defined by using logical operations on the fuzzy variables:

```
rule1 = ctrl.Rule(sentiment['negative']
        & review_length['short'] &
        model_probability['high'],
        fake_review['likely'])
rule2 = ctrl.Rule(sentiment['positive']
        & review_length['long'] &
        model_probability['low'],
        fake_review['unlikely'])
rule3 = ctrl.Rule(sentiment['neutral'] &
        review_length['medium'] &
        model_probability['medium'],
        fake_review['possible'])
rule4 =
    ctrl.Rule(model_probability['high'],
    fake_review['likely'])
rule5 =
    ctrl.Rule(model_probability['low'],
    fake_review['unlikely'])
```

v. Defuzzification is the reverse process of fuzzification, transforming fuzzy values into crisp values. The 'compute' method handles defuzzification by processing fuzzy inputs through the control system and producing a crisp output. The 'compute' method applies the fuzzy rules and then de-fuzzifies the output to give a final probability of the review being fake.

In summary fuzzy logic implementation takes a review's text, its length, and the model's probability (fakeRoBERTa) to compute the likelihood of it being fake as illustrated in Figure 4. This approach combines the strengths of linguistic and probabilistic analysis to enhance detection accuracy. In summary, when analyzing a review, the sentiment score is computed based on its content, its length is measured, and the initial probability of being fake is obtained from the primary model (fakeRoBERTa). These inputs are then fuzzified, and the fuzzy rules are applied to determine the final probability of the review being fake.

Lastly, standard metrics including accuracy, precision, recall, and F1 score are used to evaluate the models. A confusion matrix is a popular and valuable technique in data science, machine learning, and deep learning for comparing and evaluating the performance of various models. It consists of four components which illustrated in Figure 5: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). The matrix represents the classification outcomes in terms of precision, recall, f-measure, and classification accuracy.

Accuracy is used to assess the overall performance of the model in classifying both fake and real reviews. Precision focuses on minimizing false positives, ensuring that the model correctly identifies fake reviews without misclassifying genuine reviews as fake. Recall emphasizes the model's ability to detect all fake reviews, minimizing false negatives. The F1-score, as a harmonic mean of precision and recall, is particularly helpful when balancing these two metrics in cases of skewed datasets, although the dataset in this study is balanced.

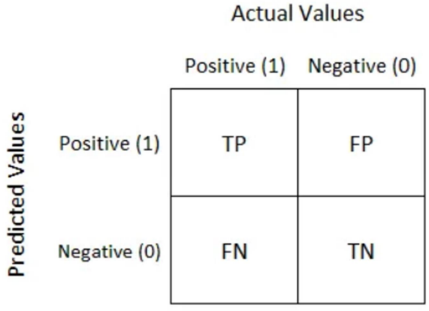


*Figure 5: Confusion Matrix*

Precision (1) is computed by taking the number of correctly identified positive outcomes and dividing

it by the number of positive results predicted by the model.

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

Recall (2) is the percentage of all positive results divided by the number of accurately detected positive discoveries.

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

The F1-score (3) finds the precision and recall harmonic mean, which helps to accurately capture the efficacy of prediction models.

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

Classification (4) accuracy is computed by taking the total number of samples and dividing it by the number of valid predictions.

$$Classification\ accuracy = \frac{(TP+TN)}{(TP+FP+TN+FN)} \quad (4)$$

This method is particularly helpful for evaluating how well the model performs when the dataset is skewed. The symbols TP, FP, TN, and FN represent true positive, false positive, true negative, and false negative values.

In the context of fake review detection, a "true positive" occurs when the review's projected value is "fake," and the original labeled review data is also "fake." The model by [8] will compare the results of the proposed implementation of fuzzy logic for fake review detection.

In this study, we employ common evaluation metrics for machine learning classification tasks, including accuracy, precision, recall, and F1-score. Each metric provides unique insights into the performance of the classification model and helps identify areas for improvement, particularly when dealing with imbalanced datasets, a common challenge in real-world scenarios.

It is important to clarify a potential contradiction regarding the use of the F1-score. The F1-score is often highlighted as a useful metric for cases where datasets are imbalanced because it provides a harmonic mean of precision and recall, ensuring that both false positives and false negatives are considered. Despite this, the dataset used in this study is balanced, containing an equal number of fake and real reviews (20,000 each). This might lead to the impression that the F1-score is not necessary in such a case. However, the F1-score remains highly relevant even in balanced datasets for several key reasons.

While accuracy measures the overall correctness of predictions (i.e., the proportion of total correct predictions out of all predictions), it does not offer insights into the balance between false positives and false negatives. In tasks such as fake review detection, the trade-off between false positives and false negatives is critical. For instance, misclassifying a real review as fake (false positive) can harm customer trust in the review system, while failing to detect a fake review (false negative) can lead to misleading product perceptions.

The F1-score is particularly useful in such cases because it balances the model's performance across precision (the ability to correctly identify fake reviews) and recall (the ability to detect all fake reviews). In cases where precision and recall are not equally optimized, focusing solely on accuracy can obscure underlying performance issues.

Although the dataset used in this study is balanced, it is critical to note that many real-world datasets are often skewed or imbalanced. For example, in a typical e-commerce environment, the proportion of fake reviews may be much smaller than the proportion of genuine reviews. In such cases, accuracy can be misleadingly high if the model predominantly predicts the majority class (real reviews). However, by optimizing for precision, recall, and F1-score, the model can be prepared to handle imbalanced future datasets where the F1-score will help ensure that performance is robust across both classes.

By employing the F1-score alongside accuracy, the model can optimize not just for balanced datasets but also make it robust enough to generalize well to imbalanced datasets where false positives and false negatives are likely to have different impacts. Therefore, the use of F1-score in this study is justified, as it not only ensures balanced performance in the current dataset but also prepares the model for future, potentially skewed datasets where the F1-score becomes even more critical.

In conclusion, a confusion matrix and its metrics provide a robust framework for evaluating, comparing, and improving fuzzy logic models for fake review identification. It is an essential tool in the data scientist's toolkit, especially in the era of digital commerce and online reviews.

## 4. 4. RESULTS AND DISCUSSIONS

The outcomes of applying the fuzzy logic technique to identify fake reviews on e-commerce sites. Primary goal for this research is to examine the performance of the fuzzyfakeRoBERTa model and compare it with the original fakeRoBERTa model. The results obtained from the original fakeRoBERTa model as reported [8]. Subsequently, provide the results from replicating the fakeRoBERTa model on a local machine to ensure reproducibility. Finally, the paper evaluates the fuzzyfakeRoBERTa model, highlighting its improvements in accuracy, precision, recall, and F1 score over the baseline.

### 4.1 fakeRoBERTa Model

Salminen et al.'s study uses a dataset of 40,000 reviews, including 20,000 real product reviews and an equal number of fake reviews, to identify fake reviews [8]. The study evaluates the performance of the fakeRoBERTa, NBSVM, and OpenAI models in distinguishing between real and fake reviews which is shown in Table 2. The fakeRoBERTa model outperforms all other models, with accuracy of 96.64%, precision of 97.35%, recall of 96.17%, and F1-score of 0.97. NBSVM has higher precision but poorer recall, resulting in a lower F1-score of 0.95. OpenAI's high recall but low precision and accuracy score of 73.02% and 83% respectively, reduces its usefulness. The study highlights the need for balanced machine learning models for fake review identification, optimizing precision and recall for high accuracy and resilient performance.

*Table 2: Result obtain from fakeRoBERTa, NBSVM and OpenAI [8]*

| Models | Accuracy (%) | Precision (%) | Recall (%) | F1-Score |
|---|---|---|---|---|
| fakeRoBERTa | 96.64 | 97.35 | 96.17 | 0.97 |
| NBSVM | 95.82 | 97.53 | 94.53 | 0.95 |
| OpenAI | 83.00 | 73.02 | 92.41 | 0.82 |

### 4.2 fakeRoBERTa implemented in local machine

The study replicated the method and dataset used [8] on a local machine to verify reproducibility and consistency. The system requirements and computational discrepancies were compared to ensure any differences in outcomes were due to methodological enhancements rather than computational discrepancies. The code used in the study was modified to run on the latest system requirements on the local machine. These adjustments included updating the code to be suitable for the CPU, ensuring no shuffle in the validation data, adding a step for saving the model after training, saving time and computational resources, and adding the function "valid" to evaluate the model's performance on unseen data. Additional steps were added in the local implementation for model testing and the PyTorch version was updated from 1.7.1 to 2.2.2 to match the latest compatibility with the rest of the software stack.

The confusion matrix obtained is as shown in Figure 6.

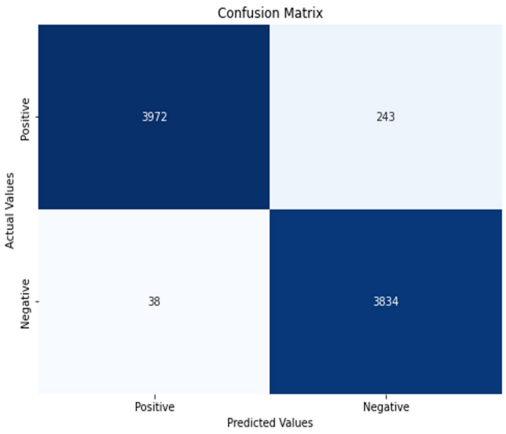


*Figure 6: Confusion Matrix for fakeRoBERTa Using Local Machine*

*Table 3: Performance Metrics for fakeRoBERTa Using Local Machine*

| Model name | Accuracy (%) | Precision (%) | Recall (%) | F1-Score |
|---|---|---|---|---|
| fakeRoBERTa | 96.53 | 99.05 | 94.23 | 0.9658 |

Table 3 shows the result obtain from fakeRoBERTa using local machine however the result obtained was differed from those reported by the original study [salminen]. The differences can be attributed to differences in system environments and software versions. The original study used a GPU for training and evaluation, while the replication used a CPU. GPUs' parallel processing can improve training efficiency, but the absence of shuffling in the validation set during replication

may have caused systematic biases in data batch processing. The local implementation also included additional steps for model validation, which may have introduced variability in the results. Consistent computing settings and hardware are crucial for machine learning research reproducibility. By ensuring consistent baseline results, the research can confidently evaluate the impact of incorporating fuzzy logic with the fakeRoberta model, ensuring observed improvements are due to enhancements rather than computational discrepancies.

### 4.3 fakeRoBERT Fuzzy Logic Implementation

The integration of fuzzy logic technique with fakeRoberta was applied to a dataset used in previous experiments. This hybrid approach aims to improve the accuracy and dependability of fake review detection by leveraging fuzzy logic's strengths in handling uncertainty and imprecision. The confusion matrix obtained is illustrated in Figure 7. The confusion matrix provides insight into the model's behavior in terms of false positives (FP) and false negatives (FN). Upon analysis, the model demonstrates a slight tendency toward false positives, where genuine reviews are misclassified as fake. This result is not uncommon in tasks involving nuanced, borderline cases, where legitimate reviews may contain characteristics (e.g., overly enthusiastic language or specific stylistic markers) that resemble those typically found in fake reviews.

Conversely, the model performs relatively well in minimizing false negatives, meaning it successfully identifies most fake reviews. However, the slight imbalance between FP and FN highlights the ongoing challenge of distinguishing between genuine reviews and sophisticated, fabricated reviews that closely mimic real content. The false positive pattern observed here suggests that further tuning of the model may be required to reduce the misclassification of real reviews as fake, without compromising its ability to detect fake reviews.

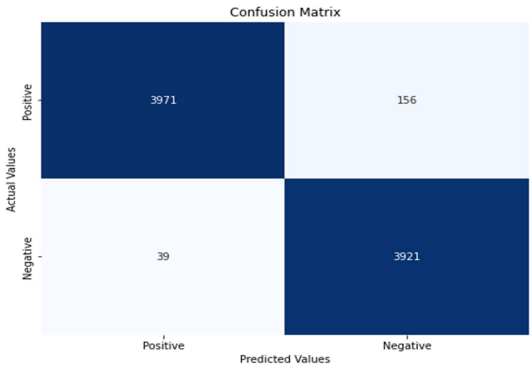


*Figure 7: Confusion Matrix for fuzzyfakeRoBERTa using Local Machine*

*Table 4: Performance Metrics Using Local Machine and its improvement.*

| Model name | fakeRoBERTa | fuzzyfakeRoBERTa | Improvement (%) |
|---|---|---|---|
| Accuracy | 96.64% | 97.59% | +0.95 |
| Precision | 97.35% | 99.03% | +1.68 |
| Recall | 96.17% | 96.22% | +0.05 |
| F1-Score | 0.97 | 0.9760 | +0.62 |

Table 4 shows the performance improvement of the fuzzyfakeRoBERTa model compared to previous models, fakeRoBERTa, is demonstrated. For clarity, the improvements in accuracy, precision, recall, and F1-score are now expressed in percentage terms. The fuzzyfakeRoBERTa model shows notable improvements across all metrics, particularly in terms of precision (+1.68%) and accuracy (+0.95%). These gains underscore the effectiveness of integrating fuzzy logic with the fakeRoBERTa architecture, leading to more accurate and reliable detection of fake reviews. The relatively smaller improvement in recall (+0.05%) suggests that the primary benefit of the new model lies in its ability to better distinguish genuine reviews from fake ones (i.e., higher precision), without significantly affecting the detection rate of fake reviews (i.e., recall).

Journal of Theoretical and Applied Information Technology
31st August 2025. Vol.103. No.16

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

### 4.4 Discussion

A clear comparison of the training dynamics of the fakeRoBERTa and fuzzyfakeRoBERTa models is shown in Figure 8, which also shows the training loss and accuracy of both models across 40 training steps.

The models are learning and improving their performance over time, as seen by the declining trend in training loss for both models in Figure 8(a). As training goes on, the fakeRoBERTa model's initial loss steadily drops from its higher starting point. Like fakeRoBERTa, the fuzzyfakeRoBERTa model likewise consistently reduces training loss, but at a somewhat steeper decrease. This quicker fall implies that the fuzzy logic integration improves the model's learning efficiency, enabling it to minimize the loss during training at a faster rate.

As training goes on, both models' training accuracy trends upward, indicating that they are becoming more accurate in classifying reviews. This is seen in Figure 8(b). The accuracy of the fakeRoBERTa model increases steadily, beginning at around 66.33% and reaching at 95.27% by the 40th step. By contrast, the fuzzyfakeRoBERTa model shows a little greater rate of progress, beginning at 64.23% and reaching about 95.29% accuracy by the 40th step. The incremental accuracy gains for fuzzyfakeRoBERTa that is shown at every iteration implies that fuzzy logic improves the model's performance, leading to more accurate classifications through the training process.

All things considered, these line graphs offer a thorough visual depiction of how both models advance over time, with fuzzyfakeRoBERTa outperforming fakeRoBERTa in terms of efficiency in minimising loss and attaining more accuracy. The advantages of adding fuzzy logic to the FakeRoBERTa model are demonstrated by this comparison, which leads to better training dynamics and overall performance.

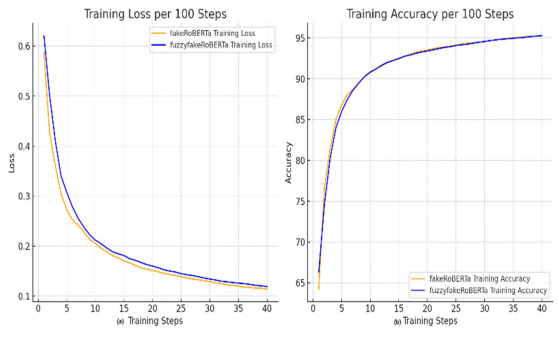


*Figure 8 The (a) training loss and (b) accuracy of the `fakeRoBERTa` and `fuzzyfakeRoBERTa`*

*Table 5: Summary of Performance Metrics for The Original Fakeroberta Model, The Replication On The Local Machine, And The Fuzzy Logic Combined With Fakeroberta:*

| Model name | Accuracy (%) | Precision (%) | Recall (%) | F1-Score |
|---|---|---|---|---|
| OpenAI [8] | 83.00 | 73.02 | 92.41 | 0.82 |
| NBSVM [8] | 95.82 | 97.53 | 94.53 | 0.95 |
| fakeRoBERTa [8] | 96.64 | 97.35 | 96.17 | 0.97 |
| fakeRoBERTa (Local Machine) | 96.53 | 99.05 | 94.23 | 0.9658 |
| fuzzyfakeRoBERTa (Local Machine) | 97.59 | 99.03 | 96.22 | 0.9760 |

From Table 5 the local replication used a CPU, which may have affected training speed and model convergence, resulting in different performance metrics.

These issues primarily arose due to differences in the computational configurations used during different stages of model evaluation. Specifically, the original training and evaluation of the model were performed on a GPU (Graphical Processing Unit), which is well-suited for handling the computational demands of deep learning models. However, during replication, the model was run on a CPU (Central Processing Unit), which led to minor performance discrepancies.

The discrepancies in performance observed between the original and replicated results can be attributed to the differences in hardware. GPUs are designed to handle parallel computations more efficiently, which speeds up the training process and improves the model's ability to generalize. In contrast, CPUs, though capable of performing the same tasks, often take longer and may result in slightly different outcomes, particularly in complex models like fuzzyfakeRoBERTa. Despite these discrepancies, the overall performance trends were consistent across both configurations.

These factors highlight the complexities in precisely replicating machine learning experiments and the necessity of consistent computational environments to ensure reproducibility in research.

From Table 5 shows that the integration of fuzzy logic with fakeRoBERTa resulted in a hybrid model that outperformed the original model and the

local replication in terms of accuracy and F1-score. The precision remained consistently high across all implementations, suggesting the robustness of the fakeRoBERTa model in identifying fake reviews. However, the recall metric showed variability, which was improved upon by the fuzzy logic integration. The hybrid approach of combining fuzzy logic with fakeRoBERTa enhances the model's performance in identifying fake reviews, validating the hypothesis that fuzzy logic can effectively complement deep learning techniques, leading to better performance in complex tasks.

In this study, the term "balance machine learning" does not refer to the balance of the dataset (i.e., the number of fake and real reviews), but instead to achieving robust performance across both classes. Specifically, balance here refers to the model's ability to accurately classify both fake and real reviews without favouring one class over the other. In machine learning, imbalanced performance occurs when a model achieves high accuracy for the majority class (real reviews) but performs poorly for the minority class (fake reviews).

The objective of this study is to create a model that can effectively detect fake reviews without disproportionately favouring real reviews. Achieving balanced performance means optimizing both precision and recall for each class. The model should be able to detect fake reviews while maintaining a low false-positive rate (incorrectly classifying real reviews as fake) and a low false-negative rate (failing to detect actual fake reviews). This balanced performance across both classes is a critical aspect of the fuzzyfakeRoBERTa model, ensuring that it is robust and reliable in practical applications.

## 5. CONCLUSIONS

This study successfully demonstrates that the integration of fuzzy logic with the fakeRoBERTa model improves the detection of fake reviews on e-commerce sites. The fuzzyfakeRoBERTa model outperforms the original model in all evaluated metrics, indicating its potential for more reliable fake review detection. The incorporation of fuzzy logic enhances the model's capacity to manage the inherent uncertainty and imprecision in review data, leading to higher accuracy, precision, recall, and F1 scores.

Although the fuzzyfakeRoBERTa model shows significant improvements over the original fakeRoBERTa model, several limitations remain. The integration of fuzzy logic increases the model's computational complexity, necessitating higher computational power. Additionally, the performance of the model relies on the quality and variety of the dataset, highlighting the need for more comprehensive and diverse data.

An innovative approach has been invented to enhance the detection of fake reviews on e-commerce platforms by integrating fuzzy logic with the fakeRoBERTa model. The primary aim was to leverage the strengths of fuzzy logic in handling uncertainty, vagueness and imprecision to enhance the accuracy and reliability of fake review identification. The findings from the experiments exhibited good results compared to the initial fakeRoBERTa model and its reproduction on a local device. The fuzzy logic-enhanced model achieved an accuracy of 97.59%, a precision of 99.03%, a recall of 96.22%, and an F1-score of 0.9760, highlighting the effectiveness of this hybrid approach.

Future work will focus on optimizing the model for real-time implementation, ensuring it can efficiently process and classify reviews on live e-commerce platforms. Enhancing the model's explainability is also crucial for broader acceptance among stakeholders, as it provides transparency into the decision-making process. Further research could explore the application of fuzzy logic in other areas of text classification and sentiment analysis, leveraging its strengths to address similar challenges in different domains.

## ACKNOWLEDGMENT

Thank you to the Faculty of Computer Science and Information Technology, Universiti Putra Malaysia for providing financial support under allocation vote 6236500.

## REFERENCES

[1] Suisse, N., & Secretariat, U. N. C. T. A. D. (2020). COVID-19 and e-commerce: findings from a survey of online consumers in 9 countries.
[2] Gupta, A. S., & Mukherjee, J. (2022). Long-term changes in consumers' shopping behavior post-pandemic: an exploratory study. International Journal of Retail & Distribution Management, 50(12), 1518-1534
[3] Ahsan, A. (2017). Consumer ratings-reviews and its impact on consumer purchasing behavior
[4] Cao, C. (2023). The Impact of Fake Reviews of Online Goods on Consumers BCP Business & Management, 39, 420–425. https://doi.org/10.54691/bcpbm.v39i.4208
[5] Chengzhang, J., Jinhong, Z., and Kang, D. (2014). Survey on fake review detection of e-commerce sites. J. Korea Inst. Inf. Commun. Eng. 1, 79–81.
[6] He, S., Hollenbeck, B., & Proserpio, D. (2022). The market for fake reviews. Marketing Science, 41(5), 896-921

[7] Song, Y., Wang, L., Zhang, Z., & Hikkerova, L. (2023). Do fake reviews promote consumers' purchase intention?. Journal of Business Research, 164, 113971.
[8] Salminen, J., Kandpal, C., Kamel, A. M., Jung, S., & Jansen, B. J. (2022). Creating and detecting fake reviews of online products. Journal of Retailing and Consumer Services, 64, 102771. https://doi.org/10.1016/j.jretconser.2021.102771
[9] Sharma, D. K., Singh, B., Agarwal, S., Pachauri, N., Alhussan, A. A., & Abdallah, H. A. (2023). Sarcasm Detection over Social Media Platforms Using Hybrid Ensemble Model with Fuzzy Logic. Electronics, 12(4), 937.
[10] Tian, Y., & Stewart, C. (2008). History of e-commerce. In Electronic commerce: concepts, methodologies, tools, and applications (pp. 1-8). IGI Global
[11] Zhu, F., & Zhang, X. M. (2010, March). Impact of Online Consumer Reviews on Sales: The Moderating Role of Product and Consumer Characteristics. Journal of Marketing, 74(2), 133–148. https://doi.org/10.1509/jmkg.74.2.133
[12] Patil, M. M., Nikumbh, S. N., & Parigond, A. P. (2021, March 30). Fake Product Monitoring and Removal for Genuine Product Feedback. Regular Issue, 7(1), 1–3. https://doi.org/10.35940/ijese.a2494.037121
[13] Choi, W., Nam, K., Park, M., Yang, S., Hwang, S., & Oh, H. (2023). Fake review identification and utility evaluation model using machine learning. Frontiers in artificial intelligence, 5, 1064371
[14] Gelder, K. (2023, July 11). Customer reviews: share of shoppers reading reviews 2021 | Statista. Statista. Retrieved January 25, 2024, from https://www.statista.com/statistics/1020836/share-of-shoppers-reading-reviews-before-purchase/
[15] Sumathi, V. P., Pudhiyavan, S. M., Saran, M., & Kumar, V. N. (2021, October). Fake Review Detection Of E-Commerce Electronic Products Using Machine Learning Techniques. In 2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA) (pp. 1-5). IEEE.
[16] Paul, H., & Nikolaev, A. (2021). Fake review detection on online E-commerce platforms: a systematic literature review. Data Mining and Knowledge Discovery, 35(5), 1830-1881
[17] Zhang, X., Guo, F., Chen, T., Pan, L., Beliakov, G., & Wu, J. (2023). A Brief Survey of Machine Learning and Deep Learning Techniques for E-Commerce Research. Journal of Theoretical and Applied Electronic Commerce Research, 18(4), 2188-2216.
[18] Jindal, N., & Liu, B. (2007, May). Review spam detection. In Proceedings of the 16th international conference on World Wide Web (pp. 1189-1190).
[19] Jindal, N., & Liu, B. (2008, February). Opinion spam and analysis. In Proceedings of the 2008 international conference on web search and data mining (pp. 219-230).
[20] Peng, Q. X., & Chen, J. W. (2013, July). Detecting Store Review Spammer via Review Relationship. Advanced Materials Research, 718–720, 2153–2158. https://doi.org/10.4028/www.scientific.net/amr.718-720.2153
[21] Thilagavathy, A., Therasa, P. R., Jasmine, J. J., Sneha, M., Lakshmi, R. S., & Yuvanthika, S. (2023, July). Fake Product Review Detection and Elimination using Opinion Mining. In 2023 World Conference on Communication & Computing (WCONF) (pp. 1-5). IEEE.
[22] Lin, T. Y., Chakraborty, B., & Peng, C. C. (2021, October). A study on identification of important features for efficient detection of fake reviews. In 2021 International Conference on Data Analytics for Business and Industry (ICDABI) (pp. 429-433). IEEE.
[23] Liu, W., He, J., Han, S., Cai, F., Yang, Z., & Zhu, N. (2019). A method for the detection of fake reviews based on temporal features of reviews and comments. IEEE Engineering Management Review, 47(4), 67-79.
[24] Qayyum, H., Ali, F., Nawaz, M., &; Nazir, T. (2023). FRD-LSTM: a novel technique for fake reviews detection using DCWR with the Bi-LSTM method. Multimedia Tools and Applications, 82(20), 31505-31519.
[25] Deshai, N., & Bhaskara Rao, B. (2023). Unmasking deception: a CNN and adaptive PSO approach to detecting fake online reviews. Soft Computing, 27(16), 11357-11378.
[26] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.
[27] Mohawesh, R., Xu, S., Tran, S. N., Ollington, R., Springer, M., Jararweh, Y., & Maqsood, S. (2021). Fake Reviews Detection: A Survey. IEEE Access, 9, 65771–65802. https://doi.org/10.1109/access.2021.3075573
[28] Jabeur, S. B., Ballouk, H., Arfi, W. B., & Sahut, J. M. (2023). Artificial intelligence applications in fake review detection: Bibliometric analysis and future avenues for research. Journal of Business Research, 158, 113631.
[29] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.
[30] Mohawesh, R., Xu, S., Tran, S. N., Ollington, R., Springer, M., Jararweh, Y., & Maqsood, S. (2021). Fake Reviews Detection: A Survey. IEEE Access, 9, 65771–65802. https://doi.org/10.1109/access.2021.3075573.
[31] Jabeur, S. B., Ballouk, H., Arfi, W. B., & Sahut, J. M. (2023). Artificial intelligence applications in fake review detection: Bibliometric analysis and future avenues for research. Journal of Business Research, 158, 113631.
[32] Le, Q. T., Do, T. K., Dang Thi, T. T., Luong Thi, M. T., Ngo, C. B., Tran, T. D., Hoang, G. B., Duong, D. C., D. C. (2024). Designing a Deep Learning-based Application for Detecting Fake Online Reviews. Engineering Applications of Artificial Intelligence, 134(108708), 1-14.
[33] Nour, Q., Ghadir, H., Maitha, A., Shatha, A., Waad, A., Arwa A. A. (2024). Multiscale Cascaded Domain-based Approach for Arabic Fake Reviews Detection in E-Commerce Platforms. Journal of King Saud University – Computer and Information Sciences, 36(101926), 1-15.
[34] Ronnie, D., Wasim, A., Kshitij, S., Mariann, H., Yogesh K. D., Ziqi, Z., Chrysostomos, A., Raffaele, F. (2024). Towards the Development of an Explainable E-commerce Fake Review Index: An Attribute Analytics Approach. European Journal of Operational Research, 317, 382-400.