Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.57, p. 519-528. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.3166

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

QUANTIFYING THE CREDIBILITY OF E-LEARNING SYSTEMS USING THE BERT MODEL

Piyush Singh[1], Rohan Pachisiya[2], Sagar Nehra[3] and Vibha Gaur*[4]

[1, 2, 3, 4]Department of Computer Science, Acharya Narendra Dev College, University of Delhi, New Delhi-110019, India.

[1]https://orcid.org/0009-0002-0769-8272, [2]https://orcid.org/0009-0000-6209-1866
[3]https://orcid.org/0009-0002-7907-6755, [4]https://orcid.org/0000-0001-6668-9339

Email: piyush.ae-1237@andc.du.ac.in, rohan.ae-1242@andc.du.ac.in, sagar.ae-1244@andc.du.ac.in, *vibhagaur@andc.du.ac.in

ARTICLE INFO

Article History
Received: December 18, 2025
Revised: December 20, 2025
Accepted: January 1, 2026
Published: January 31, 2026

Keywords:
User autonomy,
Dark pattern,
BERT classifier,
Educational websites,
Credibility index (CI).

ABSTRACT

Trust has become crucial in a digitized world for the sustainability of online platforms. As the web interfaces have been playing a vital role in day-to-day activities, ethical design has become essential to protect user autonomy and promote informed decision-making by them. Dark patterns are deceptive design strategies that can harm users' trust and are dangerous, especially in the e-learning environment. This paper presents a refined Bidirectional Encoder Representations of Transformers (BERT) classifier to automatically detect dark patterns on educational online systems. The framework starts with web scraping of the digital interface followed by organized preprocessing, such as content extraction, text cleansing, normalization, and tokenization. An algorithm to calculate a Credibility Index (CI) of e-learning systems is proposed based on the frequency and the severity of perceived dark patterns. The online systems are categorized into one of the three threat levels—*Safe*, *Moderate*, or C*ritical*, which gives clear indication to users regarding the trustworthiness of the site. Using the proposed framework, a customized educational website called SkillNest was developed to predict user trust. It was classified as *Moderate* due to its CI value of 0.64. This work may help developers in enhancing transparency and trust in educational technology by reducing the manipulative practices for developing e-learning systems.

I. INTRODUCTION

Dark patterns are untrustworthy and manipulative user interface (UI) design methods that are used by digital platforms to control user behavior in a way that can harm their choices, threaten trust, and break down faith [1-3]. The term dark pattern was given by Brignull [4] to describe ways of using mind tricks and psychological rules to make the platform earn the most money, usually at the cost of what's best for the users [2]. They include biases caused by social pressure, as well as the way choices are presented to push users toward performing things they wouldn't normally do [5], [6]. Also, such manipulative actions increase digital inequality, affecting digitally illiterate groups, and increasing the need for stricter rules on how platforms need to be designed [7]. Dark patterns restrict the users' rights to make a fair decision. Difficult procedures make it difficult for users to leave online services and thus limit their control over their personal data and online activities [6], [8].

These malpractices raise serious concerns about designing guidelines for developing online systems [2], [7]. Even the Department of Consumer Affairs has banned 13 different types of dark patterns. Since many people are not aware of these malpractices, there is a strong demand for authentic websites. Though various studies in literature report the types and effects of dark patterns [9], their automated identification in real time and measuring their impact has remained a challenge [3]. Manual detection of dark patterns requires a lot of time and effort and may be error-prone for complicated digital interfaces [10]. To address these issues, designers and customers need automated tools to detect and fix the causes of dark patterns [1]. Without these tools, people may find it challenging to follow the law and make informed decisions due to the increasing complexity of these dishonest strategies [11].

Finding the UI design malpractices on a huge scale is now feasible due to improvements in tools that understand natural language and learn from data, known as machine learning (ML). Large Language Models (LLMs), such as BERT [12], are capable of interpreting

the meaning and context of language [13]. This feature may be utilised in discovering elusive context-based dark patterns, which are otherwise difficult to catch by simple rule-based systems [14]. Also, these language models may offer deeper insights into these patterns of user behavior. These transformer models have also been utilised in other applications. For example, the BERT model has been used to extract and categorize large amounts of critical information from social media platforms [15]. Research by Kaur et al. [16] has showcased the application of the BERT model for quantifying the quality of online learning systems.

The proposed study presents an automated system to detect dark patterns and measure their impact to classify the educational platforms. It involves gathering of data, followed by preprocessing, including cleaning up the text and standardizing it using normal methods to make sure the data used is of high quality [17]. The refined BERT model calculates the chance that the user interface text is manipulative [18]. In addition, the proposed system also calculates the value of CI for digital platforms. The CI is a trust score that indicates the trustworthiness of a website by looking at the frequency and seriousness of the dark patterns. The digital platforms are assigned a threat level—*Safe*, *Moderate*, or *Critical*, based on the value of CI. It facilitates users' understanding of whether they can safely access the site or need to be careful or avoid the site altogether.

This work focuses on building credibility and trust by bringing it to the education world, where manipulative design practices can harm a student's confidence and engagement in e-learning platforms. Introducing the CI may help the students in using the trustworthy websites only. Though there are numerous ways of building trust. Sharma et al. [19] proposed that faith and trust in e-commerce can be developed by fixing problems like missing information and addressing bad reputations on business websites. Though this study has focused on educational websites, the framework is flexible and can be used for any digital platform. The main aim of the study is to develop an automatic system for predicting the credibility of educational websites. This research mainly addresses the problems caused by poor design on the educational platforms by focusing on the following objectives:

1. To build a system that automatically detects the dark patterns on learning websites using smart language models (like BERT).
2. To compute the CI of the online system based on the frequency and severity of the dark patterns identified.
3. To classify the online e-learning systems into three levels—*Safe*, *Moderate*, or *Critical* based on the value of CI.
4. To support fairness and transparency in digital educational environments.

The paper is divided into several parts: Section 2 provides a literature review of past research work on dark patterns and methods for finding them automatically. Section 3 presents the methodology for predicting the credibility of the educational online systems, while Section 4 discusses the test results. Finally, Section 5 concludes the paper with a summary of its main contributions.

II. LITERATURE REVIEW

Dark patterns are malign strategies that affect user trust. A lot of research is being done to identify and categorize the dark patterns. However, there are still gaps in terms of applying effective technology to identify these patterns in particular contexts. Luguri and Strahilevitz [2] were the first to recognize and categorize dishonest website layouts, such as the "sneak into basket," which produced the fundamental concepts for future studies. This classification was later expanded by Narayanan et al. [1] through interviews with experts, revealing 12 ethical issues, among which were “compulsory continuity” and “interface interference.” Mathur et al. [20] provided comprehensive evidence for this by analysing 11,000 shopping websites and arriving at 15 categories of dark patterns. Their results showed the effect of manipulative user interfaces on cognitive biases, thus emphasizing the immediate need for fast, scalable, and automated detection systems.

Gunawan et al. [21] presented that the dark patterns are significantly different between conventional sites and mobile applications, which means that lying tactics differ based on the platform. These studies provided valuable background knowledge towards the definition and description of dark patterns but failed to present an implementable, machine learning-powered, quantifiable assessment system in real time. It is in this regard that the necessity to have an automated system that constantly identifies and analyses dark patterns on digital platforms arises. Research in the field of behavioural economics has deeply searched for the psychological reasons of why dark patterns are used. A. Golandaz and U. Sharma [9] showed the way the dark patterns capitalize on cognitive biases such as loss aversion and default effects, and Greenberg et al. [8] have pointed to their usage in spatial interfaces in terms of close interactions. Mobile app analysis by Brignull [4] discovered that 95% of apps are designed in a manipulative fashion.

The work by Nouwens et al. [22] focused on the study of GDPR consent pop-ups using web scraping to reveal huge fraudulent cases. All these studies demonstrate how dark patterns restrict user autonomy, but studies continue to be limited by the differences in susceptibility by demographic, especially in educational contexts. Current methods of detection have numerous issues related to scalability and are not effective in terms of extensive tracking. Newer machine learning concepts have prospects and need to be explored fully as suggested by Mathur et al. [20], but have limitations due to the existing data sets. The proposed work bridges this gap by using the language model, which is sophisticated. Additional related literature in misinformation detection by Soni et al. [23] depicted that BERT based models and sentiment characterization can achieve good performance in identifying and interpreting manipulative information on social media using natural language processing (NLP) based approaches in detecting trust-violating patterns.

Ready-to-train transformer architectures, such as BERT [13] and RoBERTa presented by Liu et al. [14], have greatly contributed to the understanding of language formations, especially where manipulative design hints are hidden in minor textual details. These models generate deep contextual settings and support specific classification adaptation, as given by Sun et al. [24]. The introduction of the ELECTRA model by Clark et al. [25] enhances training effectiveness using substituted token detection. Though generally used for NLP, such models have not been well studied in identifying dark patterns. In this paper, these architectures are being used for autonomous identification of fraudulent user interface designs. Trust and credibility are important to the online platform, particularly in the face of fake features of designs, such as dark patterns, which possess more power to impact user behavior.

Little and Green [26] presented a framework for trust that is based on concepts like faith and leadership that doesn't depend on formal titles, reliability, and being seen as an expert. This framework shows that the developers have an opportunity to promote informed decision-making via reliable design. P. Paithane [27] stressed the importance of a structure-based framework that relies on both interaction

and behavioural data to measure faith in the online setting. Most critically, Corritore et al. [28] came up with a scale that can be used to measure customer trust on websites in accordance with the fact that a reliable tool is lacking in gauging online trust. Bedi and Gaur [18] extended the knowledge of trust among agent systems by including dimensions such as usability, performance, integrity, and security. It is on these grounds that the need to incorporate trust modelling into the process of identifying manipulative interfaces is amplified.

Lanier [29] suggested user-focused platforms that give importance to transparency and independence, while T. Nyström and A. Stibe [7] highlighted the wider risks of surveillance-based design with their ethical assessment. The Department of Consumer Affairs of the Government of India has also provided official guidelines in 2023 prohibiting certain dark patterns. As noted by S. Han and C. Anderson [30] regarding methodological biases in digital research, technical challenges with data collection persist. Scalable and integrated automated measurement tools and trust appraisal are still minimal, especially in education, where fairness is of great importance. Building upon these insights, this study examines an integrated approach, which is a combination of transformer-based classification and a CI intended to detect, measure, and report manipulative design practices, which is crucial for strong assessment of trust in e-spaces.

III. MATERIALS AND METHODS

Detection of dark patterns is crucial for the sustainability of the digital platforms. As the manual methods for identification of dark patterns are neither effective nor scalable. This research proposes an automated framework for dark pattern detection using the BERT model. The first step of the proposed work involves the extraction of the relevant e-content followed by preprocessing to clean and standardize the text. An optimally adjusted BERT-based model is utilised to identify instances of dark patterns with accuracy. To assess the general reliability of the online platforms, the Credibility Index algorithm is presented that considers the frequency and the severity of the discovered patterns. Figure 1 demonstrates the suggested framework to predict the credibility of online platforms. The framework comprises of the following steps: Data Extraction, Preprocessing of data, Pre-Trained BERT Model, Fine-Tuning of Model, Segregation of Dark Patterns, and Computation of CI, which are explained below.

III.1 DATA EXTRACTION

Data extraction involves downloading the raw HTML content of the URL on the online platform with the help of web scraping software like BeautifulSoup and Scrapy. This may consist of descriptive information, user reviews and policy documents. The raw HTML content is then changed to structured form.

III.2 PREPROCESSING OF DATA

The extracted data is processed using preprocessing operations to make it suitable for machine learning tasks. The raw content is first cleaned, i.e. HTML tags, special characters and other irrelevant content such as advertisements or navigation menus are removed. Linguistic noise is eliminated by normalizing the text, turning all letters to lowercase, and eliminating stop words. Lastly, the text is divided into a set of separate words or phrases called tokens, which can be interpreted by the language model.

III.3 PRE-TRAINED BERT MODEL

BERT is a language model that provides the semantic meaning of text by looking at the preceding and succeeding words simultaneously. It is a two-way text-processing model that processes the text in both directions at once and enables to see the whole picture of each token. It has been trained on large-scale corpora, such as BookCorpus and English Wikipedia, to predict and identify relationships between sentences. In this paper, the BERT-base-uncased model is used because of its good performance in natural language understa nding and high transfer learning, making it appropriate for detection of the dark patterns. The extracted data is initially processed with cleaning and tokenizing operations and then sorted into a prediction set for fine-tuning the model.

III.4 FINE-TUNING BERT MODEL

Even though the pre-trained BERT model provides an outstanding initial point for understanding language in general, it needs to be fine-tuned for detecting and classifying dark patterns. Fine-tuning involves adjusting the learning parameters of a model based on domain-specific knowledge so that it can learn to pick up the finer linguistic details and contextual differences that may signal deceptive interface designs.

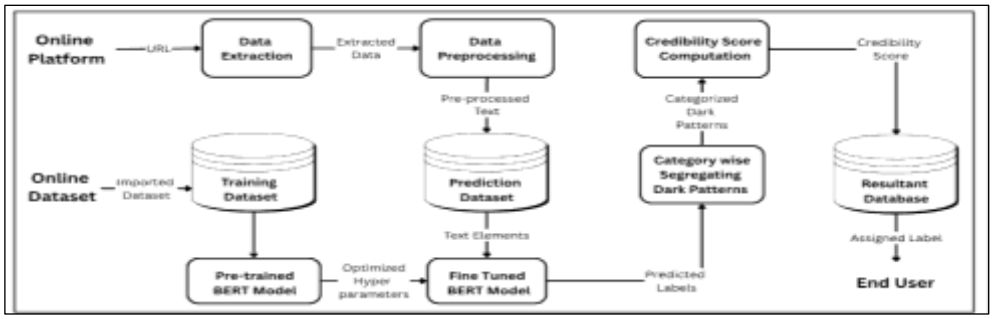


Figure 1: Proposed framework for determining the credibility of an online system.
Source: Authors, (2026).

At this step, the model does not change the transformer encoder architecture but instead changes the output layer by a task-specific classification head that is adapted to binary classification. The annotated dataset is trained with the optimal set of hyperparameters

so that the model can change the representations it learns based on the language and contextual clues that are commonly used in deceptive design.

III.5 SEGREGATION OF DARK PATTERNS

After identifying dark patterns with the help of the fine-tuned BERT model, these are classified according to the predefined taxonomies. This step is important in going beyond identification to find out the cause of the manipulative design. A labelled classification schema is used to map each instance to one of a set of predefined categories that enables the analysis without sacrificing the context of each manipulation strategy. This taxonomy helps to evaluate the behavioural capabilities of the interface of websites not only by their occurrence but specifically by the types of dark patterns used.

III.6 COMPUTATION OF CREDIBILITY INDEX (CI)

CI is a composite measure that shows the overall interface credibility of a platform based on the occurrence, frequency, and nature of dark patterns. It is calculated through the aggregation of weighted scores of identified dark patterns. Each dark pattern has been assigned a weight, which is based on empirical analysis and corresponds to the extent of loss the given dark pattern can bring. Weights may be procured from the domain experts. The proposed method uses the frequency (F) of each dark pattern and the weight of the dark pattern to derive a Severity Score. The values of the severity score and CI are computed using Eq. (1) and (2), respectively.

$$\text{Severity } Score = \left\{ \sum_{i=1}^{\lambda} (F_i \times W_i) \right\} \tag{1}$$

$$Credibility\ Index = 1 - Severity\ Score \tag{2}$$

Where:

- F_i – Frequency of the i^{th} dark pattern
- W_i – Weight of the i^{th} dark pattern
- λ – Type of unique dark patterns

Algorithm 1: Computation of credibility index.

Let Fi be the frequency of the i^{th} dark pattern, λ represent the maximum possible number of dark pattern types detectable by the system, and let W_i be the weight of the i^{th} dark pattern on the webpage.

for i = 1 to λ do

*SeverityScore ← SeverityScore + (Fi * W_i)*

end

CredibilityIndex ← 1 – SeverityScore

Return CredibilityIndex ∈ [0, 1]

Source: Authors, (2026).

Essentially, the Severity Score measures the disbelief on a digital platform in terms of the footprint of dark patterns, while CI is the extent to which the interface of a platform is trustworthy. The pseudocode of appraising the CI on the basis of the occurrence and severity of the dark patterns identified is given in Algorithm 1. The final value of CI lies between 0 and 1, which indicates the credibility of an online platform. More value of CI indicates a highly trustworthy site and vice-versa. It helps users to make wise choices and promotes the establishment of more equitable and transparent digital landscapes. A threat level is assigned to each website using CI to support the interpretation. CI is computed for each of the webpages of the site, and an average is taken to assign an overall threat level to the platform. Three levels of threat—*Safe*, *Moderate* and *Critical* are assigned, with *Critical* being the most manipulative design. The classification enables users to be aware of potential risks associated with the online systems.

IV. IMPLEMENTATION AND RESULTS

To see the application of the proposed framework, the bespoke website *SkillNest*, consisting of three webpages with a variety of dark patterns to allow more comprehensive and realistic evaluation, was created using Python at the backend for extracting and processing data and running algorithms. The user interface was designed using JavaScript, HTML, and Flask. And MySQL was used as the structured data storage, which guarantees the efficiency and scalability. The proposed framework aimed at categorizing seven classes of dark patterns, which are chosen due to their significance and frequency in digital user interfaces [20]. These are discussed in brief.

1. *Sneaking*: The deliberate concealment or omission of critical information to mislead users, such as hiding additional charges until the final checkout stage.
2. *Urgency*: It induces fake time constraints and forces them to take impulsive decisions.
3. *Misdirection*: Visual or textual indicators are provided to distract their attention away from relevant content, often leading them to take inadvertent actions.

4. *Social Proof*: It exploits users' psychological nature to follow others in unseen or unknown situations by showing fake or exaggerated data influencing the users' decisions.
5. *Scarcity*: It persuades the users by displaying limited stock available regardless of the actual availability.
6. *Obstruction*: It creates impediments to user workflows, thus intentionally obscuring crucial actions like subscription cancellations or privacy setting modifications.
7. *Forced Action*: It compels users to perform tasks such as mandatory account creation before accessing content or services. that are not directly relevant for accessing the primary intent,

These design patterns are often used to influence the user choices in a subtle or coercive manner. Each category of dark pattern targets a particular form of deception and could be beneficial for interface-based manipulation. Such a categorization schema is essential for model training, as well as credibility computation. The framework allows analysing manipulative design strategies in a granular and consistent manner by mapping identified patterns to particular behavioural strategies and improving the intelligibility and the responsibility of the detection system. The next step after the classification is the technical implementation of the framework with a strong system architecture. The proposed system used the following architecture:

1. *Content Extraction & Preprocessing*: Raw HTML content of the website SkillNest is scraped, parsed, and cleaned to retain relevant text and semantic elements.
2. *Model Integration*: A fine-tuned BERT model identifies and classifies textual elements indicative of dark patterns.
3. *Database Storage*: Model predictions, metadata, and user feedback are stored in a MySQL database.
4. *Interface Layer*: The front end is a Flask-based application that links the users to the backend services through the use of the HTTP paths so that users interact seamlessly with the real-time outcomes.

An uncased BERT model was fine-tuned utilising TensorFlow 2.0 plus the Transformers library in a Python environment and was used for the classification process. A snapshot of the implementation is presented in Figure 2.

```
train_InputExamples, validation_InputExamples = convert_data_to_examples(
    train, test, DATA_COLUMN, LABEL_COLUMN
)

train_data = convert_examples_to_tf_dataset(list(train_InputExamples), tokenizer, max_len=128)
train_data = train_data.shuffle(100).batch(16)

validation_data = convert_examples_to_tf_dataset(list(validation_InputExamples), tokenizer, max_len=128)
validation_data = validation_data.batch(16)

model.compile(optimizer=tf.keras.optimizers.experimental.AdamW(learning_rate=3e-5, weight_decay=0.01),
              loss=tf.keras.losses.SparseCategoricalCrossentropy(from_logits=True),
              metrics=['accuracy'])

model.fit(train_data, epochs=3, validation_data=validation_data)
```

Figure 2: The BERT classifier configuration parameters.
Source: Authors, (2026).

A labelled dataset [20] was utilised from the publicly available Yamana Lab repository [31] to fine-tune the BERT model. The resulting data set consisted of a combination of manipulative and non-manipulative elements of UI text, each of which was labelled as either dark (1) or not dark (0) by domain experts having knowledge of interface design and regulatory standards. Some of the examples of dark patterns found in the dataset were sneaking, urgency messages, social proof, forced action, and obstruction, and the non-dark patterns included transparent consent prompts and user-respecting design choices. To confirm the practical relevance of the trained model, an operating assessment pipeline was created in Python. This included the process of loading the fine-tuned classifier, preprocessing text found within HTML, and outputting predictions to find out the dark patterns. The 5-fold cross-validation was used to train the classifier so that it can generalize and minimize overfitting.

Every fold was trained on part of the data, and the remaining data was used for validation. This trial process was used to achieve the stable performance and emphasized the capabilities of the model to recognize obscure manipulative signals and context-sensitive wording that are employed in dark patterns. Table 1 provides the optimal hyper-parameter configuration that was used to fine-tune the BERT classifier, with the decisions being made based on performance and efficiency. The length of the maximum sequence was fixed to 128, which is the highest number of tokens that were handled each time a single input is given. *AdamW* optimizer is a weight decay adaptive optimization algorithm that was used in updating parameters. The classification objective function—*Cross-Entropy Loss* was used. The training was done in three epochs, and the architecture consisted of twelve transformer encoder layers. The number of samples processed by an update was set to 16.

Training was done with a fixed learning rate of 3×10^{-5}. An activation function—*GELU* was applied to allow nonlinear transformations. The model was evaluated using a five-fold stratified cross-validation method to maintain the balance in classes across folds. The regularization was carried out using a dropout rate of 0.1 in order to minimize overfitting. The system was combined into a lightweight Flask web server to facilitate real-time interaction. The requests were received by POST, processed by the backend (BERT predictions), and re-emitted back to the frontend to be displayed. This arrangement enabled the user to study webpages as they loaded with the feedback on deceptive patterns on the page. The core of the prediction process is the specialized BERT model, which outputs whether a phrase has a dark pattern or not. In case a pattern is identified, the model also categorizes it with the use of multi-label classification. This last step bridges the gap between detection and interpretability because users and researchers can interpret not only the existence of manipulation but also the nature of the applied deceptive strategy.

A CI was calculated on a one-on-one basis for each of the webpages of the *SkillNest* website. The index was computed using Algorithm 1, and values were stored in a MySQL database, which has scaled values across websites and user sessions and allows efficient access and retrieval of values that may be analysed further. To supplement the classification process, the framework was tested with the help of 5-fold cross-validation of the fine-tuned BERT classifier. The standard measures of classifier effectiveness—Accuracy, Recall, Precision, and F1-score were utilised. Accuracy is a measure of the correct predictions, and it gives a rough measure of the effectiveness of the model. Recall, or sensitivity, is the ability of the model to identify all true instances of dark patterns, which reduces false negatives. Precision is the frequency of the patterns that are actually represented by the model and minimize false alarms.

The F1-score represents the harmonic mean of the Precision and Recall in a single value, trading off one of the types of errors. Collectively, these measures provide a balanced evaluation of the capacity of the classifier in identifying manipulative patterns of the content presented in the websites. Table 2 summarises the results. To maximize the interpretability of the CI, the framework uses the notion of normalized Severity Scores, which measure the extent of damage each type of dark pattern causes. A survey was conducted with ten participants, comprising seven academicians with expertise in e-learning systems and three domain experts in digital ethics and online consumer protection, to derive the severity scores for the educational website *SkillNest* on a scale of 0—1. The participants evaluated significant dark patterns identified by Mathur et al. [20] for their applicability to educational platforms. These patterns were assigned weights based on cost, time, and effort required to evaluate the dark pattern, where cost refers to the potential financial loss imposed on users if a dark pattern successfully misleads them. Effort and time denote the additional effort and duration users spend to identify, understand, and evade the dark pattern. The final weight values for each dark pattern are presented in Table 3.

Table 1: Hyper-parameter settings of the BERT classifier.

Hyper-Parameters	Optimal Value
Max. Sequence Length	128
Optimizer	AdamW
Loss Function	Cross Entropy Loss
No. of Epochs	3
No. of Layers	12
Batch Size	16
Learning Rate	3e-5
Activation Function	GELU
Training Strategy	5-Fold Stratified Cross Validation
Dropout	0.1

Source: Authors, (2026).

Table 2: Model evaluation measures (5-fold cross-validation).

Metric	Score
Accuracy	0.97
Recall	0.97
Precision	0.96
F1-Score	0.96

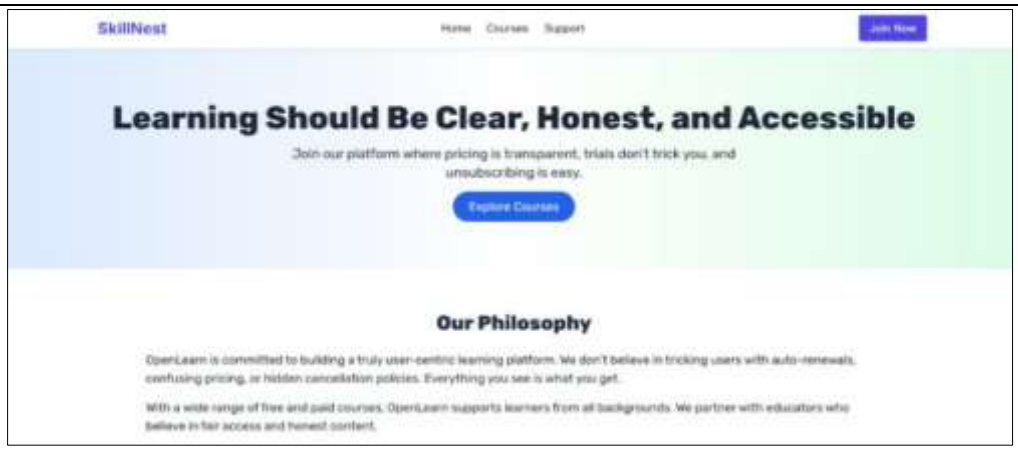
Source: Authors, (2026).

Table 3: Weights corresponding to each dark pattern type.

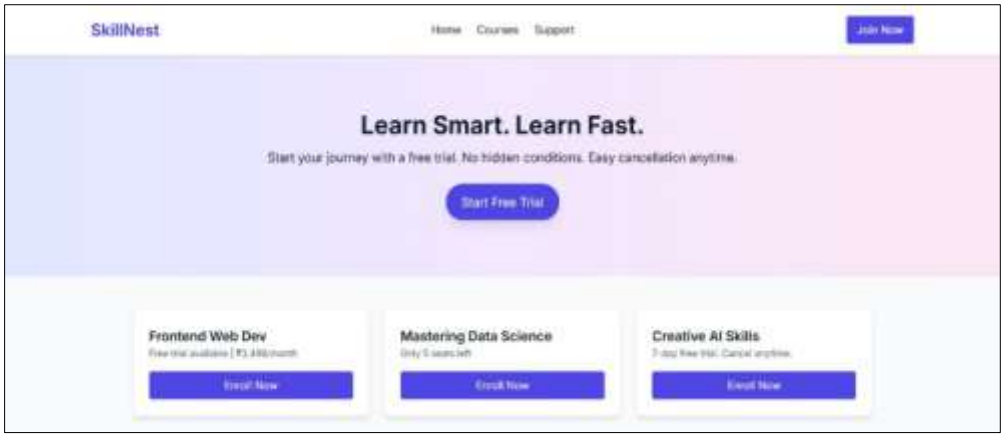
S.No.	Category of Dark Pattern	Weight of Dark Pattern
1	Sneaking	0.015
2	Urgency	0.267
3	Misdirection	0.155
4	Social Proof	0.166
5	Scarcity	0.374
6	Obstruction	0.019
7	Forced Action	0.004

Source: Authors (2026).

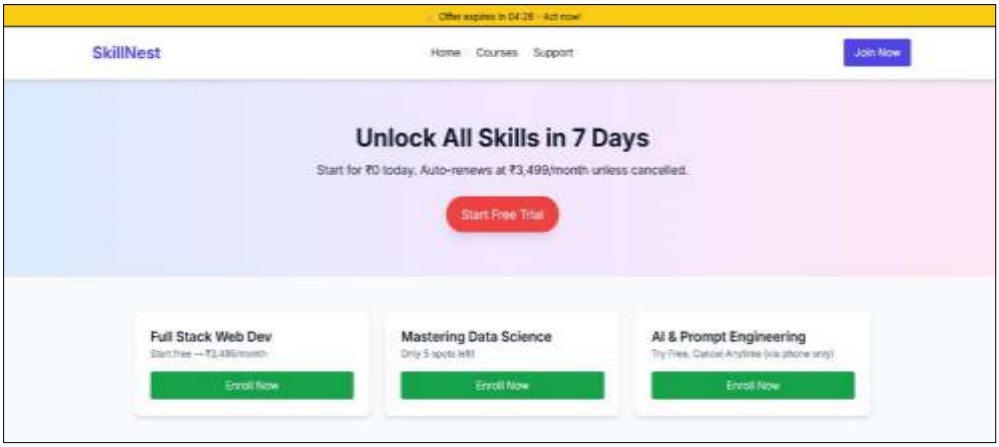
Further, the proposed framework was used for finding and analysing dark patterns in a realistic but controlled setting. The content of the websites was pre-processed in a methodical manner removing HTML tags, scripts, repeated lines, and other invalid bits in order to make clean and pertinent input in the analysis. Figure 3 shows the interface of the educational website that was developed and demonstrates the way the embedded dark patterns can be incorporated into the familiar layouts. The snapshot is a representation of design norms that can be observed in a very large variety of educational platforms. This real-world background enhances the validity of findings by showing how the manipulative design may work well in a real-world environment. The observed dark patterns were subsequently employed to calculate a CI, which indicates the fairness and reliability of the overall structure of the platform.



(a)



(b)



(c)

Figure 3: Identified segments of the dark pattern in three of the tested web pages of the custom website *SkillNest*.
(a) Webpage 1. (b) Webpage 2. (c) Webpage 3.
Source: Authors, (2026).

To give some context to the calculated credibility indices on the analysed cases, a system was prepared that categorized the websites into specific levels of threats in consultation with academicians and domain experts. This system relies on the average value of CI, which is the summation of the index values of all the webpages on the website. Based on this, every website can be rated as one of three levels— *Safe*, *Moderate*, and *Critical*, considering its general credibility and how deceptive features of design can affect the trust of the users and jeopardize online security. As shown in Table 4, a CI value between 0.86 and 1.00 is regarded as *Safe*, meaning that there is slight use of dark patterns of low severity and frequency. The interfaces in this range are usually transparent, just, and harmless for users to interact with. A CI value between 0.61 and 0.85 is classified as *Moderate* indicating the existence of moderately severe dark patterns or a comparatively concentrated mass of manipulation tactics that might quietly undermine user trust. Lastly, the score that falls below or at 0.60 is considered *Critical*, meaning that there is high usage of severe dark patterns with a high density. Potentially harmful possible interfaces in this range raise serious issues of ethical design and user control.

Table 4: Credibility Index Thresholds and Associated Threat Levels.

.No.	Mean Index Range	Threat Level	Implication
1	0.85 <	Safe	Minimal dark patterns, negligible threat.
2	0.60 < Credibility Index ≤ 0.85	Moderate	Be wary and take decisions carefully.
3	≤ 0.60	Critical	Highly manipulative and can't be trusted.

Source: Authors, (2026).

Webpage 1 shown in Figure 3(a) showed the least amount of deception, with 4 instances of the *Sneaking* dark pattern detected in 30 phrases of the content. This trend is not pronounced and is rare, but it gives a high CI value of 0.94. The content on the webpage seems open and trustworthy, which is not very threatening to users and has a high credibility with a minimum number of low-severity dark patterns. The webpage is designated as *Safe*. -Webpage 2 shown in Figure 3(b) contains a lot of deceit, with 3 instances of dark patterns *Obstruction* and 1 instance of *Social Proof*. The severity and intensity of these patterns were too high, resulting in a significant low value of CI as 0.777 and was classified as *Moderate*. The level of deception in webpage 3 shown in Figure 3(c) was observed as *Critical*, with 8 cases of dark patterns, including 5 instances of *Misdirection* and 3 instances of *Forced Action*. This is evident by the value of CI as 0.213, which shows a severely deceitful webpage, and should be avoided. The frequency and diversity of patterns are very misleading that decrease the levels of user trust. The evaluation of the website SkillNest is summarized in Table 5.

Table 5: Summary of evaluated webpages.

Webpage	**Dark Pattern Types Identified**	**Severity Score= (Frequency * Weight)**	**CI = 1-Severity Score**	**Interpretation**
Webpage 1	Sneaking-4	4*0.015=0.06	0.94	Less severe dark patterns; credibility is very high, and the web page seems to be trustworthy.
Webpage 2	Obstruction-3, Social Proof-1	3*0.019 + 1*0.166 = 0.223	0.777	Same number of patterns as in page1 but more severe; moderately deceitful.
Webpage 3	Misdirection-5, Forced Action-3	5*0.155 + 3*0.004 = 0.787	0.213	A high amount of deception and lack of credibility may be destructive.

Source: Authors, (2026).

The website SkillNest is considered to be a *Moderate* site, according to the average value of the CI as 0.643. It is prone to many threats and is less difficult to navigate without facing dark patterns. The proposed framework is able to identify and measure the impacts of dark patterns on educational sites. It offers easy-to-interpret credibility indexes by using a fine-tuned BERT model and severity-based score to enable users to make informed decisions.

V. CONCLUSIONS

This work introduced an automatic, online system for finding and analysing dark patterns that are present on online educational systems. The majority of research on manipulative interface design has focused in the fields of social media, and e-commerce, leaving their existence and impact in academic contexts largely unexplored. Since education is very dependent on trust and transparency, interpretation and preventing manipulation on educational sites is very essential. A novel framework is designed to identify dark patterns on educational systems, estimate their potential damage, and quantify trust in a definite and explicable way. The proposed system incorporated automated web scraping, high-level text processing, NLP, and a fine-tuned BERT model. This is done to identify manipulative designs that are included within education web interfaces. Besides identification of the dark pattern, the system divides these patterns into various categories to obtain a better view of the specific misleading patterns used. Additionally, a CI was introduced as a quantifiable measure that reflects the frequency and the severity of observed patterns on the websites. Weights were assigned to the dark patterns according to their potential to cause harm to user trust. Based on the value of CI, the educational systems were assigned the threat levels such as *Safe*, *Moderate* or *Critical*. These labels may help the users to make an informed decision about visiting or preventing the site.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Piyush Singh, Rohan, Sagar and Vibha Gaur.
Methodology: Piyush Singh and Rohan.
Investigation: Rohan and Sagar.
Discussion of results: Piyush Singh and Vibha Gaur.
Writing – Original Draft: Sagar.
Writing – Review and Editing: Piyush Singh and Rohan.
Resources: Sagar.
Supervision: Vibha Gaur.
Approval of the final text: Piyush Singh, Rohan, Sagar

VII. ACKNOWLEDGMENTS

The authors would like to thank the Department of Computer Science at Acharya Narendra Dev College, University of Delhi, for providing the facilities that made this research possible.

VIII. REFERENCES

[1] A. Narayanan, A. Mathur, M. Chetty, and M. Kshirsagar, "Dark patterns: Past, present, and future," ACM Queue, 2020, DOI:10.1145/3400899.3400901.

[2] J. Luguri and L. J. Strahilevitz, "Shining a light on dark patterns," Journal of Legal Analysis, vol. 13, pp. 43–109, 2021, DOI:10.2139/ssrn.3431205.

[3] C. M. Gray, Y. Kou, B. Battles, J. Hoggatt, and A. L. Toombs, "The dark (patterns) side of UX design," in Proc. 2018 CHI Conf. on Human Factors in Computing Systems, 2018, pp. 1–14, DOI:10.1145/3173574.3174108.

[4] F. Najar, "Deceptive by Design: Assessing the Impact of UX Dark Patterns on Engagement and Trust in Digital Products," 2025, DOI:10.13140/RG.2.2.25410.29120.

[5] A. Kitkowska, "The hows and whys of dark patterns: Categorizations and privacy," in Privacy and Identity Management: Fairness, Accountability, and Transparency in the Age of Big Data, 2023, pp. 173–198. DOI:10.1007/978-3-031-28643-8_9

[6] R. Selvi, S. Athinarayanan, V. Devi, M. Gobinath, M. R. Joel, and P. Shanthakumar, “Combining neural and semantic features in the analysis of being supportive in online feedback from customers,” Journal of Engineering and Technology for Industrial Applications (ITEGAM-JETIA), vol. 10, 2024, DOI:10.5935/jetia.v10i48.1002.

[7] T. Nyström and A. Stibe, “When persuasive technology gets dark?,” 2020, DOI:10.1007/978-3-030-63396-7_22.

[8] M. Belkacem, T. Boutefara, C. Abid, and M. Aberkane, “English machine learning techniques for identifying textual propaganda on social media,” ITEGAM – Journal of Engineering and Technology for Industrial Applications (ITEGAM-JETIA), vol. 11, 2025, DOI:10.5935/jetia.v11i53.1443.

[9] S. Greenberg, S. Boring, J. Vermeulen, and J. Dostal, "Dark patterns in proxemic interactions: A critical perspective," in Proc. 2014 Conf. on Designing Interactive Systems (DIS), 2014, pp. 523–532, DOI:10.1145/2598510.2598541.

[10] A. Golandaz and U. Sharma, “IoT under siege: The dark side of internet connected devices,” International Journal for Multidisciplinary Research, vol. 6, May–Jun. 2024, DOI:10.36948/ijfmr.2024.v06i03.22797.

[11] C. Nodder, Evil by Design: Interaction Design to Lead Us into Temptation. Wiley (Book), 2013.

[12] N. Adelakun and A. Baale, “Sentiment analysis of financial news using the BERT model,” Journal of Engineering and Technology for Industrial Applications (ITEGAM-JETIA), vol. 10, pp. 21–27, 2024, DOI: 10.5935/jetia.v10i48.1029.

[13] A. Rogers, O. Kovaleva, and A. Rumshisky, "A primer in BERTology: What we know about how BERT works," Transactions of the Association for Computational Linguistics, vol. 8, pp. 842–866, 2020, DOI: DOI:10.48550/arXiv.2002.12327.

[14] Y. Liu et al., "RoBERTa: A robustly optimized BERT pretraining approach," arXiv preprint, arXiv:1907.11692, 2019, DOI:10.48550/arXiv.1907.11692.

[15] A. Varshney, Y. Kapoor, V. Chawla, and V. Gaur, “A novel framework for assessing the criticality of retrieved information,” International Journal of Computing and Digital Systems, vol. 11, no. 1, pp. 1–11, 2022, DOI: 10.12785/ijcds/1101100.

[16] R. Kaur, R. Sharma, A. A. Jha, and V. Gaur, "A quantitative approach for appraising quality of online education," Journal of Engineering Education Transformations, vol. 38, no. 2, pp. 17–33, 2025, DOI:10.16920/jeet/2024/v38i2/24187.

[17] R. Mitchell, Web Scraping with Python: Collecting Data from the Modern Web. O’Reilly Media (Book), 2018.

[18] P. Bedi and V. Gaur, "Trust based prioritization of quality attributes," International Arab Journal of Information Technology, vol. 5, pp. 223–229, 2008, https://iajit.org/portal/PDF/vol.5,no.3/2-113.pdf.

[19] N. K. Sharma, V. Gaur, and P. Bedi, "Improving trustworthiness in e-market using attack resilient reputation modeling," International Journal of Intelligent Information Technologies, vol. 10, no. 3, pp. 57–82, 2014, DOI:10.4018/ijiit.2014070104.

[20] A. Mathur, G. Acar, M. Friedman, E. Lucherini, J. Mayer, M. Chetty, and A. Narayanan, "Dark patterns at scale: Findings from a crawl of 11k shopping websites," Proc. ACM on Human–Computer Interaction, vol. 3, no. CSCW, pp. 1–32, 2019, DOI:10.1145/3359183.

[21] J. Gunawan, A. Pradeep, D. Choffnes, W. Hartzog, and C. Wilson, "A comparative study of dark patterns across web and mobile modalities," Proc. ACM on Human–Computer Interaction, vol. 5, no. CSCW2, pp. 1–29, 2021, DOI:10.1145/3479521.

[22] M. Nouwens, I. Liccardi, M. Veale, D. Karger, and L. Kagal, "Dark patterns after the GDPR: Scraping consent pop-ups and demonstrating their influence," in Proc. 2020 CHI Conf. on Human Factors in Computing Systems, 2020, pp. 1–13, DOI:10.1145/3313831.3376321.

[23] A. Soni, S. Kaur, S. Jain, M. Karki, and V. Gaur, "Topic modelling, classification and characterization of critical information," International Journal of Computing and Digital Systems, vol. 14, 2023, DOI:10.12785/ijcds/140112.

[24] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to fine-tune BERT for text classification?," in Proc. 5th Workshop on Chinese Computational Linguistics and Natural Language Processing, 2019, pp. 194–206, DOI:10.48550/arXiv.1905.05583.

[25] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, "ELECTRA: Pre-training text encoders as discriminators rather than generators," in Proc. Int. Conf. on Learning Representations (ICLR), 2020, DOI:10.48550/arXiv.2003.10555.

[26] D. Little and D. A. Green, "Credibility in educational development: Trustworthiness, expertise, and identification," Higher Education Research & Development, vol. 41, no. 3, pp. 804–819, 2021, DOI:10.1080/07294360.2020.1871325.

[27] P. Paithane, “Trust aware recommendation using deep matrix factorization model”, Journal of Engineering and Technology for Industrial Applications, vol. 10, pp. 115–121, 2024, DOI: 10.5935/jetia.v10i48.941.

[28] C. L. Corritore, R. P. Marble, S. Wiedenbeck, B. Kracher, and A. Chandran, "Measuring online trust of websites: Credibility, perceived ease of use, and risk," in Proc. AMCIS 2005, 2005, p. 370.

[29] J. Lanier, Ten Arguments for Deleting Your Social Media Accounts Right Now. Henry Holt, 2018, DOI:10.1386/eme_00262_5.

[30] S. Han and C. Anderson, "Web scraping for hospitality research: Overview, opportunities, and implications," Cornell Hospitality Quarterly, vol. 62, 2020, Art. no. 193896552097358, DOI:10.1177/1938965520973587.

[31] Yamana Lab., "EC-DarkPattern: A dataset for dark pattern detection," 2023. Available: https://github.com/yamanalab/EC-darkpattern.