Received: 21 March 2025 | Revised: 15 October 2025 | Accepted: 3 December 2025 | Published online: 25 March 2026

RESEARCH ARTICLE

Journal of Data Science and Intelligent Systems
2026, Vol. 00(00) 1–10
DOI: 10.47852/bonviewJDSIS62025752

BON VIEW PUBLISHING

NLP Framework to Safeguard Youngsters Online Using Advanced Transformer-Based Models

Hisham AbouGrad[1,*], Sankar Santhosh[1] and Salem Alsaid[2]

[1]*Department of Computer Science & Digital Technologies, University of East London, UK*

[2]*Department of Computer Networks, Sebha University, Libya*

Abstract: Ensuring the safety and well-being of youngsters in online environments has become increasingly challenging, particularly with the rise of harmful and inappropriate content on social media platforms. This research study focuses on developing a natural language processing (NLP) framework designed to monitor online interactions and identify online harmful conversations. The framework integrates emotion detection with an advanced age verification model to analyze communication patterns and detect inappropriate behavior. When repeated sexually inappropriate behavior exceeds a predefined threshold, the system responds by blocking accounts and issuing notifications to guardians. The NLP framework was evaluated using two traditional machine learning algorithms alongside two advanced models, namely, BERT and RoBERTa, to assess their effectiveness in detecting harmful interactions. These models were trained and tested on datasets containing emotional patterns and real-world social media conversations. The proposed model's best accuracy result was 95.15%, which shows great promise in addressing inappropriate behavior during a conversation. However, several challenges were identified, including managing imbalanced data and the substantial computational resources required for model training. Despite these limitations, the framework can be used to enhance online safety for youngsters.

Keywords: sentiment analysis, emotion detection, NLP framework, BERT model, RoBERTa model, online safety, ethical AI

1. Introduction

The usage growth of online platforms has revealed more prospects, which have also brought new challenges that cannot be ignored. Indeed, digital platforms are great for supporting connections and creativity, but they have also turned into wide grounds for unsafe and inappropriate communications, especially in targeting youngsters [1]. This disturbing tendency has initiated serious concerns among many people, including parents and policymakers, which urges the need to have effective regulations and an examination framework. The truth is that improper online behaviors, such as pornography or sexually explicit interactions, can raise psychological risks to immature users and generate a tangled web of ethical and legal issues [2].

Recently, many advanced deep learning models, such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and large language models (LLMs), have achieved remarkable accuracy in text emotion detection. However, their high computational complexity and energy demands make them unsuitable for deployment on mobile devices and low-end machines. Along with that, there are limitations in a proper text sentiment model combined with strict age verification on social media conversations to recognize young users' inappropriate or harmful communications and interactions.

To tackle these issues, a natural language processing (NLP) framework has been proposed, which combines sentiment analysis and emotion detection toward a more advanced age verification model [3]. This NLP framework is designed to keep a watchful eye on online conversations and identify harmful behaviors and interactions in real time by harnessing the power of advanced machine learning models, including both traditional algorithms and cutting-edge transformer-based architectures, such as bidirectional encoder representations from transformers (BERT) and the robustly optimized BERT approach (RoBERTa), for the framework to take the necessary actions, such as blocking offending accounts or alerting guardians, to ensure a safer online environment [4]. The framework aims to enhance the accuracy of detecting harmful behaviors. Sentiment analysis is an NLP key component used to identify and interpret emotional cues in text [5]. The development of advanced models, such as BERT and RoBERTa, has significantly improved the ability to analyze user-generated content with higher efficiency and accuracy. Despite these advancements, challenges remain, including imbalanced datasets, limited interpretability, and difficulties in generalizing models across different languages.

Text emotion detection is a key approach and crucial aspect of NLP that codes, like zeroes, into emotional cues within text [6]. The emergence of more advanced and sophisticated models, such as BERT and RoBERTa, has revolutionized our ability to analyze user-generated content with remarkable efficiency and precision to recognize behaviors and identify emotions.

This research study utilized advanced models to recognize improper online content and addressed legal and ethical concerns by integrating age verification artificial intelligence (AI) models. By focusing on the safety of youngsters, the proposed framework aligns itself with all legal, privacy, and ethical standards [7]. The proposed model validation result achieved an exciting 95.1% accuracy rate using multiple datasets, which were used to prove the framework's capability.

In conclusion, this research study contributes to the growing fields of ethical AI, emotion detection, and online safety because it proposes an approach to protect youngsters in digital environments.

[*]**Corresponding author:** Hisham AbouGrad, Department of Computer Science & Digital Technologies, University of East London, UK. Email: h.abougrad@uel.ac.uk

Indeed, the proposed framework can be used to have security with more responsible online communication and behavior [8]. While this study focuses on developing a highly efficient model for English text, it is acknowledged that challenges, such as data imbalance and cross-lingual application, remain important areas for future work and research investigations.

2. Literature Review

Sentiment analysis has been a rapidly growing area of research in NLP, with numerous studies focusing on developing advanced models for sentiment classification. However, most existing studies have limitations in handling issues around inappropriate interactions, particularly involving youngsters, on online platforms [7]. Recent studies have employed various machine learning algorithms, including deep learning models, to improve accuracy. For instance, da Silva et al. [3] proposed a sentiment analysis system, with an accuracy of 87.20% with an SVM model on the best sampling with the Stanford dataset. Nandwani and Verma [9] utilized a large-scale dataset using a seven-layer deep learning CNN model with 87% accuracy.

Several studies have also explored the integration of emotion detection with sentiment analysis to improve the accuracy and performance of sentiment classification tasks. For instance, Ahmad et al. [10] achieved an accuracy of 86% on a motion-aware sentiment-aware model with the context of age verification. Several studies have employed machine learning algorithms to detect and prevent inappropriate interactions involving youngsters on online platforms. Angulu et al. [11] proposed an age verification system based on facial recognition using probabilistic neural networks (PNN), achieving an accuracy of 87% when trained with the HOIP face database. However, their study does not integrate emotion detection with age verification, which could enhance the accuracy of sentiment analysis. This limitation highlights the need for a more comprehensive approach that incorporates both age and emotional cues for improved performance.

2.1. Critical analysis of existing studies

Research gaps and challenges remain unaddressed, even though current research studies have shown significant improvements in sentiment analysis and NLP modeling (see Table 1). Traditional machine learning algorithms, such as SVM and deep learning-based CNN, have shown acceptable accuracy rates [12–14]. Conversely, these techniques are often unable to capture deeply related relationships in text, particularly in such sensitive conversations concerning young users [15]. In addition, while emotion-aware sentiment analysis and NLP models have enhanced classification accuracy, they are struggling with imbalanced qualitative datasets, which lead to biased predictions and decreased generalizability across many social media platforms. Further, current age verification models, such as facial recognition, are able to make another layer of security but lack integration with NLP-based sentiment analysis for comprehensive online safety solutions. The reliance on single-language datasets further limits the application of such NLP models in multilingual and different online environments. This research aims to bridge these gaps in multimodal sentiment analysis by combining age verification with sentiment analysis to improve online safety by utilizing advanced transformer-based architectures, such as the BERT model and RoBERTa model, for detecting harmful communication and interactions while addressing dataset imbalances, along with computational constraints.

2.2. Advancements in sentiment analysis

The latest advancements in sentiment analysis for social media platforms are exciting, especially with the new and enhanced NLP algorithms. Previously, sentiment analysis often relied on rule-based and simpler NLP models, which quietly come with limitations. Now, with the new era of deep learning architectures, such as BERT and RoBERTa, NLP models have become significantly accurate because they can understand the context and subtle meanings behind social media posts. This is crucial, especially when the informal and often chaotic nature of online communications and interactions is considered.

Further, the use of transformer-based models along with word embeddings has greatly boosted sentiment classification accuracy rates, despite dealing with complex and informal social media content. This supports scholars, researchers, and organizations to do real-time analyses of large volumes of social media data, which allows them to effectively be on top of new trends and public sentiments. Such advancements in sentiment analysis have initiated a world of many possibilities across many fields, from sales, marketing, and politics to client loyalty and feedback analysis, and therefore, this becomes a vital tool to recognize how people feel and what they think, which can support organizations to connect with their people in appropriately correct and more meaningful ways [16].

2.3. Research objectives

The study objectives are used to achieve the aims and answer the research questions, and these are the following: to conduct an inclusive

Table 1
Critical analysis: summary of the existing core research studies

Research study	Dataset	Algorithm	Accuracy	Limitations
Nandwani and Verma [9]	SemEval-2018	CNN	87%	The limited dataset size may lead to undersampling, which can negatively affect the model's performance.
Chatzakou et al. [12]	Twitter & Facebook	Lexicon-based approach	59.38%	Detection difficulties in a wider range of emotions from texts.
Martín-Valdivia et al. [14]	MuchoCine	SVM	86%	Struggles to detect a wider range of emotions from texts, beyond basic sentiments like positive/negative.
Mohammad and Turney [15]	Emolex	Crowdsourcing	66.67%	The use of Mechanical Turk and human annotation may not be scalable for creating large-scale emotion lexicons.
Proposed model	Online platforms	RoBERTa	95.1%	Data imbalance and cross-lingual applications are not covered.

sentiment analysis using social media text conversations to recognize young users' harmful or inappropriate communication and interactions, to examine current research studies on sentiment analysis methods and their effectiveness to detect young social media users' harmful or inappropriate communication and interactions, to evaluate the effectiveness of transformer-based models to recognize harmful and inappropriate content, to analyze the challenges in handling imbalanced datasets and computational resource constraints while enhancing model performance toward real-world applications, and to compare the performance of different NLP models for capturing sentiment nuances within the text.

Further, the following steps are used to conduct the study: data collection to retrieve three datasets containing text with different emotions; preprocessing will include text normalization, tokenization, and removal of noise from the dataset; sentiment analysis to apply RoBERTa, BERT, logistic regression (LR), and random forest (RF) for sentiment classification; temporal analysis will include the examination of sentiment trends over time; the event correlation phase will involve the identification of events influencing sentiment fluctuations; and the comparative analysis phase will involve the valuation of RoBERTa and BERT's performance in sentiment analysis.

3. Research Methodology

An agile-based approach is applied for integrating both traditional machine learning and deep learning algorithms to detect harmful and inappropriate online communication and interactions [13]. The methodology follows an iterative process with dataset collection, preprocessing, and NLP model training, evaluation, and deployment. In this research study, three datasets, GoEmotions, a custom lust dataset, and an emotion classification dataset, were used to train NLP models to recognize a wide range of emotions, including "lust." Data balancing algorithms, such as resampling and synthetic data generation, were applied to improve the classification accuracy rate. Traditional machine learning models, such as RF and LR, were compared with transformer-based models, namely, BERT and RoBERTa, with ensemble algorithms enhancing overall performance. The research proposed an optimized framework that was integrated into a web-based chat application for real-time monitoring. The agile methodology allowed continuous and iterative improvements to ensure online safety and adaptability for young users.

3.1. Dataset description

This research framework utilizes three well-prepared datasets that formulate the grounds for training sensitive analysis NLP models to detect various emotions. To train and evaluate a robust model for detecting inappropriate messages, a diverse and well-labeled text dataset is essential. Our research required data that cover a range of inappropriate content categories. These datasets are widely recognized: the GoEmotions dataset, a custom-designed lust dataset, and an emotion classification dataset, which is designed to improve emotional understanding and recognition.

To introduce the nuanced "lust" emotion, the lust dataset is combined separately with the GoEmotions and the emotion datasets, which are used to create two distinct combined datasets, namely, dataset 1 (Lust + GoEmotions Dataset) and dataset 2 (Lust + Emotions Dataset). This enriched dataset is used for model training, which enables the recognition of common and complex emotional situations. By combining these three datasets, we created a more comprehensive and challenging training and evaluation environment. This multisource mixed approach ensures that our model not only is tested against standard benchmarks but also is robust in identifying specialized and refined forms of inappropriate content, which is used to improve the model's real-world applicability.

3.1.1. Lust dataset

The lust dataset has been explicitly designed to support the detection of inappropriate and harmful online communication and interactions, particularly those concerning pornography or sexual online content (see Figure 1). This dataset, with 10,000 labeled tweets, was manually verified and categorized according to emotional labels, including the "lust" category, which is used to flag specific harmful content. The tweets were sourced from Twitter (i.e., x.com), which is a platform known for its large amount of user-generated content, because such social media dataset provides coverage of various emotional spectrums. See data availability to view the dataset.

Figure 1
Lust dataset samples

	text	emotion
9990	aint got time hoe tell hop whip shut cadillac doe	8
9991	aint bitch got license	8
9992	aint really trippin money yo shit constantly c...	8
9993	aint stressin hoes know dat	8
9994	aint talking onna phone butch straight texting...	8
9995	aint trying fuck bitch want wings	8
9996	aint mad bitches thats hoes	8
9997	aint mad thats hoes	8

The lust dataset is a single-emotion data source focused on "lust." It is utilized in combination with two other datasets to create improved training datasets, which are referred to as dataset 1 and dataset 2. These datasets facilitate broader and more nuanced emotion detection while keeping the specialized focus required to identify inappropriate online social media content.

3.1.2. GoEmotions dataset

The GoEmotions comprehensive dataset is a data source for analyzing emotions, and it has 43,410 text samples labeled with 28 different and distinct emotions. GoEmotions dataset samples are shown in Table 2. These emotions are categorized from fundamental ones, such as happiness, sadness, and anger, to more nuanced ones, such as pride, embarrassment, and relief. The dataset has been gathered through online sources, including Reddit, which is used to capture a spectrum of emotional expressions, and such variety makes it highly effective for tasks involving deep understanding of emotions.

Dataset 1 is a combination of the GoEmotions and lust datasets, which resulted in a comprehensive dataset with 28 emotions. This combination increases the research dataset's ability to detect both general emotions and context-specific emotions and adds the newly integrated "lust" category to the overall dataset 1.

3.1.3. Emotions dataset

The emotions dataset is another highly valued dataset to deeply analyze and understand a different and broad range of emotions through text. Emotions dataset samples are shown in Table 3. This dataset comes with more than 400,000 Twitter messages and provides a diversity of real-life expressions, and each of these is labeled using one of the six core selected emotions: sadness is 0, joy is 1, love is 2, anger is 3, fear is 4, and surprise is 5. This structure is used to help researchers and developers realize how emotions are conveyed through words, which is very useful for tasks such as sentiment analysis and various emotions detection.

Table 2
GoEmotions dataset samples

Label	Emotion	Count
0	Admiration	4130
1	Amusement	2328
2	Anger	1567
3	Annoyance	2470
4	Approval	2939
5	Caring	1087
6	Confusion	1368
7	Curiosity	2191
8	Lust	10641
9	Disappointment	1269
10	Disapproval	2022
11	Disgust	793
12	Embarrassment	303
13	Excitement	853
14	Fear	596
15	Gratitude	2662
16	Grief	77
17	Joy	1452
18	Love	2086
19	Nervousness	164
20	Optimism	1581
21	Pride	111
22	Realization	1110
23	Relief	153
24	Remorse	545
25	Sadness	1326
26	Surprise	1060
27	Neutral	14219

Table 3
Emotions dataset samples

Label	Emotion	Count
0	Sadness	121187
1	Joy	141067
2	Love	34554
3	Anger	57317
4	Fear	47712
5	Surprise	14972
6	Lust	10000

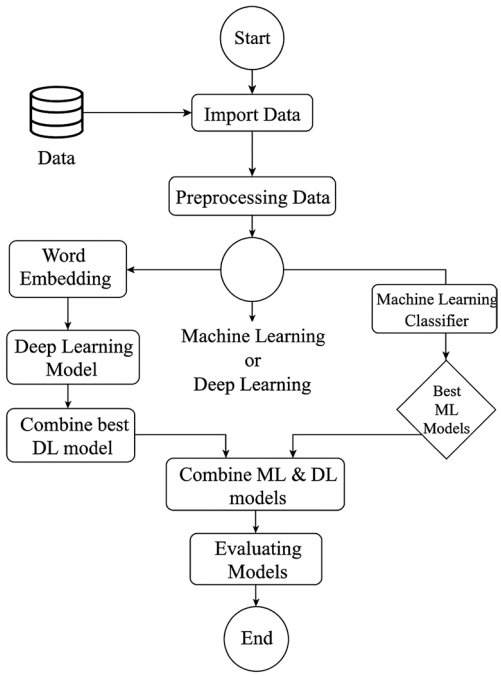
The dataset is made up of two main columns: one with the text of each tweet and another with the emotion label that best represents the feeling behind the message. The six emotion categories cover a broad emotional spectrum, from happiness and excitement to frustration and fear. This makes it a great starting point for anyone looking to explore how people express their feelings online, offering a clear and manageable framework for emotion classification. Whether you are working on machine learning models or simply curious about how emotions show up in text, this dataset provides a solid foundation for your work.

3.2. Research design

The proposed model utilizes a hybrid approach by combining traditional machine learning and deep learning algorithms to classify text based on seven different and distinct emotional categories, including a custom "lust" category. This sentiment analysis NLP approach guarantees a balanced trade-off between interpretability and performance using the strengths of both conventional and transformer-based models.

The design of this research study follows a sequential pipeline, as represented in the Figure 2 model flowchart, which systematically integrates data preprocessing, model training, and evaluation to improve performance [17]. The training phase applies multiple models, which are used to compare traditional algorithms, such as RF and LR, and advanced deep learning algorithms, such as BERT and RoBERTa. Figure 2 shows a detailed illustration of each step in the proposed sentiment analysis NLP model.

Figure 2
The proposed sentiment analysis NLP model



3.2.1. Data import and preprocessing

The model begins by importing (data import) a labeled dataset containing text samples categorized into seven emotion classes to ensure adequate representation of the custom "lust" category (Figure 2).

The preprocessing step of the proposed model involves essential actions to prepare the text data to be analyzed. First is the cleaning process, which removes noise from the text, such as special characters, HTML tags, and redundant whitespaces, to ensure that the dataset is clean from unnecessary content. Afterward, normalization is used to standardize the text by converting it to lowercase characters and applying stemming or lemmatization to limit words to their base forms and simplifying variations in word usage. Tokenization then follows by splitting the text into individual tokens (words or subwords) that can be processed by machine learning algorithms. Finally, to address the issue

of class imbalance, the proposed model utilizes techniques, such as resampling by synthetic minority oversampling technique (SMOTE), to ensure that all emotion categories, including underrepresented ones, are adequately represented for effective model training. Collectively, these steps form a robust foundation for an accurate emotion classification.

3.2.2. Balancing the dataset

After preprocessing, the issue of class imbalance was addressed. For dataset 1, which is characterized by its large volume, manual undersampling was employed to create a balanced class distribution. For dataset 2, it is used to ensure fair and effective training across all 7 emotion categories. This is achieved through a combination of sampling and resampling techniques. First, the existing dataset is combined with new data, shuffled, and organized to integrate all available samples at random. For underrepresented emotions, oversampling techniques such as resampling with replacement are used to ensure that each emotion class meets a minimum threshold of samples. For overrepresented classes, random sampling is applied to limit the number of samples per class to maintain balance without data redundancy. These steps result in a balanced and shuffled dataset where all emotion categories, including the custom "lust" emotion, are effectively represented. This balanced dataset is then used for model training to prevent bias and improve classification accuracy across all classes.

3.2.3. Model selection

The machine learning pathway involves training traditional models, such as RF and LR, on such preprocessed data. Text data are transformed into numerical representations using the feature engineering TF-IDF technique. These numeric features serve as input for the machine learning models. After training, the performance of each model is evaluated using standard metrics to identify the best-performing classifier for the task.

The deep learning model pathway utilizes advanced techniques to have a better understanding of the semantics and context of text. This includes RoBERTa and BERT models, which are trained on this embedded data. RoBERTa is an optimized version of BERT to improve performance using a larger qualitative dataset and more robust training, which allows it to capture subtle nuances in language. BERT is fine-tuned to a specific dataset and excels at understanding the contextual meanings of words. Both deep learning models are particularly effective for nuanced emotion detection because they are powerful tools for analyzing complex emotional expressions in text.

3.2.4. Model combination

The proposed model identifies the best-performing models using machine learning and deep learning pathways by assessing them on metrics that include accuracy, precision, recall, and F1-score. To have the strengths of both approaches, model combination techniques are applied. For example, a stacking approach is used, where the model predictions from both models are fed into a meta-classifier to finalize the model predictions. Otherwise, ensemble methods, such as voting or blending, are used to effectively integrate predictions to improve overall performance.

3.2.5. Model evaluation

Model performance is carefully evaluated using metrics such as accuracy, precision, recall, F1-score, and specificity, with a focus on accurately classifying underrepresented categories, including "lust." Cross-validation is applied to ensure the model's robustness by examining performance across multiple subsets of the dataset, which minimizes the risk of overfitting. In addition, error analysis is conducted to recognize and analyze misclassified instances to enable further refinement of the model and develop its interpretability for practical usage.

3.2.6. Deployment

By having acceptable performance, the fine-tuned models are prepared for deployment. This involves structuring the models into a format suitable for real-world applications. An application programming interface (API) with a user interface is developed to facilitate real-time emotion classification to enable the model to be integrated into applications, e.g., behavior analysis, content moderation, or user sentiment analysis. This deployment step ensures that the model can provide practical value in various business settings.

3.3. Model functionality

The model utilizes advanced transformer-based technology and focuses on RoBERTa, which is a powerful version developed based on BERT and designed to enhance performance by training using large datasets with improved settings. Both RoBERTa and BERT models are bidirectional algorithms, and they review the words by checking before and after each word to understand it. This is important when trying to capture the subtle emotions in such complex sentences. For example, the same word has various emotions depending on its context, and the RoBERTa algorithm is magnificent at understanding such different meanings by analyzing the whole sentence [18, 19].

Alongside RoBERTa, the proposed NLP framework also uses other models, such as RF and LR, for performance comparisons. These models are great for baseline comparisons because they use TF-IDF and bag-of-words techniques to turn text into numbers. While RF and LR are effective in certain cases, they do not capture the deeper and complex relationships between words and their meaning like RoBERTa does.

What makes the framework better is the fine-tuning process, which is applied to the RoBERTa algorithm. The proposed model is trained on a custom dataset that includes seven different emotions along with a "lust" category for applications like content moderation. This fine-tuning supports RoBERTa in better understanding more specific emotions. RoBERTa uses its transformer mechanism to focus on important words or phrases in a sentence, which supports detecting emotions in a more accurate way.

With the combination of RoBERTa and machine learning traditional algorithms, such as RF and LR, the proposed model finds a better balance between complexity and practicality, which makes it more effective for tasks like sentiment analysis, content moderation, and recognizing inappropriate online interactions.

3.3.1. The RoBERTa model

RoBERTa is an advanced variant of the BERT model designed to improve its performance by training on a larger dataset and fine-tuning its hyperparameters. RoBERTa builds on BERT's transformer architecture, but it takes a more aggressive approach in terms of training, including removing the next sentence prediction task and training with much larger mini-batches and longer sequences. This allows RoBERTa to learn more nuanced relationships between words, making it even more powerful for tasks like sentiment analysis and other text classification tasks [4].

The framework fine-tuned the RoBERTa model on our emotion classification dataset, adjusting it to detect seven different emotions, including a custom "lust" category. RoBERTa's ability to capture deep contextual meanings and its strong understanding of the relationships between words in a sentence made it the ideal choice for this task. It outperformed other models, including traditional machine learning models like RF and LR, by better understanding the complex emotional undertones in text. Owing to its robust training and architecture, RoBERTa achieved the highest accuracy in emotion detection, leading us to choose it as the primary model for further development and experimentation.

3.3.2. Random forest

The RF model is a powerful ensemble machine learning method that is particularly well suited for classification tasks. It works by building multiple decision trees during training and then predicting the class that is most predicted by individual trees. We chose this model as one of our traditional machine learning baselines because it is robust against overfitting and can handle high-dimensional data with ease. The RF model was trained using preprocessed text data that had been converted into numerical format using TF-IDF. While it performed reasonably well in terms of recall, it struggled to keep up with more advanced models like BERT when it came to understanding the nuances of language and detecting complex emotions [20].

3.3.3. Logistic regression

LR is another traditional machine learning model that we used as a baseline. It works by modeling the probability of a sample belonging to a particular class by fitting a logistic function to the input data. In this case, we applied LR using a one-vs-rest scheme to handle the multiclass classification of emotions. Like the RF, the LR model was trained and converted into a numerical format using TF-IDF. Despite its simplicity and efficiency, LR had a challenging time capturing the subtleties of emotional context in the text data, which resulted in lower accuracy compared to the deep learning models. However, it still provided a useful benchmark for comparison with more complex methods [18].

3.3.4. The BERT model

Bidirectional encoder representations from transformers (BERT) is a state-of-the-art pretrained deep learning model that excels at understanding the context of words within a sentence by considering both the left and right context simultaneously [19]. Unlike traditional models that treat each word independently, BERT uses a transformer architecture to capture the relationships between words, making it highly effective for tasks like sentiment analysis and emotion detection. For our task, we fine-tuned the model on the emotion classification dataset, customizing it to detect seven emotions, including the custom "lust" category.

3.4. Getting results

The proposed emotion classification NLP model performance was assessed using commonly accepted metrics to determine its effectiveness in accurately recognizing many emotions. These metrics have measurements, including accuracy, precision, recall, F1-score, specificity, and a confusion matrix, to analyze the proposed model performance.

3.4.1. Accuracy

Accuracy measures the overall correctness of the model by calculating the ratio of correctly classified instances (emotions) to the total number of instances. It is given by the formula:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}. \quad (1)$$

1) TP (true positive): the number of correct predictions for a given emotion class. TP presents emotion predictions.
2) TN (true negative): the number of correctly classified instances that do not belong to the given class. TN is the true negative correct, recognized as not belonging to a particular emotion.
3) FP (false positive): instances incorrectly classified as the given class. FP presents emotion predictions.
4) FN (false negative): instances of the given class incorrectly classified as another class. FN presents negative emotion predictions.

3.4.2. Precision

Precision measures the proportion of correctly predicted instances of an emotion to the total instances predicted for that emotion. It highlights the accuracy of positive predictions and is given by the formula:

$$Precision = \frac{TP}{TP+FP}. \quad (2)$$

3.4.3. Recall

Recall, also referred to as sensitivity or the true positive rate, measures the proportion of actual instances of a specific emotion that were correctly predicted, given by the formula:

$$Recall = \frac{TP}{TP+FN}. \quad (3)$$

3.4.4. F1-score

The F1-score is the value of the harmonic mean of precision and recall, which results in a balanced metric when there is an uneven distribution between positive and negative predictions. The F1-score is calculated as:

$$F1\ Score = 2^{*}\frac{Precision^{*}Recall}{Precision+Recall}. \quad (4)$$

3.4.5. Specificity measurement

Specificity is used to measure the model's ability to correctly recognize negative cases of an emotion and is used to refer to the TN rate, which is given by:

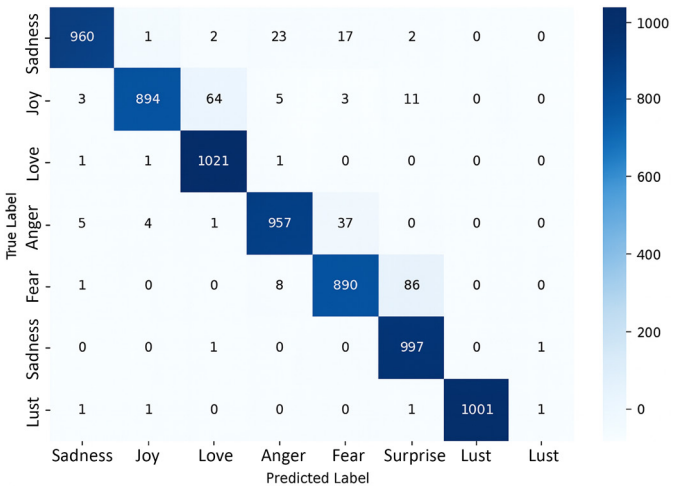
$$Specificity = \frac{TN}{TN+FN}. \quad (5)$$

3.4.6. Class-wise evaluation

The confusion matrix, as shown in Figure 3, provides a detailed breakdown of the model's predictions across all emotion categories. It is a table that presents the following for each class: TP, TN, FP, and FN.

By visualizing the confusion matrix, a high overall fidelity for the RoBERTa model across emotional categories has been found (Figure 3). The model exhibited exceptional performance on highly distinct emotional states, achieving near-perfect classification for "love" with 1021/1023 TP and "surprise" with 997/998 TP. Similarly, the model proved robust in detecting the key negative and inappropriate classes, with high identification rates for "lust" and "anger," with 1001

Figure 3
Confusion matrix for the RoBERTa model



TP. These results confirm that the linguistic features of an emotion are clearly delineated, and the RoBERTa architecture is highly effective at extraction and classification.

By looking at the critical "lust" category from Table 4, the model's performance was outstanding. This class achieved a perfect 100% across all key metrics, precision, recall, and F1-score, indicating that the model made zero misclassifications for this type of content in the test set. Given that identifying inappropriate content was a primary goal of this work, this result confirms the framework's maximum effectiveness and reliability for such an important task. While the overall F1-score is 95% due to slight confusion in other emotion classes like "fear" and "joy," an F1-score of 100% for "lust" validates the system's core ability to protect against highly sensitive content.

Table 4
Class-wise results

Emotion	Precision	Recall	F1-score	Support
Sadness	99%	96%	97%	1005
Joy	99%	91%	95%	980
Love	94%	100%	97%	1024
Anger	96%	95%	96%	1004
Fear	94%	90%	92%	985
Surprise	91%	100%	95%	998
Lust	100%	100%	100%	1004

Despite the strong overall performance, there are some limitations. The primary source of misclassification occurred between high-arousal negative states, specifically the confusion of "fear" with "surprise" (86 instances). Further, the model struggled to differentiate between positive sentiments, misclassifying 64 instances of "joy" as "love," suggesting a common challenge in separating closely related positive affect.

3.5. Multifactor evaluation

To evaluate the performance of the selected sentiment analysis models, a multifaceted approach has been applied by comparing four models, namely, LR, RF, BERT, and RoBERTa. Each model was validated by evaluating its ability to classify text into one of the seven emotional categories, including the custom "lust" category. These are the steps during the training and evaluation of the models. See Table 5 for the results and comparisons.

3.5.1. Training and evaluation

This combined research dataset has been divided into 80% training, 10% testing, and 10% validation datasets to have a robust evaluation process. These splits are used to train the transformer-based models, i.e., BERT and RoBERTa, while the traditional machine learning models, i.e., LR and RF, were trained using cross-validation techniques. In addition, five-fold cross-validation has been used to ensure performance consistency [20, 21].

3.5.2. Insights from results

RoBERTa model classification results have outperformed other models in all metrics, which proves its remarkable capability to handle such complex emotional detection scenarios and text understanding. As shown in Table 5, RoBERTa has achieved the highest accuracy of 95.15% with 95.30% precision and 95.13% recall, which confirms its effective performance in detecting and recognizing a wide range of emotional categories.

While the BERT model is slightly behind RoBERTa, it still performed well, with an accuracy of 95.08%, a precision of 95.20%, and a recall of 95.04%. This shows that BERT is still a powerful model for emotion detection, but RoBERTa's improvements in training make it a bit more accurate. Conversely, LR and RF performed less effectively. LR achieved an accuracy of 91.34%, and RF had an accuracy of 88.00%. Both models struggled with the complexity of this multilabel emotion detection task, as evidenced by their lower precision and F1-scores compared to RoBERTa and BERT. This suggests that while these models are useful for simpler tasks, they face challenges in capturing the deep nuances required for emotion classification.

A comparison analysis has been conducted, as shown in Table 6, which provides detailed comparisons for various recent and related studies in text classification and emotion detection. This research study applied diverse methods, which ranged from older lexicon-based models as described by Angulu et al. [11] to traditional machine learning by Martín-Valdivia et al. [14], as well as Mohammad and Turney [15], studies, and modern deep learning models, such as LSTM and transformer architectures [22, 23]. This pragmatic research philosophy provides a broad and robust contextual benchmark for assessing our results.

The data clearly demonstrate the effectiveness of leveraging large pretrained models. Our framework, utilizing the fine-tuned RoBERTa and BERT architectures on our combined datasets, achieves a significantly higher accuracy of 95.15%. This score not only surpasses older methods, which struggled with complex emotions and scored below 70%, but also exceeds hybrid deep learning approaches, such as the study reported by Brauwers and Frasincar [21] and contemporary transformer-based models, such as the study reported by Abdullah and Ahmet [22]. While direct comparison is complicated by differences in target classes and dataset domains, these results firmly position our NLP framework as a highly competitive, state-of-the-art solution for multifaceted and sensitive text content analysis.

Overall, RoBERTa showed the best performance for emotional detection purposes, as illustrated in Figure 4, with 95% across all levels in the classification matrix.

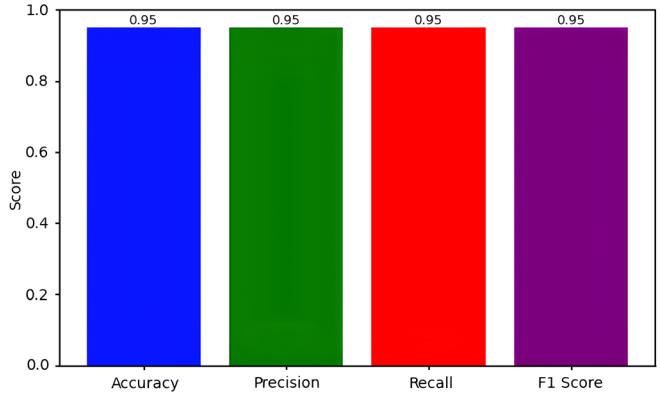
Table 5
Classification results of different sentiment analysis models

Model	Embedding	Dataset	Accuracy	Precision	F1-score	Recall
RoBERTa	RoBERTa's transformer-based embeddings	Dataset 1	53.5%	51.7%	53.5%	52.0%
		Dataset 2	**95.15%**	**95.30%**	**95.15%**	**95.13%**
Logistic regression	TF-IDF	Dataset 1	45.84%	45.51%	45.84%	45.39%
		Dataset 2	**91.34%**	**91.40%**	**91.34%**	**91.28%**
Random forest	TF-IDF	Dataset 1	42.83%	41.42%	42.83%	41.87%
		Dataset 2	**88.00%**	**87.96%**	**88.00%**	**88.87%**
BERT	BERT embeddings	Dataset 1	53.86%	52.30%	53.86%	52.69%
		Dataset 2	**95.08%**	**95.20%**	**95.08%**	**95.04%**

Table 6
Comparison analysis of recent studies

Research study	Year	Dataset	Core methodology	Accuracy
Our research study	2025	All datasets	RoBERTa & BERT	95.15%
Nandwani and Verma [9]	2020	SemEval-2018	CNN	87%
Chatzakou et al. [12]	2017	Twitter & Facebook	Lexicon-based approach	59.38%
Martín-Valdivia et al. [14]	2013	MuchoCine	SVM	86%
Mohammad and Turney [15]	2020	Emolex	Crowdsourcing	66.67%
Brauwers and Frasincar [21]	2022	Online phone review	LSTM + LDA	89.5%
Abdullah and Ahmet [22]	2022	Sentihood	BERT	94%
Sachin et al. [23]	2020	Amazon review	LSTM	70%

Figure 4
Classification matrix for the RoBERTa model



4. Discussion and Findings

Evaluating diverse models for detecting emotions revealed that the RoBERTa model has achieved outstanding performance by reaching the highest metrics in accuracy at 95.15%, precision at 95.30%, and recall at 95.13%. Such results emphasize RoBERTa's success in recognizing and classifying emotional categories by showing its high capability in handling the complexities of different emotional detection. RoBERTa's transformer-based architecture has contributed to such high performance, which is able to capture deep contextual relationships in text and sentence. The RoBERTa model can process text bidirectionally, which allows it to find word meaning based on its surrounding context. Such a capability to interpret subtle relationships between words is important to detect emotions, where minor differences in language regularly signal diverse states of emotions.

BERT, while slightly behind RoBERTa, also performed strongly with an accuracy of 95.08%, a precision of 95.20%, and a recall of 95.04%. This performance confirms that transformer-based models, such as BERT, are highly effective for emotion-detection tasks [24]. However, RoBERTa's refinements in training contributed to its slight edge over BERT in this task. In contrast, traditional machine learning models like LR and RF performed less effectively in this multilabel emotion detection task. LR achieved an accuracy of 91.34%, and RF had an accuracy of 88%. Although these models are computationally efficient and easier to implement, they struggled to capture the complex contextual relationships present in the data, leading to lower precision, recall, and F1-scores compared to RoBERTa and BERT.

As shown in Table 5, we observed a significant positive shift in dataset 2 (Lust + Emotions dataset) compared to those obtained with dataset 1 (Lust + GoEmotions dataset). This change mainly happened because of several factors, including its larger size, the presence of more complex emotional expressions, and its unbalanced distribution.

Finally, it is obvious that RoBERTa has achieved high performance for emotion detection purposes, which revealed the advantages of transformer-based architectures for understanding and capturing different and complex emotional nuances in text. This proves the importance of utilizing advanced contextual understanding in models for high accuracy and interpretability emotion detection applications.

5. Limitations

Even though the current models achieved promising results, it has been recognized that they might not be the absolute best fit for text sentiment analysis and emotion detection. Our work was constrained by several practical limitations: difficulties managing the sheer volume of data, performance dips when classifying a wide spectrum of emotions, the need for substantial time investments, and insufficient computational resources to fully explore and fine-tune state-of-the-art architectures, such as HuBERT, LSTM, and LLMs.

In addition to the primary limitations discussed, our team managed several smaller, practical issues. Finding the optimal model architecture for our specific application was a complex selection process that required considerable time. The study faced the persistent difficulty of locating an ethically and technically suitable offensive dataset, a resource crucial for the research's scope. On the logistical side, typical research bottlenecks emerged, such as minor time losses when multiple experiments had to queue for limited computational power and the demanding effort required for team-based debugging when combining codes from different contributors.

6. Recommendations and Further Research

It is recommended that further studies to enhance the proposed NLP model must be conducted by combining emotion detection using text and audio [25]. With the growth of voice messages and audio communication and interactions on social media, integrating audio emotion detection becomes important to get a complete analysis of online interactions. The use of advanced audio processing models, such as speech-to-text and sentiment analysis models tailored for voice, can help the model detect emotions expressed vocally [20].

Fine-tuning state-of-the-art sentiment analysis NLP models, such as GPT-4 and multilingual models, e.g., multilingual BERT (mBERT), can be used to boost text analysis, and combining them with advanced audio emotion recognition tools could improve the model's accuracy and adaptability across different languages and modes of communication. In addition, training these models using powerful GPUs can enhance the model's analytic capability for both written and spoken contents [26].

Expanding the proposed model's language abilities to support multiple languages, including in-text and audio formats, can formulate an enhanced version of the model that is more inclusive and relevant to a wider audience. Further, making the model detect various accents, dialects, and speech patterns is vital for global applications.

To robustly address challenges stemming from imbalanced data and the need for greater methodological innovation, this research plans to investigate generative adversarial network (GAN)-based techniques, such as DiGAN, to effectively mitigate class imbalance through high-fidelity synthetic data generation, thereby complementing traditional oversampling methods. Concurrently, the research will implement cost-sensitive learning techniques to ensure that the model places higher importance on the critical, minority classes, such as "lust" and "anger," for promoting fairness and greater accuracy. Further, exploring capsule-based architectures will be key to handling subtle, multilayered emotional patterns more robustly [24–26]. Finally, conducting real-world practical testing, continuous validation, and actively gathering user feedback will be essential to prove the framework's performance and ensure its adaptability to new trends and linguistic changes.

In conclusion, it is important to address ethical considerations, such as bias reduction and safeguarding privacy, to ensure reliable model deployment. By combining emotion detection from text and audio, the model can be a comprehensive method to examine online interactions and ensure the safety of online users.

7. Conclusion

This research study sheds light on the importance of integrating emotion detection with the age verification model to enhance youngsters' online safety. The NLP framework applied the RoBERTa model to monitor online communication and interactions to classify harmful conversations, such as sexually improper behavior. By analyzing emotional patterns in text conversations, the framework can be utilized to detect potential threats and act, for instance, to block accounts and notify guards when needed to protect youngsters. The proposed NLP framework has been evaluated using advanced models, including the RoBERTa and BERT models, which achieved outstanding accuracy levels and high-performance metric results. The RoBERTa model had an accuracy of 95.15%, which demonstrated its capability to effectively classify and recognize different ranges of emotional behaviors. While the model's performance is promising, many other challenges were recognized, and these include handling imbalanced data and the significant computational resources required for training such advanced models.

Further research studies can be conducted to make extra enhancements to the model's accuracy and efficiency along with audio emotion detection and integration of multilanguage supporting models, which expands its abilities to support multiple languages and ensures fairness and transparency in its decision-making processes. This research lays the groundwork to use advanced NLP technologies to protect online environments for youngsters and provides valuable insights for future developments.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

The data that support the findings of this study are openly available in Kaggle at https://www.kaggle.com/datasets/mrmorj/hate-speech-and-offensive-language-dataset, https://www.kaggle.com/datasets/debarshichanda/goemotions, and https://www.kaggle.com/datasets/bhavikjikadara/emotions-dataset.

Author Contribution Statement

Hisham AbouGrad: Conceptualization, Methodology, Validation, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration. **Sankar Santhosh:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Visualization. **Salem Alsaid:** Writing – review & editing, Visualization.

References

[1] Greyson, D., Chabot, C., Mniszak, C., & Shoveller, J. A. (2023). Social media and online safety practices of young parents. *Journal of Information Science*, *49*(5), 1344–1357. https://doi.org/10.1177/01655515211053808

[2] Kloess, J. A., Seymour-Smith, S., Hamilton-Giachritsis, C. E., Long, M. L., Shipley, D., & Beech, A. R. (2017). A qualitative analysis of offenders' modus operandi in sexually exploitative interactions with children online. *Sexual Abuse*, *29*(6), 563–591. https://doi.org/10.1177/1079063215612442

[3] da Silva, N. F. F., Hruschka, E. R., & Hruschka, E. R. (2014). Tweet sentiment analysis with classifier ensembles. *Decision Support Systems*, *66*, 170–179. https://doi.org/10.1016/j.dss.2014.07.003

[4] Pranay Kumar, B. V., & Sadanandam, M. (2024). A fusion architecture of BERT and RoBERTa for enhanced performance of sentiment analysis of social media platforms. *International Journal of Computing and Digital Systems*, *15*(1), 51–66.

[5] O'Carroll Bantum, E., Elhadad, N., Owen, J. E., Zhang, S., Golant, M., Buzaglo, J., …, & Giese-Davis, J. (2017). Machine learning for identifying emotional expression in text: Improving the accuracy of established methods. *Journal of Technology in Behavioral Science*, *2*(1), 21–27. https://doi.org/10.1007/s41347-017-0015-5

[6] Holtgraves, T. (2022). Implicit communication of emotions via written text messages. *Computers in Human Behavior Reports*, *7*, 100219. https://doi.org/10.1016/j.chbr.2022.100219

[7] Hartikainen, H., Iivari, N., & Kinnula, M. (2019). Children's design recommendations for online safety education. *International Journal of Child-Computer Interaction*, *22*, 100146. https://doi.org/10.1016/j.ijcci.2019.100146

[8] Kalyan, K. S. (2024). A survey of GPT-3 family large language models including ChatGPT and GPT-4. *Natural Language Processing Journal*, *6*, 100048. https://doi.org/10.1016/j.nlp.2023.100048

[9] Nandwani, P., & Verma, R. (2021). A review on sentiment analysis and emotion detection from text. *Social Network Analysis and Mining*, *11*(1), 81. https://doi.org/10.1007/s13278-021-00776-6

[10] Ahmad, Z., Jindal, R., Ekbal, A., & Bhattachharyya, P. (2020). Borrow from rich cousin: Transfer learning for emotion detection using cross lingual embedding. *Expert Systems with Applications*, *139*, 112851. https://doi.org/10.1016/j.eswa.2019.112851

[11] Angulu, R., Tapamo, J. R., & Adewumi, A. O. (2018). Age estimation via face images: A survey. *EURASIP*

Journal on Image and Video Processing, *2018*(1), 42. https://doi.org/10.1186/s13640-018-0278-6

[12] Chatzakou, D., Vakali, A., & Kafetsios, K. (2017). Detecting variation of emotions in online activities. *Expert Systems with Applications*, *89*, 318–332. https://doi.org/10.1016/j.eswa.2017.07.044

[13] AbouGrad, H., Chakhar, S., & Abubahia, A. (2023). Decision making by applying machine learning techniques to mitigate spam SMS attacks. In *Key Digital Trends in Artificial Intelligence and Robotics: Proceedings of 4th International Conference on Deep Learning, Artificial Intelligence and Robotics*, 154–166. https://doi.org/10.1007/978-3-031-30396-8_14

[14] Martín-Valdivia, M.-T., Martínez-Cámara, E., Perea-Ortega, J.-M., & Ureña-López, L. A. (2013). Sentiment polarity detection in Spanish reviews combining supervised and unsupervised approaches. *Expert Systems with Applications*, *40*(10), 3934–3942. https://doi.org/10.1016/j.eswa.2012.12.084

[15] Mohammad, S. M., & Turney, P. D. (2010). Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, 26–34.

[16] Wadawadagi, R. S., & Pagi, V. B. (2022). Sentiment analysis on social media: Recent trends in machine learning. In Information Resources Management Association (Ed.), *Research anthology on implementing sentiment analysis across multiple disciplines* (pp. 780–799). IGI Global.

[17] Bharti, S. K., Varadhaganapathy, S., Gupta, R. K., Shukla, P. K., Bouye, M., Hingaa, S. K., & Mahmoud, A. (2022). Text-based emotion recognition using deep learning approach. *Computational Intelligence and Neuroscience*, *2022*(1), 2645381. https://doi.org/10.1155/2022/2645381

[18] Sinha, A., Rout, B., Mohanty, S., Mishra, S. R., Mohapatra, H., & Dey, S. (2024). Exploring sentiments in the Russia-Ukraine conflict: A comparative analysis of KNN, decision tree and logistic regression machine learning classifiers. *Procedia Computer Science*, *235*, 1068–1076. https://doi.org/10.1016/j.procs.2024.04.101

[19] Dias Souza, F., & Baptista de Oliveira e Souza Filho, J. (2022). BERT for sentiment analysis: Pre-trained and fine-tuned alternatives. In *Computational Processing of the Portuguese Language: 15th International Conference*, 209–218. https://doi.org/10.1007/978-3-030-98305-5_20

[20] Sanchez-Medina, J. J. (2024). *Sentiment analysis and random forest to classify LLM versus human source applied to scientific texts*. arXiv. https://doi.org/10.48550/ARXIV.2404.08673

[21] Brauwers, G., & Frasincar, F. (2023). A survey on aspect-based sentiment classification. *ACM Computing Surveys*, *55*(4), 65. https://doi.org/10.1145/3503044

[22] Abdullah, T., & Ahmet, A. (2023). Deep learning in sentiment analysis: Recent architectures. *ACM Computing Surveys*, *55*(8), 159. https://doi.org/10.1145/3548772

[23] Sachin, S., Tripathi, A., Mahajan, N., Aggarwal, S., & Nagrath, P. (2020). Sentiment analysis using gated recurrent neural networks. *SN Computer Science*, *1*(2), 74. https://doi.org/10.1007/s42979-020-0076-y

[24] Wu, S., & Dredze, M. (2020). Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, 120–130.

[25] Zhang, H., Huang, H., Zhao, P., Zhu, X., & Yu, Z. (2024). CENN: Capsule-enhanced neural network with innovative metrics for robust speech emotion recognition. *Knowledge-Based Systems*, *304*, 112499. https://doi.org/10.1016/j.knosys.2024.112499

[26] P R, J. D., Venkatraman, S., Sharma, V., & Malarvannan, S. (2024). *Multimodal emotion recognition using audio-video transformer fusion with cross attention*. arXiv. https://doi.org/10.48550/arXiv.2407.18552

How to Cite: AbouGrad, H., Santhosh, S., & Alsaid, S. (2026). NLP Framework to Safeguard Youngsters Online Using Advanced Transformer-Based Models. *Journal of Data Science and Intelligent Systems*. https://doi.org/10.47852/bonviewJDSIS62025752