International Journal of Computational and Experimental Science and ENgineering (IJCESEN)

Vol. 11-No.3 (2025) pp. 5535-5541
http://www.ijcesen.com

ISSN: 2149-9144

Research Article

From Lexical to Semantic: A Systematic Review of off-Topic Detection Evolution in Automated Assessment

Hanadi Hamed Abdelrahman[1*], Salma A. Mahmood[2]

[1] Computer science department, College of Computer Science and Information Technology, Basrah, Iraq.
*** Corresponding Author Email**: hanadi.hamed@uobasrah.edu.iq **- ORCID:** 0000-0002-5247-7150

[2] Computer Information Systems Department, College of Computer Science and Information Technology, Basrah, Iraq.
Email: Salma.mahmood@uobasrah.edu.iq**- ORCID:** 0000-0002-5247-7250

Article Info:

DOI: 10.22399/ijcesen.3382
Received : 05 May 2025
Accepted : 06 July 2025

Keywords

Lexical to Semantic
off-Topic Detection Evolution
Automated Assessment

Abstract:

Off-topic detection is a vital challenge in natural language processing, especially as online learning, automated assessments, and AI-powered content platforms continue to expand – it ensures that learner responses or user-generated content align with intended prompts or topics Its applications span far beyond essay scoring to include spoken response assessment, educational dialogue systems, business writing analysis, and even harmful content moderation on social platforms This research aims to develop and enhance an off-topic detection model using AraBERT embeddings for Arabic student responses, focusing on boosting automatic assessment accuracy and elevating the overall robustness of educational AI systems. The findings demonstrate that our model reliably identifies relevance in short-answer tasks, achieving precision and recall comparable to or better than existing methods, suggesting that future work integrating deeper semantic, structural, and discourse-level features may further enhance assessment capabilities. [1], [2], [3], [4].This review aims to survey a range of studies in this domain, identifying the methods employed, assessing their significance, exploring the limitations they face, and highlighting emerging directions in the evolution of off-topic detection techniques within automated assessment systems.

1. Introduction

Detection of off-topic content is a crucial problem in natural language processing (NLP), particularly in the context of automated systems that evaluate written essays. The key task involves determining whether a response contains information that does not relate to the set topic. This function is important for maintaining the accuracy and usefulness of evaluation tools in school settings. This work is crucial for maintaining the accuracy and integrity of automated assessments, as responses that are grammatically correct but nonsensical could result in inflated scores without it.

Off-topic content identification techniques have traditionally analyzed content using sentence length metrics and word occurrence statistics, which are usually framed in vector space models. These simplistic approaches fail to capture the deep semantic meaning of the text, hence their inability to accurately evaluate it. More advanced methods, utilizing deep learning and neural networks, have been developed to address these challenges. Exploring the integration of neural attributes and surface characteristics using convolutional neural networks (CNNs), Siamese networks, and attention mechanisms can enhance semantic understanding. For instance, a model that integrates a Siamese network with a cross-attention mechanism has demonstrated exceptional efficacy on benchmark datasets, achieving precision and recall metrics that surpass 90%.

In a similar vein, the application of Sentence-BERT alongside cosine similarity has yielded markedly improved outcomes in delineating the semantic correspondence between an essay and its corresponding prompt, significantly outperforming conventional models [4], [5], [6], [7]

Automated Essay Scoring (AES):

Off-topic detection is essential for improving the reliability of automated essay grading systems by

ensuring that irrelevant yet grammatically correct responses do not receive inflated scores. This is especially valuable in large-scale educational assessments [5].

Social Media Content Moderation:

Identifying off-topic or harmful content helps maintain healthy digital environments by filtering spam, misinformation, and irrelevant posts. This has become increasingly important due to the rapid growth of user-generated content [4].

Spoken Response Evaluation:

Off-topic detection is used in language learning systems to automatically assess the relevance of spoken answers, thereby enhancing the quality of oral exams and language proficiency tests [7].

IT Support and Helpdesk Systems:

In technical support conversations, detecting off-topic utterances ensures that automated chatbots and agents focus on relevant user issues, improving response efficiency and customer satisfaction [6].
The current research is organized into several key sections. The first part presents a general introduction to the topic, followed by a review of related previous works. The study concludes with a summary of key findings, an outline of current limitations, and suggestions for future improvements.

2. Related Work

Automated off-topic response detection as a domain has undergone significant changes in terms of methods, transitioning from traditional, linguistically based algorithms to deep learning architectures. This survey of the relevant literature highlights many notable milestones within this domain by systematically investigating the trajectory of the field, from its initial solutions in search engines to present-day hybrid models that combine document representation techniques with rich semantic analysis of context. This survey indicates a shift in methodology from simpler, essentially statistical models to increasingly sophisticated AI systems, accompanied by a notable increase in accuracy (i.e., F1 scores increased from 77% to 97%); however, considerable obstacles remain, such as model explainability, algorithmic bias, and cost. Through a critical review of the recent history of automated off-topic response detection, we develop a systematic taxonomy of methods and review and contrast performance standards, identifying gaps and issues for future research. The research framework in this survey summarizes the state of knowledge and offers further pathways toward more effective, socially equitable, and practically deployable solutions for automated assessment systems.
Goharian and Platt (2007) decisively established the Detection of Off-Topic Search (DOTS) methodology, a powerful tool for identifying off-topic searches through the innovative clustering of search engine results and user profile comparisons. This methodology utilizes agglomerative hierarchical clustering to organize results into distinct topical clusters systematically. These clusters are rigorously evaluated against either predefined or dynamically generated user profiles to measure their relevance. The system effectively retains only the most significant clusters. It ranks them based on their similarity to the user's query, employing a term frequency and normalized inverse document frequency (TF-IDF) weighting schema. The effectiveness of the method was monitored using the traditional performance metrics precision, recall, and F1-score, and the results from the NIST TREC dataset provide clear evidence of a significantly improved detection accuracy about the regular RF2 process, with a slight reduction in coverage, demonstrating yet again the often-important trade-offs between accuracy and coverage. Statistical tests confirmed that these performance enhancements are not just noteworthy but statistically significant. While the system has proven successful, it faces specific challenges, including the computational complexity of performing real-time hierarchical clustering and the demand for highly accurate user profiles, which can lead to an over-reliance on the clustering selection process. Nonetheless, this study indicates that result clustering represents a viable and effective strategy for enhancing off-topic detection accuracy within search systems.
Li, Wen, and Pan (2017) introduced an unsupervised methodology for off-topic essay detection based on a dual-level semantic analysis: the first level measures essay-to-target prompt similarity, while the second compares this similarity against reference prompts to generate an on-topic score. Tested on six Kaggle datasets, this approach demonstrated superior accuracy compared to traditional single-prompt methods through its comparative semantic framework that identifies topical distinctions without requiring labeled training data. However, the method's effectiveness remains contingent upon the quality of underlying language representation models and suffers from adaptive inflexibility. In addition, its generalizability is limited due to reliance on a single competition dataset and single

prompts. However, this study represents a significant methodological improvement and can inform future work, including the combination of supervised systems or the improvement of semantic modeling for enhanced flexibility and generalizability.

Qu et al. (2018) developed an algorithm specifically designed for detecting off-topic compositions in English writing instruction systems within the Chinese context. This algorithm combines Latent Dirichlet Allocation (LDA) for topic modeling with word2vec embeddings to understand the semantic relationships among words. The methodology includes several preprocessing steps, such as segmenting words, removing stop words, and eliminating punctuation. After these steps, Latent Dirichlet Allocation (LDA) is used to model document topics. Additionally, word2vec is utilized to enhance semantic understanding, which enables the calculation of cosine similarity between word vectors and the feature words of the topics. Gibbs sampling is utilized to estimate the parameters of LDA proficiently. The efficacy of the algorithm was substantiated through validation against a dataset comprising 1,230 college English compositions spanning six distinct themes, achieving an average accuracy of 91.86%. A recall rate of 88.78% and an F-measure of 89.81% were attained in this study. These findings exceed those yielded by conventional TF-IDF methodologies, which achieved a mean F-measure of 77.4%. The investigation further revealed that selecting five feature words per topic, alongside specifying the number of topics (K) as 15, maximized detection precision. The results demonstrate the effectiveness of the approach used, but also reveal methodological challenges related to determining the optimal number of topics and adopting subjective manual evaluation criteria. This suggests that future research should prioritize developing more objective quantitative assessment measures and evaluating the algorithm's efficiency across various writing contexts.

Yang et al. (2018) proposed a hybrid model for off-topic text detection, integrating neural network-based semantic representations with conventional textual features. The model architecture incorporates a three-layer convolutional neural network (CNN) followed by three max-pooling layers designed to extract high-level semantic patterns from input texts. Word embeddings were generated using the skip-gram model, enabling the encoding of lexical semantics into dense vector representations that preserve contextual relationships. To enhance feature representation, the model combines deep neural features derived from the CNN with surface-level features generated via TF-IDF, forming a multi-dimensional text representation framework. These fused features are subsequently classified through a SoftMax layer for final prediction. Their model was trained on an annotated dataset of 8,106 texts sourced from Kaggle and evaluated using standard performance metrics, including precision, recall, and F1-score. Comparative analysis across four distinct writing prompts demonstrated the model's superiority over conventional CNN baselines, with the F1-score increasing by 3.88 percentage points on Prompt #3 and achieving an average improvement of 1.7 percentage points across all prompts. The method of adding TF-IDF features has reliably provided better classification accuracy, confirming the value of classical feature engineering methods that contribute to and complement neural models in various ways. Nevertheless, this study has also identified several of its limitations. The observed reduction in recall performance for Prompts #1 and #2 indicates inherent difficulties in identifying off-topic content within highly divergent thematic contexts. Furthermore, the dependence on manually annotated datasets introduces potential subjectivity in labeling, while the model's structural complexity—arising from the integration of multiple features—compromises its interpretability. Limitations to dataset diversity and potential overfitting also limit the model's scalability. To address these limitations, the authors suggest future work to (1) create richer feature representations and (2) implement structured evaluation frameworks to enhance learning experience off-topic detection performance across different educational contexts.

Li and Wen (2019) present an unsupervised methodology for detecting off-topic essays based on topic clustering analysis. The approach initially extracts keywords from both student compositions and writing prompts and then calculates a topic correlation coefficient between them. The system utilizes the results of the cluster analysis to determine which classification thresholds to use, enabling text to be partitioned into topic-relevant and irrelevant categories without requiring labeled training data. Although the methodology offers substantial benefits in terms of scalability and applicability across diverse situations, it has some methodological limitations. The main limitations are the reliance of the methodology on the validity of keyword extraction, the sensitivity of classification thresholds to different dataset characteristics, and a lack of evidence regarding cross-topic generalizability. The study suggests that future work should include the development of broad-scope comparative studies and explorations into possible accuracy improvements, as well as potential engagement with semi-supervised systems, to support further effectiveness. This approach

represents a valuable contribution to automated educational assessment, though further development is necessary to ensure reliability in practical applications. Key areas for future research include the optimization of threshold techniques, the development of hybrid models, and the extensive validation of models across multiple subject domains and writing styles to establish robust performance benchmarks. The unsupervised nature of the algorithm makes it particularly suitable for deployment scenarios where labeled training data is scarce or unavailable.

Wang et al. (2019) developed an automated system for detecting off-topic spoken responses in high-stakes assessments using deep convolutional neural networks (CNNs), which transform the response-prompt relationship into a similarity grid representation analyzed through Inception networks. The system demonstrated superior performance (F1-score: 92.8%) compared to word embedding-based baselines (85.5%), highlighting its effectiveness in topical relevance detection. While this approach enhances assessment validity, it presents three notable limitations: (1) significant computational demands that may constrain real-time implementation, (2) inherent interpretability challenges characteristic of deep learning models, and (3) exclusion of other critical speaking proficiency dimensions (e.g., fluency, pronunciation). Despite these limitations, the research suggests that combining deep convolutional neural networks and similarity grid analysis can be leveraged to develop automated assessment systems. Moreover, the research suggests that future work can be done to optimize computation as well as combine multi-dimensional assessments without losing detection accuracy. The findings contribute to the development of more sophisticated automated assessment tools, though further refinement is needed for practical educational applications.

Shahzad and Wali (2022) systematically investigated automated off-topic essay detection through a comprehensive comparison of embedding techniques and classification models. Their methodology employed five distinct semantic representation approaches—Word Mover's Distance (WMD), Average Word Embeddings (AW), IDF Weighted Embeddings (IW), Similarity Grid Model (SGM), and Siamese Convolutional Neural Networks (SCNN)—which were subsequently processed by four classifiers (Logistic Regression, Random Forest, Support Vector Machine, and CNN) across ten benchmark datasets. The optimal configuration, combining WMD, AW, and IW with Random Forest, achieved 93.5% accuracy, while the SCNN-SVM pairing demonstrated superior performance with an average accuracy of 97%. Although the results confirm the technical feasibility and scalability of automated detection systems, particularly for online education applications, the study identified three critical limitations: (1) variability in dataset characteristics affecting generalizability, (2) substantial computational requirements for deep learning implementations, and (3) sensitivity to prompt distribution patterns. The results provide a key advance in the field as they are an empirical validation of embedding classifier collaborative advantages while addressing significant practical considerations for application in educational assessment contexts. This work presents a methodological framework for future research that examines more hybridized approaches to automated essay evaluation

Fan and colleagues (2023) introduced an advanced method for off-topic detection utilizing a Siamese network architecture combined with cross-attention mechanisms. The Siamese architecture utilizes two inputs and is effective for tasks that require pairwise comparisons. In the authors' case, the prompt and student response are treated as two inputs of the Siamese architecture with dynamic cross-attention between both inputs to align with key semantic tokens in the prompt and response. The student's response is not simply matched against the prompt's features and, thus, significantly outperforms traditional feature-based approaches but instead represents a far more contextual understanding. Evaluation on the Mohler dataset demonstrated exceptional performance (F1: 91.9%, Recall: 92.5%, Precision: 94.1%), outperforming BERT-based benchmarks by 3.7% in F1-score and 3.0% in precision. However, the model exhibits three principal constraints: (1) substantial computational requirements limiting practical deployment, (2) restricted generalizability due to single-dataset validation, and (3) inherent interpretability challenges from its complex "black box" architecture. While proving particularly effective for short-answer assessment, its applicability to extended responses remains unverified. This research presents a significant advancement in educational assessment technology through the innovative integration of attention mechanisms with comparative neural architectures, though further investigation is required to enhance interpretability, expand validation across diverse datasets, and adapt the framework for comprehensive response evaluation scenarios. The study lays a foundation for possible future developments in systems of automated essay assessment, explicitly highlighting the success of the attention-comparison hybrid model for gauging student essays in an educational context.

Das, Vadi, and Yadav (2024) propose an integrated transformer-based framework (AOES) that jointly addresses automated essay scoring and off-topic detection through three key innovations: (1) a BERT architecture enhanced with a Topic Regularization Module (TRM) replacing conventional regression layers, (2) an unsupervised Mahalanobis distance metric operating on transformer latent features for off-topic detection, and (3) a multi-task learning paradigm with hybrid loss (MSE + topic regularization) optimizing both tasks simultaneously. Evaluation across the ASAP-AES and Psywar-Essay datasets demonstrates superior performance (F1 improvement = 0.18 ± 0.03) and adversarial robustness against content manipulation strategies. The TRM-hybrid loss combination proves particularly effective, achieving 92.7% topic discrimination accuracy while maintaining scoring consistency (κ = 0.85). However, three critical limitations emerge: (i) on-topic training data dependency reduces generalization to high off-topic prevalence environments (performance degradation δ = 0.22 at 40% off-topic content), (ii) unaddressed potential societal biases due to absent demographic metadata, and (iii) computational complexity ($O(n^3)$ for Mahalanobis calculations) potentially limiting large-scale deployment. The study makes significant theoretical contributions by demonstrating how shared representation learning can effectively couple assessment and relevance detection tasks while highlighting important practical considerations for educational technology implementations regarding bias mitigation and computational feasibility. Future research directions should investigate demographic-aware training protocols and optimized distance metrics to address current scalability constraints.

Zhao (2024) developed a comprehensive framework for off-topic composition detection leveraging advanced natural language processing and machine learning techniques. The methodology employs a multi-stage approach: (1) Biterm-LDA topic modeling extracts latent thematic structures from a composition corpus, (2) Doc2vec generates semantic document embeddings, and (3) feature fusion creates unified representations capturing both topical relevance and contextual meaning. A twin-network multilayer perceptron architecture performs simultaneous dimensionality reduction and feature enhancement, enabling classification based on vector space proximity to dynamically calculated topic centroids. The structure implements adaptive thresholding through the optimization of a ROC curve at the second layer of the network, thus automating variance to achieve a consistent performance across domains or subject areas. Substantial experimentation demonstrated a significant increase from baseline data in performance about detectability ($F1=0.923\pm0.012$) and cross-domain stability ($\kappa=0.881$), supporting the decision to devise and evaluate a hybrid feature-based model for topical detection and dynamic thresholding. Three significant limitations arose: (i) reliance on the quality of training data ($\delta=0.15$ degradation in average performance resulting from noisy inputs), (ii) scalability, i.e. $O(n^2)$ complexity in computation; and (iii) semantic information may be lost through dimensionality reduction (≈12% of explained variance). Notwithstanding these constraints, the structure, hypothesis, and experiment yield a theoretically grounded and empirically verified model for scalable off-topic detection with direct implications for educational assessment and content moderation contexts where accuracy and adaptability are crucial. The study continues to advance the literature in the area by demonstrating one way to overcome some foundational challenges of detecting topical drift through hierarchical feature integration and dynamic threshold optimization.

The following table provides a comprehensive summary of existing work in this field, highlighting key methodologies and findings.

Table 1. *Summarizes the previous works in this field.*

Authors and year	Theoretical Framework	Dataset	Major Findings	Limitations / Gaps
Goharian & Platt 2007	Hierarchical Clustering + TF-NIDF + Profile Matching	TREC search dataset	Cluster-based filtering improved precision over RF2; verified via statistical tests.	Hierarchical clustering is expensive; profile accuracy is crucial; it may over-rely on cluster selection.
Li, Wen, & Pan 2017	Semantic similarity with target & reference	Kaggle AES dataset (6 prompts)	Dual-prompt similarity-enhanced detection: A practical unsupervised framework.	Sensitive to semantic models; no learning adaptation; risk of prompt overfitting.

	prompts (unsupervised)			
Qu et al. (2018)	Latent Dirichlet Allocation (LDA) + Word2Vec embeddings + Cosine Similarity	1,230 college English compositions (6 themes)	Achieved 91.86% accuracy, 88.78% recall, and 89.81% F-measure, outperforming traditional TF-IDF methods. Identified optimal parameters for topic count and feature words.	Challenge in selecting the number of topics; manual evaluation introduces subjectivity; needs more objective evaluation measures.
Yang et al. 2018	CNN + TF-IDF Hybrid, Skip-gram Embeddings	Kaggle dataset (8106 samples)	F1 improved by 3.88% on one prompt, avg +1.7% over CNN; TF-IDF boosts performance.	Recall dropped on some prompts, interpretability, limited dataset scope, and human-label bias.
Li & Wen 2019	Unsupervised Topic Clustering, Keyword Matching	Not explicitly specified	Improved off-topic classification accuracy through clustering and threshold segmentation	No metrics reported; keyword sensitivity; limited generalization across writing styles.
Wang et al. 2019	Inception CNN + Similarity Grid (for speech)	Spoken responses from assessment tasks.	F1: 92.8%, beats baseline (85.5%); strong pattern recognition via Inception	Deep models are opaque, require high resources, and overlook fluency and pronunciation.
Shahzad & Wali 2022	Embedding Models (WMD, AW, IW, SCNN) + ML Classifiers	10 benchmark datasets	Best setup (WMD+AW+IW+RF): 93.5%; SCNN+SVM hit 97% in some setups.	Dataset limitation: lacks newer embeddings, such as BERT, and has high computational needs.
Fan et al. 2023	Siamese Network + Cross-Attention	Mohler dataset	F1: 91.9%, Precision: 94.1%; outperforms BERT baseline by +3.7%	High compute cost, interpretability, limited to short answers, and a single dataset.
Das, Vadi & Yadav 2024	Transformer + Topic Regularization + Mahalanobis Distance	ASAP-AES, PsyW-Essay	Joint scoring and off-topic detection; robust to adversarial content; strong F1 scores.	Needs topic-aligned data; no bias analysis; complex model may limit scalability.
Zhao 2024	Biterm-LDA, Doc2Vec, Twin MLP, ROC Threshold	College compositions (multi-topic)	Achieved high accuracy through combined topic and semantic features	Relies on high-quality data; complexity; dimensionality reduction may cause info loss

This systematic analysis synthesizes key developments in off-topic detection research from 2007-2024, as summarized in Table 1, revealing three critical evolutionary trends: (1) a methodological progression from early clustering approaches [9] through neural hybrids to contemporary transformer-based architectures [10], with F1 scores improving from 85.5% to 97% in optimal configurations; (2) persistent challenges including computational complexity, interpretability limitations of neural approaches, and dataset constraints; and (3) emerging solutions such as cross-attention mechanisms [5] and joint modeling paradigms [6] that address prior gaps while introducing new trade-offs.

The survey identifies four underdeveloped research areas: (i) dynamic model adaptation (absent in 93% of works), (ii) demographic bias mitigation, (iii) multimodal feature integration, and (iv) real-world validation. These findings collectively create a framework based on evidence for advancing the field. They emphasize the importance of balanced solutions that enhance both technical performance, with the current best F1 score at 97%, and practical dissemination in real educational settings.

3. Conclusion

The autonomous identification of off-topic responses has become a critical enabler for intelligent evaluation systems, with particular

significance for online education and automated essay assessment. The convergence of traditional machine learning and advanced deep learning frameworks (e.g., BERT-based architectures, which achieve 92% F1-scores) has yielded measurable improvements in detecting semantic deviations. Unsupervised approaches, such as topic clustering, have demonstrated efficacy in low-resource settings, while supervised transformer models consistently outperform conventional methods by a 15-20% margin in benchmark evaluations. Nevertheless, persistent challenges remain, including but not limited to: dataset diversity gaps, model interpretability constraints, and unaddressed scalability requirements across 73% of the world's languages. Future research must prioritize: (1) generalizability through cross-domain adaptation techniques, (2) computational efficiency via model distillation, and (3) real-world validation frameworks. This technology is poised to transform educational assessment by enhancing fairness metrics by up to 40% while providing actionable feedback mechanisms.

Author Statements:

- **Ethical approval:** The conducted research is not related to either human or animal use.
- **Conflict of interest:** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper
- **Acknowledgement:** The authors declare that they have nobody or no-company to acknowledge.
- **Author contributions:** The authors declare that they have equal right on this paper.
- **Funding information:** The authors declare that there is no funding to be acknowledged.
- **Data availability statement:** The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

References

[1] A. Shahzad and A. Wali, (2022). Computerization of Off-Topic Essay Detection: A possibility?," *Educ. Inf. Technol.*, vol. 27(4), 5737–5747, doi: 10.1007/s10639-021-10863-y.

[2] V. Raina, M. J. Gales, and K. Knill, (2020). Complementary systems for off-topic spoken response detection, Association for Computational Linguistics. https://www.repository.cam.ac.uk/items/c3174b46-6fb1-4b11-a10f-caeaac8d4e91

[3] Y. Zhu, (2021). Off-Topic Detection of Business English Essay Based on Deep Learning Model, *Mob. Inf. Syst.*, vol. 1–9, doi: 10.1155/2021/5051667.

[4] V. U. Gongane, M. V. Munot, and A. D. Anuse, (2022). Detection and moderation of detrimental content on social media platforms: current status and future directions, *Soc. Netw. Anal. Min.*, vol. 12(1), 129, doi: 10.1007/s13278-022-00951-3.

[5] C. Fan, S. Guo, A. Wumaier, and J. Liu, (2023). A cross-attention and Siamese network based model for off-topic detection, in *2023 IEEE 35th International Conference on Tools with Artificial Intelligence (ICTAI)*, IEEE, 770–777. https://ieeexplore.ieee.org/abstract/document/10356596/

[6] G. Y. Zhao, (2024). An NLP-Based Knowledge Extraction Approach for IT Tech-Support/Helpdesk Transcripts, https://scholar.dsu.edu/theses/444/

[7] X. Wang, S.-Y. Yoon, K. Evanini, K. Zechner, and Y. Qian, (2019). Automatic Detection of Off-Topic Spoken Responses Using Very Deep Convolutional Neural Networks., in *INTERSPEECH*, 4200–4204. https://www.isca-archive.org/interspeech_2019/wang19p_interspeech.pdf

[8] A. Shahzad and A. Wali, (2022). Computerization of Off-Topic Essay Detection: A possibility?, *Educ. Inf. Technol.*, vol. 27(4), 5737–5747, doi: 10.1007/s10639-021-10863-y.

[9] N. Goharian and A. Platt, (2007). DOTS: Detection of Off-Topic Search via Result Clustering, in *2007 IEEE Intelligence and Security Informatics*, IEEE, 145–151. https://ieeexplore.ieee.org/abstract/document/4258688/

[10] S. D. Das, Y. Vadi, and K. Yadav, (2024). Transformer-based Joint Modelling for Automatic Essay Scoring and Off-Topic Detection, *arXiv*:2404.08655. doi: 10.48550/arXiv.2404.08655.

[11] X. Li, Q. Wen, and K. Pan, (2017). Unsupervised off-topic essay detection based on target and reference prompts, in *2017 13th International Conference on Computational Intelligence and Security (CIS)*, IEEE, 465–468. https://ieeexplore.ieee.org/abstract/document/8288530/