A Large Scale Dataset for AI-generated Review Detection

Shifali Agrahari, Sujit Kumar and Sanasam Ranbir Singh

***Abstract*—Large language models (LLMs) have demonstrated substantial proficiency in tasks such as text generation, summarization, and producing coherent responses to a wide array of user queries. However, this capability has raised concerns about their potential misuse in areas like journalism, education, business, and academia, as well as the proliferation of deepfake text across digital platforms. The sudden rise of AI-generated reviews, where LLMs can create deceptive or biased content, further complicates the landscape of digital authenticity and trustworthiness. Consequently, detecting AI-generated reviews has become a significant research challenge to ensure the authenticity of reviews on e-commerce platforms, maintain consumer trust in online product evaluations, enable informed decision-making, and uphold fairness and integrity within the marketplace. In the literature, there is a scarcity of datasets for AI-generated review detection, which are restricted to domains like restaurants, Amazon, and Google Play Store and are very limited in sample size. This study proposes a large-scale dataset for AI-generated review detection, consisting of human-authored and AI-generated reviews from eight large language models (LLMs), including GPT, BART, T5, Gemini, Pegasus, LLaMA, Gemma, and Mistral, across five domains: E-commerce, Books, Movies, Hotels, and Tourism. To assess the quality and reliability of the proposed dataset, we evaluate the performance of a baseline model trained on this dataset against AI-generated content detection datasets from the literature, covering diverse domains and applications. The observations from our experimental results and assessments indicate that the proposed dataset is reliable for detecting AI-generated reviews across various domain applications, enhancing the model's ability to generalize to unseen domains and effectively detect AI-generated content across diverse applications.**

***Index Terms*—Article submission, IEEE, IEEEtran, journal, LaTeX, paper, template, typesetting.**

I. Introduction

RECENT advancements in natural language generation have enabled Large Language Models (LLMs) to produce text with exceptional linguistic quality. Large Language Models have exhibited remarkable capabilities across various domains, including logical reasoning, fluent language generation, and extensive factual knowledge [1], [2]. The emergence of large language models like LLaMA [3], BLOOM [4], and ChatGPT [5] has demonstrated the ability to achieve human-level performance across various domains [6]. These capabilities bring new challenges, such as a tendency to hallucinate new information [7], introduce biases [8], and violate privacy [9]. Text generated by LLMs that appears authentic to human readers is called deepfake text, also known as AI-generated or synthetic text [10], [11]. As noted in the studies [12], [13], modern advanced language models can generate deceptive content to influence public opinion, which includes the production of fake news articles [13], counterfeit application reviews [14], fraudulent app reviews [14], and misleading and manipulative social media posts [15].

Online product reviews are customer evaluations posted on websites or platforms that sell products or provide product information. These reviews cover product quality, functionality, durability, customer service, and delivery experience. Which also gives detailed insights into the strengths and weaknesses of a product, helping consumers make informed purchasing decisions. Products with higher ratings and more positive reviews are more likely to be purchased again than those with lower ratings and fewer reviews Spiegel Research Center. These reviews also serve as feedback for brands or companies, and over 80% of consumers rely on online reviews before purchasing B2B SaaS Reviews. Over the past decade, online reviews have remained a key component of e-commerce platforms, as consumers increasingly rely on reviews for purchasing decisions [16], [17]. Nearly 90% of online shoppers report that authentic reviews greatly influence their choices, often outweighing the importance of price importance of review. However, this positive reliance on reviews has been tainted by the generation of fake reviews which endorse the product, app, or place, boosting sales, improving product rankings, or damaging competitors. This unethical behaviour has led to a surge in counterfeit reviews, severely impacting the reputation and integrity of platform-based economies [18]. As reported in the study [19], there are two ways to create fake reviews. (i) *Fake review* human content writers are hired to write a large number of positive and favourable reviews for various products. These reviews appear authentic but are not genuine; they are written by product customers but are authored by hired content writers. (ii) *AI generated review* Large Language models used to generate product reviews. As reported in the study [19], AI-generated reviews can be persuasive and challenging to distinguish from human-written reviews due to advancing techniques and language models used to craft deceptive reviews. Recently, the prevalence of AI-generated reviews has sparked concerns about their potential to mislead consumers, known in the literature as fake reviews, review fraud, review opinion spam, or deceptive reviews [20]. Detecting AI-generated reviews is essential for the marketing and e-commerce sectors for the following reasons: (i) AI-generated reviews can erode trust in online reviews, which have a negative impact and significant market implications [21]. (ii) The widespread presence of AI-generated reviews could distort perceptions of product quality and customer satisfaction [18]. (iii) The prevalence of fake or AI-generated

Authors affiliation: Department of Computer Science and Engineering, Indian Institute of Technology Guwahati, India

reviews on online platforms can mislead potential buyers regarding a product's true performance and features, resulting in customer dissatisfaction and increased product returns. (iv) Detecting and combating AI-generated reviews is essential for maintaining fair competition among businesses [22]. Genuine reviews foster a level playing field where product reputation is based on honest customer feedback rather than fabricated endorsements. This encourages companies to prioritize product quality and customer satisfaction, promoting healthy market dynamics. Consequently, detecting AI-generated reviews has become a significant research challenge to ensure the authenticity of reviews on e-commerce platforms, maintain consumer trust in online product reviews, enable informed decision-making, and uphold fairness and integrity within the marketplace [18], [20]–[24].

Initial studies [18], [25]–[28] on detecting AI-generated reviews have primarily relied on feature-based methods, such as perplexity, counts of parts of speech (POS), and other feature indicators of word distribution along with sentiment of review text. However, features based methods struggle to identify subtle differences between human-written and AI-generated content and are less effective at detecting sophisticated AI-generated texts that closely mimic human writing styles [29]. Subsequent studies have employed sequential encoding [30]–[32] and transformer-based models [33], [34] for fake review detection. While many datasets have been curated for fake review detection [35]–[39], there are comparatively few datasets available specifically for detecting AI-generated reviews. In the literature, studies [40], [41] have curated datasets for AI-generated review detection in the restaurant and Google Play Store domains by combining authentic reviews with reviews generated using OpenAI's ChatGPT. Similarly, study [30] has also curated datasets for AI-generated fake review detection in the online product domain by combining authentic reviews with reviews generated using GPT-2 and Universal Language Model Fine-tuning (ULMFiT). However, these datasets are limited in number and lack diversity in review across different domains. Furthermore, the fake samples are generated using only GPT-2 and Universal Language Model Fine-tuning (ULMFiT). Given the availability of numerous large language models for text generation and summarization, the existing AI-generated review detection datasets are limited by their reliance on fake samples generated using only GPT-2 and Universal Language Model Fine-tuning (ULMFiT), making them inadequate for detecting review-generated by LLMs other than GPT and ULMFiT and also not effective in detecting review generated by recent advanced large language models. Datasets curated in the literature are limited in scope, primarily focusing on authentic and AI-generated reviews within specific domains, such as online product reviews from Amazon, Google Play Store reviews, and restaurant reviews. Consequently, AI-generated models trained on these datasets are often inadequate at detecting AI-generated reviews in other domains, such as travel, hotel, and book purchase reviews. To address the limitations of existing AI-generated review detection datasets in the literature, this study proposes a dataset that includes reviews from diverse domains, with fake reviews generated by numerous advanced large language models. We curated AI generated review detection datasets across diverse domains and applications, including book reviews, IMDB movie reviews, TripAdvisor reviews, hotel reviews, and women's clothing e-commerce reviews. Similarly, our proposed datasets also incorporate numerous and diverse large language models, including *Generative Pre-trained Transformer*(GPT-3) [42], *Bidirectional and Auto-Regressive Transformers*(BART) [43], *Text-To-Text Transfer Transformer* (T5) [44], Gemini [45], Pegasus [46], Llama [3], Gemma [47], and Mistral [48], to generate reviews across various domains. To provide a comprehensive understanding of AI-generated reviews and their detection, we examine the following two research questions through empirical studies and investigations.

- **RQ1** Is the effectiveness of AI-generated review detection dependent on the domain?. Can AI review detection models trained on reviews from one domain effectively detect AI-generated reviews in other domains?
- **RQ2** Is AI-generated review detection influenced by the specific LLM used to generate the reviews?. can a model trained on reviews generated by one LLM effectively detect reviews generated by other LLMs?
- **RQ3** Given the same source text and prompt as input, do all LLMs generate similar output, or how different are the texts produced by various LLMs over the same source and same prompts?

By investigating **RQ1**, we aim to study whether an AI-generated review detection model trained on datasets containing human-written reviews and fake reviews from specific domains can effectively detect fake reviews generated by the same or different LLMs across various domains. Simiarly, the prime motivation behind investigating **RQ2**, we aim to explore whether an AI-generated review detection model trained on datasets containing human-written reviews and fake reviews generated by specific LLMs can effectively detect fake reviews generated by other LLMs in the same or different domain. In addition, the motivation behind exploring **RQ3** is to examine how similar or different the reviews or texts generated by various LLMs are when provided with the same source text and prompt.

The remainder of this paper is organized as follows. Section II presents the works related to AI-generated review detection. In Section III, we present our dataset curation methods and setup. Section IV presents experiments setup followed by empirical studies and analysis to answer the research question, and Section VI concludes the paper with future directions.

II. Related Work

Online product reviews, a key form of electronic Word-of-Mouth (eWOM), significantly influence consumer purchase decisions [49]–[53]. In the United States, over 80% of consumers report consulting online reviews before making a purchase [54].

Detecting AI-generated and fake reviews is crucial for maintaining the integrity of online platforms. Researchers have explored various detection methods, including machine learning (ML), graph-based approaches, and LSTM models [10],

[55]–[58]. Recently, there has been a shift toward using Large Language Models (LLMs) for detection [28], [59]–[62]. Some literature treats deep fake text detection as a classification problem, solvable by Pre-trained Language Models (PLMs) [10], [11], [19].

Fake review detection is not a new issue, with various datasets available, each posing unique challenges. For example, [36] collected 42 fake and 40 true hotel reviews, which are insufficient for effective ML training. [38] compiled a dataset of 9,000 labeled reviews, and there is also the reputable Amazon Review Data (2018) [39]. extensive and reputable [39]. [19] created a dataset for classifying fake reviews, considering those generated by ChatGPT as fake, using prompts for dataset generation.

With advancements in LLMs, creating large volumes of AI-generated text has become easy, making detection increasingly difficult due to the variety of models and domains. Recent studies have focused on datasets for AI-generated text detection across various models like GPT-3 [42], BART [43], T5 [44], and Gemini [45], spanning domains such as Wikipedia, Reddit, and news. However, there remains a lack of review-specific domain datasets

Current research on AI-generated reviews often lacks diversity in datasets, primarily focusing on a narrow range of domains or relying on limited model outputs. Many existing studies use datasets generated by a single LLM or focus on specific application areas, which hampers the generalizability of their findings [19]. Furthermore, there is a gap in understanding how AI-generated reviews differ across multiple domains and models, which limits the ability to develop robust detection mechanisms.

This paper aims to address these gaps by proposing a comprehensive dataset that encompasses AI-generated reviews from four distinct domains, utilizing various LLMs. By doing so, we seek to investigate the effects of domain and model diversity on the effectiveness of AI review detection systems. Our objectives focus on creating a more representative dataset and evaluating the performance of detection algorithms across different contexts and generated content.

To tackle these issues, we will create a dataset comprising human-written reviews and their corresponding AI-generated counterparts across multiple domains, including books, movies, hotels, and clothing. Our methodology involves employing multiple state-of-the-art LLMs to ensure a wide variety of generated reviews. Additionally, we will explore detection models that can generalize across different domains and AI generation techniques, leveraging advanced machine learning approaches, including fine-tuning pre-trained transformers like RoBERTa [63]. In conclusion, our study contributes to the literature by providing a rich and diverse dataset for AI-generated reviews, facilitating research on detection techniques. We aim to demonstrate that a well-rounded dataset enhances the effectiveness of AI review detection across different domains and models. By addressing the limitations of existing studies, we hope to pave the way for more accurate and reliable detection systems that can adapt to the rapidly evolving landscape of AI-generated content.

III. Methodology

As discussed in Section I, the primary objective of this study is to propose datasets for AI-generated review detection across diverse domains. The proposed dataset is intended to include samples from various domains, with fake reviews generated by various LLMS to facilitate the detection of AI-generated reviews effectively.

A. Generation of AI-Generated Review Datasets

Given a human-written true review, $\mathcal{R}$, the objective of the AI review generation process is to produce a corresponding fake review, $\mathcal{R}_a$, utilizing large language models (LLMs) guided by prompt-based instructions. Considering the effectiveness of large language models in text generation and summarization, we utilize the *Generative Pre-trained Transformer* (GPT-3) [42], *Bidirectional and Auto-Regressive Transformers* (BART) [43], *Text-To-Text Transfer Transformer* (T5) [44], Gemini [45], *Pre-training with Extracted Gap-sentences for Abstractive Summarization* (PEGASUS) [46], *Large Language Model Meta AI* (Llama) [3], Gemma [47], and Mistral [48] models to generate fake reviews. To curate a comprehensive AI-generated fake review dataset suitable for training a model to detect AI-generated reviews across diverse domains, we include datasets from various sources. These domains encompass a wide range of reviews which include Amazone Book Reviews[1][2] [64]–[66], Internet Movie Database Movie Reviews[3][4] [**?**], Trip Advisor Hotel Reviews[5][6] [67], Tourist Reviews[7], and the Women's Clothing E-Commerce dataset[8][9] [68], [69]. We also provide prompt instructions to the models, offering specific guidance for generating fake reviews. Though we tried several prompt instructions to guide the model to generate the fake review, the prompt instruction for which we received the best results is as follows: *Generate a Fake Review using True Review* , which directs the model to create a fabricated review based on a human written true review. Figure 1 presents a working diagram outlining the process of generating fake reviews.

B. Method for Fake Review Classification

Given a review $\mathcal{R}$, the task is to classify it as $Y \in T, F$, where T denotes that the review $\mathcal{R}$ is true and human-authored, and F signifies that the review $\mathcal{R}$ is fake or AI-generated. The objective is to learn an AI-generated review detection function $f : (\mathcal{R}) \rightarrow Y$. Recognizing the superior performance of the transformer-based model RoBERTa [63] in tasks such as text similarity, natural language inference (NLI), question answering, and recognizing textual entailment (RTE), which capture deep semantic relationships, complex negations,

[1] Amazon Books Reviews Dataset

[2] Kaggle Amazon Books Reviews Repository

[3] Kaggle IMDb Movie Reviews Repository

[4] IMDb Movie Reviews Repository

[5] Kaggle Trip Advisor Hotel Reviews Repository

[6] Trip Advisor Hotel Reviews Source Repository

[7] Kaggle Tourist Review Repository

[8] Women's Clothing E-Commerce Dataset

[9] Women's Clothing E-Commerce Dataset

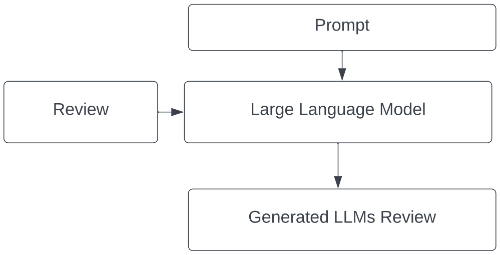


Fig. 1. The working diagram outlines the process of generating fake reviews. A human-written review (**Review**) and a set of guiding instructions (**Prompts**) are provided as inputs to a (**Large Language Model**) LLMs. The LLMs use these inputs to produce a corresponding fake review, emulating the characteristics of an artificially generated review.

and propositional content between texts, this study fine-tunes a pre-trained RoBERTa[10] for review detection. Another key motivation for fine-tuning pre-trained RoBERTa models for AI-generated review detection is that RoBERTa is trained on large and diverse datasets, including the *Colossal Clean Crawled Corpus*, *News Corpus*, *BooksCorpus*, and *English Wikipedia*. This extensive training enhances the model's deep contextual understanding of text. Given that these models are pre-trained on corpora covering a wide range of topics and domains, RoBERTa is well-suited for detecting AI-generated reviews across various subjects and domains.

Dataset	**Train**	**Dev**	**Test**	**Total**
E-Commerce	32,880	4,697	9,394	46,972
Book	4,200,000	600,000	1,200,000	6,000,000
Hotel	28,687	4,098	8,196	40,982
Movie	70,000	10,000	20,000	100,000
Tourist	10,179	1,454	2,908	14,542

TABLE I
DATASET STATISTICS

TABLE II
LEXICAL FEATURE ANALYSIS OF CREATED DATASET

Features	Human	Bert	Gemini	Pegasus	T5	GPT2	Llama	Gemma	Mistral
#sent	6.89	5.02	5.01	3.89	3.01	4.25	7.45	6.00	6.23
#quota	3.44	1.09	2.09	0.03	0.01	0.91	4.34	3.01	2.22
#uni wor	5	4.1	3.75	0.34	0.02	2.24	5.09	2.12	4.12
#spe cha	2.19	1.92	0.34	0.02	0	1.23	2.34	1.10	1.92
#pron	5.23	3.21	2.43	1.33	2.13	3.23	5.34	2.34	4.56

IV. EXPERIMENTAL RESULTS AND DISCUSSIONS

A. Dataset

As discussed in subsection III-A, we curated a fake review dataset by collecting human-authored reviews from five different domains: book reviews, movie reviews, TripAdvisor reviews, hotel reviews, and women's clothing e-commerce reviews. Fake reviews (AI-generated) were produced using eight different LLMs, including (GPT-3) [42], (BART) [43], (T5) [44], Gemini [45], Pegasus [46], Llama [3], Gemma [47], and Mistral [48], across all five domains. Consequently, we propose a comprehensive collection of AI-generated review datasets, resulting in forty datasets across five different domains, each created using eight distinct models. Since a fake review is generated by LLMs for every human-authored review, each proposed dataset contains an equal number of LLM-generated (AI-generated) reviews and human-written reviews. Therefore, our proposed datasets are balanced in terms of the number of reviews generated by humans and those generated by LLMs within each dataset. We curated a large-scale AI-generated review dataset named (**Large AI-Generated Review Dataset**) by merging forty individual datasets, which consist of human-authored reviews and LLM-generated reviews spanning five domains and generated by eight different LLMs. Table I presents the characteristics of the dataset across five different domains. While eight different LLMs generate fake reviews for each domain, the number of samples remains consistent across training, test, and development sets across the forty datasets. Additionally, each dataset maintains an equal distribution of human-authored and AI-generated text. Similarly,

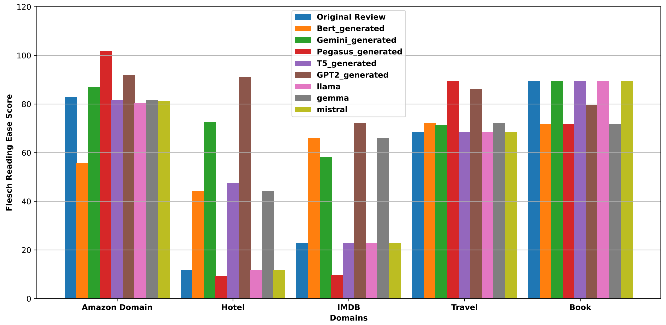


Fig. 2. Flesch Reading Ease Scores Across Different Domains and Generations

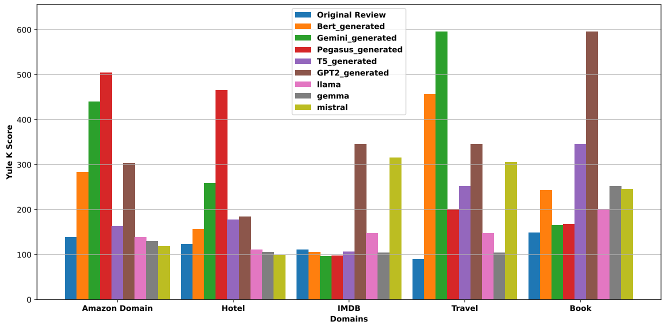
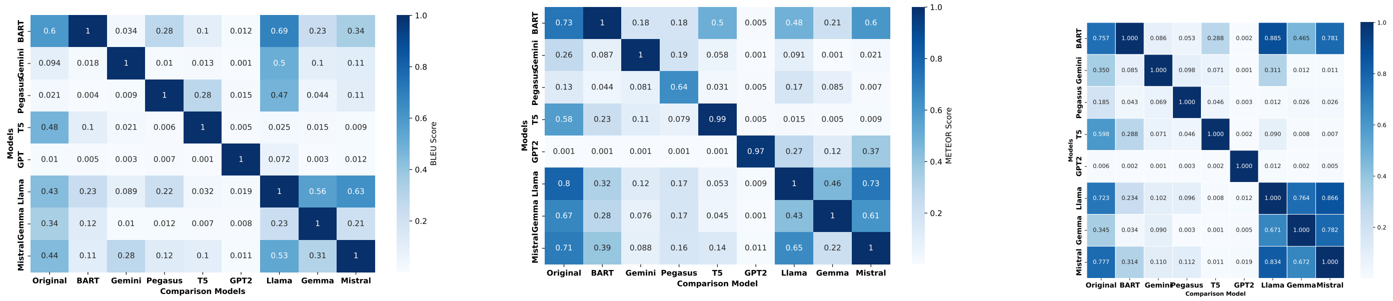


Fig. 3. Yule K Scores Across Different Domains and Generations

To evaluate the performance of the model trained on our proposed dataset, we additionally consider seven datasets from the literature, Restaurant Review [70], Wikipedia [71], Medicine [71], Essay Kaggle, Amazon Review [19], Applied Statistics [72], and M4 [73]. A description of these datasets is provided in Table V.

[10] RoBERTa Large Model



(a) BLUE Score (b) Meteor Score (c) Rouge Score

Fig. 4. **Make it GPT-3 and its PEGASUS not pegasus make these changes throuhgt paper**

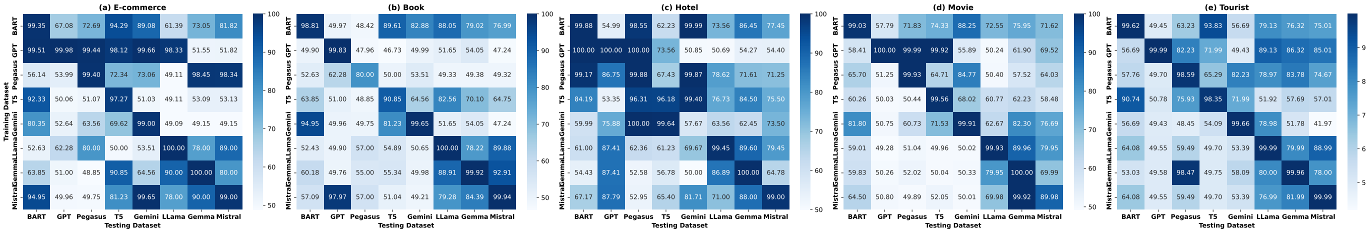


Fig. 5. Accuracy of cross-model classification using Roberta model: train and test on different across five domain, over machine-generated text from 8 LLMs models. **Split these Figure (each heatmap seperatelt) into 5(a) , 5(b)...similar to setup of Figure 4**

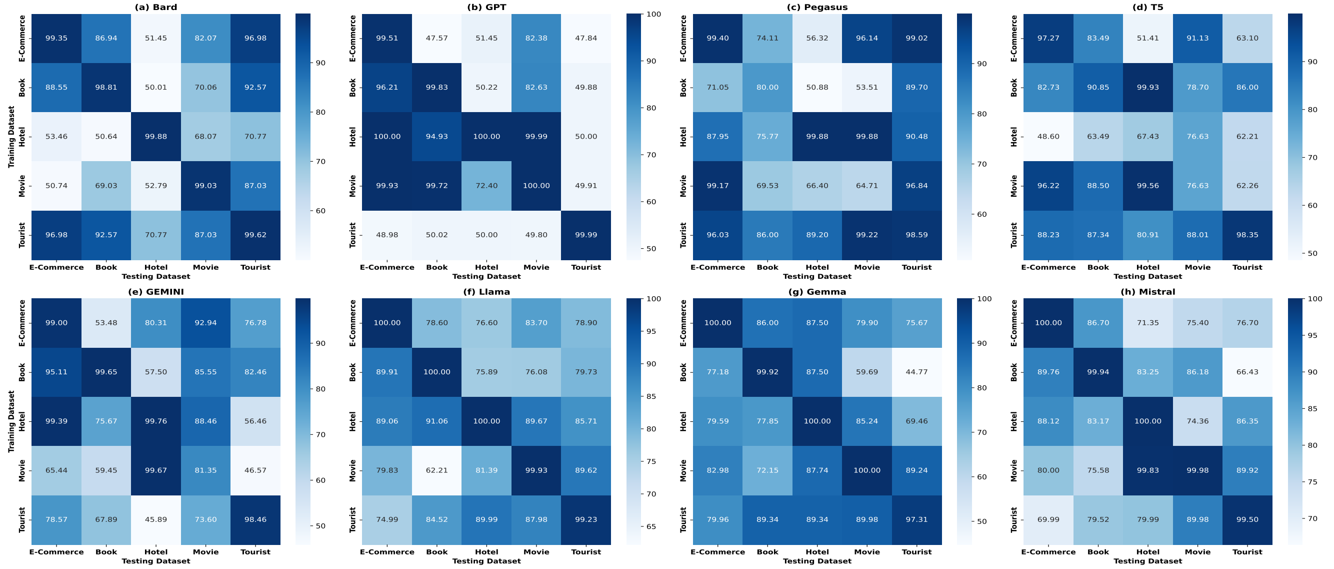


Fig. 6. Accuracy of cross-domain experiments classification using Roberta model: given generations from 8 different LLMs Model, train on a single domain and test across all different domains within the same Model. **Split these Figure (each heatmap seperatelt) into 5(a) , 5(b)...similar to setup of Figure 4**

B. Experimental Setups

This study considers accuracy (Acc.) and F-measure (F.) as performance metrics to evaluate classification models. Additionally, we calculate certain linguistic features of the dataset, as mentioned in the study by [74], to understand lexical diversity across different models alongside the original human reviews. Table II presents key metrics, including the average number of sentences per review, quotations, and unique words per review. Additional indicators, such as special characters and personal pronouns per review, help identify conversational tones. Readability scores and measures of vocabulary diversity assess the reading level of a text, often based on the education level required for easy comprehension. Human texts consider factors like cultural relevance, context, and style, whereas LLM models focus on textbook definitions, which can lead to differences even when conveying the same meaning. Fig 2 shows Flesch Reading Ease scores, revealing distinct patterns in readability. Yule's K (Fig 3) provides further insight into vocabulary richness, with higher values indicating more complexity. Both measures are analyzed

	BART			**Gemini**			**Pegasus**			T5			**GPT2**			Llama			**Gemma**			**Mistral**		
Models	**P**	**R**	**F1**	**P**		**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
BART	1	1	1	0.851	0.790	0.819	0.844	0.766	0.803	0.872	0.810	0.839	0.700	0.753	0.725	0.720	0.780	0.750	0.830	0.800	0.815	0.670	0.700	0.685
Gemini	0.790	0.851	0.819	1	1	1	0.815	0.824	0.819	0.825	0.837	0.830	0.703	0.809	0.753	0.780	0.800	0.790	0.670	0.640	0.655	0.600	0.660	0.630
Pegasus	0.766	0.844	0.803	0.844	0.766	0.803	1	1	1	0.802	0.823	0.812	0.694	0.790	0.739	0.800	0.810	0.805	0.710	0.760	0.730	0.690	0.710	0.700
T5	0.810	0.872	0.839	0.810	0.872	0.839	0.837	0.825	0.830	1	1	1	0.695	0.790	0.739	0.710	0.730	0.720	0.600	0.640	0.620	0.750	0.780	0.765
GPT2	0.753	0.700	0.725	0.700	0.753	0.725	0.809	0.703	0.753	0.790	0.694	0.739	1	1	1	0.670	0.650	0.660	0.700	0.730	0.715	0.620	0.650	0.635
Llama	0.810	0.850	0.830	0.720	0.770	0.740	0.670	0.740	0.700	0.680	0.720	0.700	0.710	0.760	0.735	1	1	1	0.823	0.802	0.813	0.803	0.824	0.815
Gemma	0.720	0.770	0.740	0.700	0.740	0.720	0.680	0.710	0.695	0.640	0.670	0.655	0.670	0.700	0.685	0.730	0.710	0.720	1	1	1	0.819	0.851	0.790
Mistral	0.670	0.740	0.700	0.680	0.720	0.700	0.710	0.760	0.735	0.700	0.740	0.720	0.680	0.710	0.695	0.790	0.695	0.739	0.851	0.790	0.819	1	1	1

TABLE III

CAPTION AND TABLE ORIENTAL IS NOT PROPER

across original reviews and text generated by various models for five domains: Amazon, Hotel, IMDB, Travel, and Book. Other measures the Bilingual Evaluation Understudy (BLEU) Score [75] and the Metric for Evaluation of Translation with Explicit Ordering (METEOR) Score [76] are used to analyze the relationship between human-written reviews and those generated by LLMs, as well as to study the differences and similarities between reviews generated by different LLMs. Table IV presents the details of the hyperparameters used to produce the results in this paper. Although we experimented with various hyperparameter settings, Table IV outlines the specific hyperparameters for which our classification model achieved optimal performance. **Availability of Codes and Proposed Datasets:** We will provide a web link to the source code of baseline models and proposed models in the final version of our submission upon acceptance of the paper.

TABLE IV

PRESENTS THE DETAILS OF EXPERIMENTAL HYPERPARAMETERS

Hyperparameters	**Value**
Batch Size	16
Learning Rate	0.001
Maximum # Words in Review	500
Epochs	20
Dropout Rate	0.5
Optimizer	Adam
Weight Decay	0.0001

C. Results and Discussion

Figure 5 depicts the performance of the RoBERTa model, trained on reviews generated by a specific LLM and separately evaluated on reviews produced by various other LLMs (including GPT, BART, T5, Gemini, Pegasus, LLaMA, Gemma, and Mistral). The evaluation covers human-authored and LLMs-generated reviews from various domains such as E-commerce, Books, Movies, Hotels, and Tourism. Accordingly, a heatmap is provided for each domain: Figures 5 (a), (b), (c), (d), and (e) present the evaluation of the RoBERTa model on human-authored and LLMs generated review from the E-commerce Book, Hotel, Movie, and Tourism domains, respectively. In Figure 5, the Roberta model is trained on human-authored reviews generated by a specific LLM and subsequently evaluated on human-authored reviews and reviews generated by different LLMs within the same domain. Although the training and test samples are generated by different LLMs, both training and testing set reviews belong to the same domain. Each cell in the i^{th} row and j^{th} column of Figure 5 represents the scenario where the RoBERTa model is trained on reviews from the LLM in the i^{th} row and tested on reviews generated by a different LLM within the same domain in the j^{th} column. **RQ2** Is AI-generated review detection influenced by the specific LLM used to generate the reviews?. can a model trained on reviews generated by one LLM effectively detect reviews generated by other LLMs?

The diagonal of the heatmaps in Figure 5 (a), (b), (c), (d), and (e) demonstrates that the RoBERTa model performs optimally when the training and test datasets comprise reviews generated by the same LLMs. While examining the non-diagonal elements of the heatmaps in Figure 5 (a), (b), (c), (d), and (e), it becomes apparent that the model's performance is either inconsistent or diminished when the reviews in the training and test sets are generated by different LLMs. From such observation of the performance of the RoBERTa model, we can conclude that the model trained over reviews generated by specific LLMs may not be effective in detecting reviews generated by other LLMs. The observations from Figure 5 (a), (b), (c), (d), and (e) regarding the performance of the RoBERTa model indicate that the effectiveness of AI-generated review detection is strongly generative models (LLMs) dependent. A model trained on reviews generated by a specific generative model (LLM) experiences a significant drop in performance when detecting reviews produced by different LLMs, even within the same domain. This suggests that AI review detection models face difficulties in generalizing when detecting reviews generated by different LLMs. Consequently, we conclude that models trained on reviews generated by specific LLMs may not effectively detect reviews generated by others. Therefore, ensuring consistency between the LLMs used to generate training and test samples is crucial for achieving optimal performance in AI-generated review detection. **RQ3** *Given the same source text and prompt as input, do all LLMs generate similar output, or how different are the texts produced by various LLMs over the same source and same prompts?* As observed from the non-diagonal elements of the heatmaps in Figure 5 (a), (b), (c), (d), and (e), the performance of the RoBERTa model is either inconsistent or diminished when the reviews in the training and test sets are generated by different LLMs. To further explore the reasons behind such inconsistency or decline in performance, we analyze the relationship between the text generated by different LLMs. As discussed in Section III-A and Subsection IV-A, we use a similar prompt to generate fake reviews from all LLMs for each human-authored review across all domains. Accordingly, we investigate the following aspects: (i) Given the same true review authored by a human and an identical prompt instruction as input to different LLMs, we examine how similar or dissimilar the text generated by the LLMs is in terms of contextual similarity and lexical overlap. we analyze linguistic features of the dataset, such as lexical diversity, readability scores, and vocabulary diversity, as discussed in [74]. Table II compares the lexical diversity and conversational tone of reviews generated by various LLMs to original human-written reviews. Both the original reviews and Llama-generated text display higher lexical diversity, with more sentences, quotations, and unique words, reflecting richer language. In contrast, Pegasus and T5 generate shorter, simpler reviews with fewer sentences and unique words, resulting in more straightforward language. Additionally, the frequent use of special characters and personal pronouns in the original reviews and Llama's outputs suggests a more conversational tone. Overall, Llama and the original reviews exhibit more complexity and naturalness, while Pegasus and T5 favor simplicity and directness.Figure 2 highlights significant readability differences across LLMs using Flesch Reading Ease scores. Pegasus shows the widest range, indicating potential difficulty in certain domains, whereas GPT-2 consistently generates highly readable text, and Gemini performs similarly. Bert and T5 display moderate readability, while Llama, Gemma, and Mistral closely resemble original reviews. This demonstrates the importance of model choice, with GPT-2 and Gemini excelling in accessibility for readable content.Figure 3 shows Yule's K scores for vocabulary diversity. Original reviews have moderate to high diversity, while model-generated texts vary. Pegasus stands out with exceptionally high diversity, followed by Gemini. T5 and GPT-2 strike a balance, with GPT-2 adapting well to domain-specific content. Bert shows variable performance, often surpassing the originals. Llama, Gemma, and Mistral align closely with original reviews, indicating more natural language use. Overall, Pegasus and Gemini lead in vocabulary diversity, while T5, GPT-2, and Bert show more variation.In answer to the investigation, when provided with the same human-authored review and prompt, LLMs show varying degrees of contextual similarity and lexical overlap. Llama and the original reviews closely mirror each other in terms of complexity and diversity, while Pegasus and T5 tend toward simpler, less diverse language styles, underscoring the importance of model selection based on the desired text characteristics.

(ii) Given a true review authored by a human and the same prompt instruction, we analyze how similar or dissimilar the text generated by the LLMs is to the original true review. This study utilizes ROUGE [77], METEOR [76], and BLEU [75] scores to assess the similarity between reviews generated by LLMs in terms of lexical overlap. Additionally, BERTScore [78] is employed to evaluate the contextual similarity between the reviews generated by different LLMs. Table III presents the similarity between reviews generated by LLMs given the same source review and prompt as input. Table III shows that the BERT scores between reviews generated by different LLMs range from seventy to eighty-five percent. This suggests that reviews generated by different LLMs, given the same prompt and source review, exhibit high contextual similarity to one another. Similarly, Figure 4 (a), (b), and (c) present the BLEU, METEOR, and ROUGE scores, respectively, between reviews generated by different LLMs when provided with the same prompt and source review. It is evident from Figure 4 that these scores are either low or average, indicating that the reviews generated by different LLMs exhibit less similarity in terms of token overlap, even when given the same source review and prompt. Intuitively, it is also possible because there is less vocabulary overlap between text generated by different LLMs. Subsequently, we also examine the relationship between the original true review and the fake reviews generated by the LLMs corresponding to that original review. As shown in Figure 4 (a), (b), (c), the similarity between the source review and the reviews generated by GPT-3, Gemini, and PEGASUS is relatively low in terms of BLEU, ROUGE, and METEOR scores. This indicates that the reviews generated by the GPT-3, Gemini, and PEGASUS models are not identical to their corresponding original review inputs to the LLMs, particularly in terms of lexical overlap. In contrast, reviews generated by other LLMs show moderate to average similarity to their corresponding input review based on these metrics. This moderate similarity may be attributed to the fact that while the LLMs are instructed to generate fake reviews, they may retain sentiment and polarity words (e.g., *Excellent*, *Good*, *Best*) that indicate a positive or negative sentiment, which may or may not be altered in the fake review. Based on the above observations, we can conclude that reviews generated by different language models (LLMs) using the same source review and prompts are contextually similar. However, there is minimal overlap in words used between reviews generated by different LLMs. The minimal lexical overlap and unique vocabulary in review generated by different LLMs contribute to the diminishing and declining performance of the Roberta model when the training and test datasets are composed of reviews generated by different LLMs.

RQ1 *Is the effectiveness of AI-generated review detection dependent on the domain?. Can AI review detection models trained on reviews from one domain effectively detect AI-generated reviews in other domains?* Figure 6 shows the performance of the RoBERTa model, trained on human-authored and LLM-generated reviews from a specific domain, and tested on human-authored and LLM-generated reviews from a different domain. The evaluation encompasses reviews from various domains. To present these evaluations, a heatmap is provided for each domain: Figures 6 (a), (b), (c), (d), (e), (f), (g), and (h) display the performance of the RoBERTa model on human-authored and LLM-generated reviews across different domains, produced by the same LLMs—BART, GPT-3, PEGASUS, T5, GEMINI, LLaMA, Gemma, and Mistral, respectively. In Figure 6, the RoBERTa model is trained on human-authored and LLM-generated reviews from one domain, and subsequently evaluated on human-authored and LLM-generated reviews from a different domain, but generated by the same LLMs. Although the training and test samples are generated by the same LLMs, they originate from different domains. Each cell in the i^{th} row and j^{th} column of Figure 6 represents a scenario where the RoBERTa model is trained on reviews from the domain in the i^{th} row and tested on reviews generated from a different domain in the j^{th} column. The diagonals in Figures 6 (a), (b), (c), (d), (e), (f), (g), and (h) indicate that the RoBERTa model achieves the best performance when both the training and test datasets are sourced from reviews within the same domain. Conversely, when reviewing the off-diagonal of heatmaps in Figures 6 (a), (b), (c), (d), (e), (f), (g), and (h), it becomes evident that the model's effectiveness is either inconsistent or decreases when the training and test sets, although produced by the same LLMs, pertain to different domains. Additionally, the following insights can be drawn from Figures 6. (i) From Figure 6 (a), it is evident that the RoBERTa model has difficulty detecting BART-generated reviews from the hotel domain when trained on BART reviews from the clothing, book, travel, and movie domains. Additionally, it fails to recognize BART-generated reviews from the book domain when trained on hotel reviews and struggles with identifying clothing domain reviews when trained on BART reviews from the movie and hotel domains. (ii) it is apparent from Figures 6 (b) that the RoBERTa model fine-tuned on GPT-3 generated reviews for the travel domain is unable to accurately detect GPT-3 generated reviews from other domains. Likewise, when trained on reviews from different domains, the model struggles to identify those from the travel domain. Furthermore, a model trained on GPT-generated reviews for the clothing domain fails to recognize reviews in both the book and hotel domains. A similar pattern occurs when trained on book domain reviews, resulting in an inability to detect hotel domain reviews. When the RoBERTa model is trained on PEGASUS-generated reviews from the book domain in Figures 6 (c), it cannot identify reviews from the hotel and movie domains. Similarly, training on Pegasus reviews from the movie domain results in the model being ineffective at identifying reviews across the book, hotel, and movie domains. The RoBERTa model faces challenges in identifying Gemini-generated reviews from the book and travel domains when trained on reviews from other domains, and it also struggles to detect hotel domain reviews when trained on Gemini-generated reviews from the book domain. Similarly, the model exhibits poor performance in detecting LLaMA-generated reviews from the book and travel domains when trained on reviews from the clothing and travel domains, respectively. Furthermore, it is unable to detect Gemini-generated reviews from the travel domain when trained on book domain reviews, and its ability to identify Gemini-generated reviews from the movie and travel domains decreases when trained on clothing domain reviews. Lastly, the model fails to recognize Mistral-generated book domain reviews when trained on movie domain data and also struggles to detect movie domain reviews when trained on Mistral-generated reviews from the clothing and hotel domains. Considering the above observations from Figure 6 regarding the performance of the RoBERTa model when human-authored and AI-generated reviews in the training and test dataset are from different domains, we conclude that the performance of models in detecting AI-generated reviews may decline or diminish when the training-and-test datasets contain reviews from differing domains.

TABLE V
REAL ONLINE TEST DATASET STATISTICS

Test Dataset	Test	AI Label	Human Label
Restaurant Review [70]	119	55	55
Wikipedia [71]	1,684	842	842
Medicine [71]	2,585	1,337	1,248
Essay Kaggle	29,145	11,637	17,508
Amazon Review [19]	40,432	20,216	20,216
Applied Statistics [72]	4,178	2,137	2,041
M4 [73]	261,758	163,430	98,328

V. EVALUATION OF PROPOSED DATASET

To evaluate the quality and reliability of the proposed dataset, we examine the performance of the RoBERTa model trained using the proposed dataset over AI-generated content detection datasets from the existing literature. We consider several datasets from the literature containing human-written content as true samples, as well as AI-generated content produced by various LLMs across diverse domains and tasks. Table IV-C illustrates the performance of the RoBERTa model when trained on the different setups of the proposed dataset and evaluated on AI-generated text detection datasets across various domains -Restaurant Review [70], Wikipedia [71], Medicine [71], Essay Kaggle, Amazon Review [19], Applied Statistics [72], and M4 [73]. Table IV-C presents the performance of the RoBERTa model trained under the following setups: (i) The model is trained on a mix of human-authored reviews, and AI-generated reviews generated by specific LLMs across five domains, including E-commerce, Books, Movies, Hotels, and Tourism. For instance, ”RoBERTa (BART)” denotes a RoBERTa model trained on a dataset containing human-authored reviews and fake reviews explicitly generated by the BART model only across the five domains. Similarly, ”RoBERTa (-)” indicates a model trained on human-authored reviews and reviews generated by a particular LLM. Here, can be replaced by any of the LLMs, including (GPT, BART, T5, Gemini, Pegasus, LLaMA, Gemma, or Mistral). (ii) Reviews generated by all eight LLMs across the five domains: RoBERTa (LLMs) in Table IV-C refers to a model trained on a dataset consisting of both human-authored and AI-generated reviews from all eight LLMs including (GPT, BART, T5, Gemini, Pegasus, LLaMA, Gemma, or Mistral) across five domains including E-commerce, Books, Movies, Hotels, and Tourism. From Table IV-C, it is evident that the RoBERTa models trained on the proposed dataset effectively detect AI-generated reviews across a range of domains and applications, even when different LLMs generate these reviews. Additionally, by analyzing the model's performance in Table IV-C following observations can be made. (i) The RoBERTa model demonstrates superior performance on domain and application-specific datasets (such as restaurant, Amazon reviews, Wikipedia, Medicine, Applied Statistics, and essays) when trained on reviews generated by specific LLMs (e.g., T5 for restaurants, BART for Amazon reviews, Applied Statistics, and essays, and Gemini for medicine) across five domain including E-commerce, Books, Movies, Hotels, and Tourism. (ii) Despite the absence of restaurant and Amazon reviews in the training dataset, the model effectively detects AI-generated reviews from the restaurant and Amazon domains. Consequently, we can conclude that the proposed training dataset enables the model to generalize well to un-

TABLE VI
FEW ENTRY ARE IN DECIMAL AND FEW ENTRY ARE IN PERCENTAGE. CONVERT EVERYTHING INTO DECIMAL LIKE LAST ROW

All Model	Restaurant Review		Wikipedia		Medicine		Essay		Amazon Review		Applied Statistics		M4	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
RoBERTa (BART)	60.37	0.60	**74.81**	0.90	**79.02**	**0.82**	68.36	**0.95**	**80.65**	**0.85**	**80.56**	**0.95**	**85.11**	**0.97**
RoBERTa (Gemini)	69.54	0.50	70.83	**0.91**	68.26	0.48	80.03	0.74	70.58	0.74	70.92	0.87	59.62	0.96
RoBERTa (GPT-3)	70.46	0.61	70.00	0.60	68.30	0.58	80.08	0.74	81.23	0.84	68.79	0.74	58.17	0.45
RoBERTa (PEGASUS)	69.54	0.60	69.41	0.60	68.33	0.60	79.69	0.80	54.56	0.74	67.61	0.64	60.07	0.51
RoBERTa (T5)	**70.46**	**0.80**	70.00	0.60	68.33	0.74	80.07	0.74	65.60	0.48	68.79	0.58	58.16	0.45
RoBERTa (Llama)	70.46	0.61	70.06	0.60	69.46	0.61	80.33	0.75	75.67	0.61	68.67	0.58	57.70	0.45
RoBERTa (Gemma)	70.46	0.61	70.59	0.62	69.57	0.61	**80.48**	0.75	73.56	0.56	69.17	0.61	58.28	0.46
RoBERTa (Mistral)	70.45	0.49	70.00	0.61	68.22	0.57	80.00	0.74	80.63	0.72	68.79	0.61	58.16	0.74
RoBERTa (LLMs)	**0.803**	**0.8**	**0.848**	**0.9**	**0.89**	**0.82**	**0.883**	**0.95**	**0.886**	**0.85**	**0.885**	**0.95**	**0.891**	**0.97**

seen domains. (iii) The model trained on reviews generated by BART and Gemini performs exceptionally well on the Wikipedia and Medicine datasets. This suggests that the model trained on our proposed dataset is also effective in detecting AI-generated answers to questions and distinguishing between AI-generated answers and human-authored answers. Such observations regarding the model's performance when trained on our proposed dataset are also effective in detecting AI-generated content in other applications. (iii) When trained on BART-generated reviews, the RoBERTa (BART) model significantly improves its capability to detect whether answers to questions from the Applied Statistics dataset are generated by humans or LLMs. (iv) The RoBERTa model, when trained on review generated by BART or Gemini, effectively differentiates between human-written and AI-generated essays and effectively detects AI-generated essays. Furthermore, from such observation, we can conclude that training the RoBERTa model on a dataset containing human-authored and specific LLM-generated reviews across various domains enhances its ability to detect AI-generated content effectively across different domains, applications and types of questions, even when these domains were not part of the original training data. Next, consider the RoBERTa (LLMs) model—where the RoBERTa model is trained on both human-authored reviews and AI-generated reviews from eight different LLMs (including GPT, BART, T5, Gemini, Pegasus, LLaMA, Gemma, and Mistral) across five domains (E-commerce, Books, Movies, Hotels, and Tourism)—Table IV-C shows that the performance of RoBERTa (LLMs) is significantly high and highly proficient in detecting AI-generated reviews within the Amazon and restaurant review dataset, as well as in identifying AI-generated content in question-answering tasks related to Medicine, Wikipedia, Applied Mathematics, and M4. The superior performance of RoBERTa(LLMs) is not surprising because here RoBERTa is trained on a dataset which contains reviews authored by humans and reviews generated by eight LMMs across five domains, whereas in the case RoBERTa(-), the RoBERTa model is human authored and review generated by a specific LLMs (One of the LLMs among GPT, BART, T5, Gemini, Pegasus, LLaMA, Gemma, and Mistral). From such observations, we conclude that our proposed data is reliable for detecting AI-generated reviews in different domain applications.

VI. CONCLUSION

In conclusion, this study highlights the challenges in detecting AI-generated reviews, particularly the significant performance decline observed in cross-model and cross-domain scenarios. While the RoBERTa model performs effectively when both training and test datasets originate from the same language model (LLM), its detection capability diminishes markedly when exposed to differing LLMs or domains. Despite contextual similarities in LLM outputs, the low lexical overlap contributes to inconsistent detection outcomes. These findings underscore the necessity for consistent training datasets to enhance detection efficacy, suggesting that the proposed dataset can help RoBERTa generalize better across unseen domains, thereby improving the reliability of AI-generated content detection.

For future work, we plan to investigate the integration of additional features, such as sentiment analysis and stylistic markers, to bolster detection accuracy. Moreover, exploring Advance methods that Few Sort Finetuning, Zero Shot method and LLMs itself as a detector could further enhance performance across various LLMs and domains, providing a more robust solution to the challenges of detecting AI-generated content in cross-model and cross-domain contexts.

ACKNOWLEDGMENTS

REFERENCES

[1] Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[2] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

[3] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[4] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. Bloom: A 176b-parameter open-access multilingual language model. 2023.

[5] Konstantinos I Roumeliotis and Nikolaos D Tselikas. Chatgpt and open-ai models: A preliminary review. *Future Internet*, 15(6):192, 2023.

[6] Tim Jansen, Yangling Tong, Victoria Zevallos, and Pedro Ortiz Suarez. Perplexed by quality: A perplexity-based method for adult and harmful content detection in multilingual heterogeneous web data. *arXiv preprint arXiv:2212.10440*, 2022.

[7] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*, 2023.

[8] Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. Towards understanding and mitigating social biases in language models. In *International Conference on Machine Learning*, pages 6565–6576. PMLR, 2021.

[9] Hannah Brown, Katherine Lee, Fatemehsadat Mireshghallah, Reza Shokri, and Florian Tramèr. What does it mean for a language model to preserve privacy? In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, pages 2280–2292, 2022.

[10] Yisroel Mirsky and Wenke Lee. The creation and detection of deepfakes: A survey. *ACM computing surveys (CSUR)*, 54(1):1–41, 2021.

[11] Jiameng Pu, Zain Sarwar, Sifat Muhammad Abdullah, Abdullah Rehman, Yoonjin Kim, Parantapa Bhattacharya, Mobin Javed, and Bimal Viswanath. Deepfake text detection: Limitations and opportunities. In *2023 IEEE symposium on security and privacy (SP)*, pages 1613–1630. IEEE, 2023.

[12] Yongqiang Ma, Jiawei Liu, Fan Yi, Qikai Cheng, Yong Huang, Wei Lu, and Xiaozhong Liu. Is this abstract generated by ai? a research for the gap between ai-generated scientific text and human-written scientific text. *arXiv preprint arXiv:2301.10416*, 2023.

[13] Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. Defending against neural fake news. *Advances in neural information processing systems*, 32, 2019.

[14] Daniel Martens and Walid Maalej. Towards understanding and detecting fake reviews in app stores. *Empirical Software Engineering*, 24(6):3316–3355, 2019.

[15] Tiziano Fagni, Fabrizio Falchi, Margherita Gambini, Antonio Martella, and Maurizio Tesconi. Tweepfake: About detecting deepfake tweets. *Plos one*, 16(5):e0251415, 2021.

[16] Hongpeng Wang, Rong Du, Wenqi Shen, Liangfei Qiu, and Weiguo Fan. Product reviews: A benefit, a burden, or a trifle? how seller reputation affects the role of product reviews. *How Seller Reputation Affects the Role of Product Reviews (June 23, 2021). Forthcoming in MIS Quarterly*, 2021.

[17] Dezhi Yin, Triparna de Vreede, Logan M Steele, and Gert-Jan de Vreede. Decide now or later: making sense of incoherence across online reviews. *Information Systems Research*, 34(3):1211–1227, 2023.

[18] Sherry He, Brett Hollenbeck, and Davide Proserpio. The market for fake reviews. *Marketing Science*, 41(5):896–921, 2022.

[19] Joni Salminen, Chandrashekhar Kandpal, Ahmed Mohamed Kamel, Soon-gyo Jung, and Bernard J Jansen. Creating and detecting fake reviews of online products. *Journal of Retailing and Consumer Services*, 64:102771, 2022.

[20] Raffaele Filieri. What makes an online consumer review trustworthy? *Annals of Tourism Research*, 58:46–64, 2016.

[21] Souheila Kaabachi, Selima Ben Mrad, and Maria Petrescu. Consumer initial trust toward internet-only banks in france. *International Journal of Bank Marketing*, 35(6):903–924, 2017.

[22] Ackerloff George et al. The market for lemons: Quality uncertainty and the market mechanism. 1970.

[23] Yuanshun Yao, Bimal Viswanath, Jenna Cryan, Haitao Zheng, and Ben Y Zhao. Automated crowdturfing attacks and defenses in online review systems. In *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security*, pages 1143–1158, 2017.

[24] Sunil Sharma and Satish Kumar. Insights into the impact of online product reviews on consumer purchasing decisions: A survey-based analysis of brands' response strategies. *Scholedge International Journal of Management & Development*, 10(1), 2023.

[25] Daria Plotkina, Andreas Munzel, and Jessie Pallud. Illusions of truth—experimental insights into human and algorithmic detections of fake online reviews. *Journal of Business Research*, 109:511–523, 2020.

[26] Ramadhani Ally Duma, Zhendong Niu, Ally S Nyamawe, Jude Tchaye-Kondi, and Abdulganiyu Abdu Yusuf. A deep hybrid model for fake review detection by jointly leveraging review text, overall ratings, and aspect ratings. *Soft Computing*, 27(10):6281–6296, 2023.

[27] Petr Hajek, Lubica Hikkerova, and Jean-Michel Sahut. Fake review detection in e-commerce platforms using aspect-based sentiment analysis. *Journal of Business Research*, 167:114143, 2023.

[28] Jiwei Luo, Guofang Nan, Dahui Li, and Yong Tan. Ai-generated review detection. *Available at SSRN 4610727*, 2023.

[29] AJ Alvero, Jinsook Lee, Alejandra Regla-Vargas, René F Kizilcec, Thorsten Joachims, and Anthony Lising Antonio. Large language models, social demography, and hegemony: comparing authorship in human and synthetic text. *Journal of Big Data*, 11(1):1–28, 2024.

[30] Julien Fontanarava, Gabriella Pasi, and Marco Viviani. Feature analysis for fake review detection through supervised classification. In *2017 IEEE international conference on data science and advanced Analytics (DSAA)*, pages 658–666. IEEE, 2017.

[31] N Gobi and A Rathinavelu. Analyzing cloud based reviews for product ranking using feature based clustering algorithm. *Cluster Computing*, 22(Suppl 3):6977–6984, 2019.

[32] Christopher G Harris. Comparing human computation, machine, and hybrid methods for detecting hotel review spam. In *Digital Transformation for a Sustainable Society in the 21st Century: 18th IFIP WG 6.11 Conference on e-Business, e-Services, and e-Society, I3E 2019, Trondheim, Norway, September 18–20, 2019, Proceedings 18*, pages 75–86. Springer, 2019.

[33] Rami Mohawesh, Haythem Bany Salameh, Yaser Jararweh, Mohannad Alkhalaileh, and Sumbal Maqsood. Fake review detection using transformer-based enhanced lstm and roberta. *International Journal of Cognitive Computing in Engineering*, 5:250–258, 2024.

[34] Majd AbedRabbo, Cathryn Hart, Fiona Ellis-Chadwick, and Zeina AlMalak. Towards rebuilding the highstreet: Learning from customers' town centre shopping journeys. *Journal of Retailing and Consumer Services*, 64:102772, 2022.

[35] Nitin Jindal and Bing Liu. Opinion spam and analysis. In *Proceedings of the 2008 international conference on web search and data mining*, pages 219–230, 2008.

[36] Kyung-Hyan Yoo and Ulrike Gretzel. Comparison of deceptive and truthful travel reviews. In *Information and communication technologies in tourism 2009*, pages 37–47. Springer, 2009.

[37] Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T Hancock. Finding deceptive opinion spam by any stretch of the imagination. *arXiv preprint arXiv:1107.4557*, 2011.

[38] Vlad Sandulescu and Martin Ester. Detecting singleton review spammers using semantic similarity. In *Proceedings of the 24th international conference on World Wide Web*, pages 971–976, 2015.

[39] Jianmo Ni, Jiacheng Li, and Julian McAuley. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 188–197, 2019.

[40] Alessandro Gambetti and Qiwei Han. Combat ai with ai: Counteract machine-generated fake restaurant reviews on social media. *arXiv preprint arXiv:2302.07731*, 2023.

[41] Arwa Bader, Yazan Suhweil, Bushra Alhijawi, Saleh Abu-Soud, et al. Detecting chatgpt generated fake reviews using supervised machine learning. In *2023 14th International Conference on Information and Communication Systems (ICICS)*, pages 1–5. IEEE, 2023.

[42] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[43] M Lewis. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

[44] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

[45] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[46] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International conference on machine learning*, pages 11328–11339. PMLR, 2020.

[47] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

[48] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

[49] Paulo Duarte, Susana Costa e Silva, and Margarida Bernardo Ferreira. How convenient is it? delivering online shopping convenience to enhance customer satisfaction and encourage e-wom. *Journal of Retailing and Consumer Services*, 44:161–169, 2018.

[50] Seiji Endo, Jun Yang, and JungKun Park. The investigation on dimensions of e-satisfaction for online shoes retailing. *Journal of Retailing and Consumer Services*, 19(4):398–405, 2012.

[51] Kapil Kaushik, Rajhans Mishra, Nripendra P Rana, and Yogesh K Dwivedi. Exploring reviews and review sequences on e-commerce

platform: A study of helpful reviews on amazon. in. *Journal of retailing and Consumer Services*, 45:21–32, 2018.

[52] Sandra MC Loureiro, Luisa Cavallero, and Francisco Javier Miranda. Fashion brands on retail websites: Customer performance expectancy and e-word-of-mouth. *Journal of Retailing and Consumer Services*, 41:131–141, 2018.

[53] Gina A Tran and David Strutton. Comparing email and sns users: Investigating e-servicescape, customer reviews, trust, loyalty and e-wom. *Journal of Retailing and Consumer Services*, 53:101782, 2020.

[54] Aaron Smith and Monica Anderson. Online shopping and e-commerce. 2016.

[55] Qing Cao, Wenjing Duan, and Qiwei Gan. Exploring determinants of voting for the "helpfulness" of online user reviews: A text mining approach. *Decision Support Systems*, 50(2):511–521, 2011.

[56] Kolli Shivagangadhar, H Sagar, Sohan Sathyan, and CH Vanipriya. Fraud detection in online reviews using machine learning techniques. *International Journal of Computational Engineering Research (IJCER)*, 5(5):52–56, 2015.

[57] Youli Fang, Hong Wang, Lili Zhao, Fengping Yu, and Caiyu Wang. Dynamic knowledge graph based fake-review detection. *Applied Intelligence*, 50:4281–4295, 2020.

[58] Ning Wang, Jun Yang, Xuefeng Kong, and Ying Gao. A fake review identification framework considering the suspicion degree of reviews with time burst characteristics. *Expert Systems with Applications*, 190:116207, 2022.

[59] Xiaomin Yu, Yezhaohui Wang, Yanfang Chen, Zhen Tao, Dinghao Xi, Shichao Song, and Simin Niu. Perceived authenticity mediates the relationships between online review source (human vs chatgpt) and online review trust and online review usefulness. *arXiv preprint arXiv:2405.00711*, 2024.

[60] Clinton Amos and Lixuan Zhang. Consumer reactions to perceived undisclosed generative ai usage in an online review context. *Telematics and Informatics*, page 102163, 2024.

[61] Konstantinos F Xylogiannopoulos, Petros Xanthopoulos, Panagiotis Karampelas, and Yiorgos Bakamitsos. Is ai-assisted paraphrase the new tool for fake review creation? challenges and remedies. *Challenges and Remedies*.

[62] Chaka Chaka. Reviewing the performance of ai detection tools in differentiating between ai-generated and human-written texts: A literature and integrative hybrid review. *Journal of Applied Learning and Teaching*, 7(1), 2024.

[63] Yinhan Liu. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[64] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pages 43–52, 2015.

[65] Ruining He and Julian McAuley. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *proceedings of the 25th international conference on world wide web*, pages 507–517, 2016.

[66] Mengting Wan and Julian McAuley. Item recommendation on monotonic behavior chains. In *Proceedings of the 12th ACM conference on recommender systems*, pages 86–94, 2018.

[67] Md Hijbul Alam, Woo-Jong Ryu, and SangKeun Lee. Joint multi-grain topic sentiment: modeling semantic aspects for online reviews. *Information Sciences*, 339:206–223, 2016.

[68] Abien Fred Agarap. Statistical analysis on e-commerce reviews, with sentiment classification using bidirectional recurrent neural network (rnn). *arXiv preprint arXiv:1805.03687*, 2018.

[69] Abien Fred Agarap. Afagarap/ecommerce-reviews-analysis: v0.1.0-alpha, March 2018.

[70] Faranak Abri, Luis Felipe Gutiérrez, Akbar Siami Namin, Keith S Jones, and David RW Sears. Linguistic features for detecting fake reviews. In *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 352–359. IEEE, 2020.

[71] Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arXiv:2301.07597*, 2023.

[72] Md Shahidul Salim and Sk Imran Hossain. An applied statistics dataset for human vs ai-generated answer classification. *Data in Brief*, 54:110240, 2024.

[73] Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Chenxi Whitehouse, Osama Mohammed Afzal, Tarek Mahmoud, Toru Sasaki, et al. M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection. *arXiv preprint arXiv:2305.14902*, 2023.

[74] Lorenz Mindner, Tim Schlippe, and Kristina Schaaff. Classification of human-and ai-generated texts: Investigating features for chatgpt. In *International Conference on Artificial Intelligence in Education Technology*, pages 152–170. Springer, 2023.

[75] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[76] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.

[77] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[78] Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020.