

Article

Mitigating Class Imbalance in Sentiment Analysis through GPT-3-Generated Synthetic Sentences

Cici Suhaeni * and Hwan-Seung Yong

Department of Computer Science and Engineering, Ewha Womans University, Seoul 03760, Republic of Korea; hsyong@ewha.ac.kr

* Correspondence: cici.suhaeny@gmail.com

Abstract: In this paper, we explore the effectiveness of the GPT-3 model in tackling imbalanced sentiment analysis, focusing on the Coursera online course review dataset that exhibits high imbalance. Training on such skewed datasets often results in a bias towards the majority class, undermining the classification performance for minority sentiments, thereby accentuating the necessity for a balanced dataset. Two primary initiatives were undertaken: (1) synthetic review generation via fine-tuning of the Davinci base model from GPT-3 and (2) sentiment classification utilizing nine models on both imbalanced and balanced datasets. The results indicate that good-quality synthetic reviews substantially enhance sentiment classification performance. Every model demonstrated an improvement in accuracy, with an average increase of approximately 12.76% on the balanced dataset. Among all the models, the Multinomial Naïve Bayes achieved the highest accuracy, registering 75.12% on the balanced dataset. This study underscores the potential of the GPT-3 model as a feasible solution for addressing data imbalance in sentiment analysis and offers significant insights for future research.

Keywords: GPT-3; imbalanced sentiment analysis; sentiment analysis; sentiment classification; synthetics review generation; text classification; text generation



Citation: Suhaeni, C.; Yong, H.-S. Mitigating Class Imbalance in Sentiment Analysis through GPT-3-Generated Synthetic Sentences. *Appl. Sci.* **2023**, *13*, 9766. <https://doi.org/10.3390/app13179766>

Academic Editors: Yunbo Rao, Yadong Wu and Zhihan Lv

Received: 1 August 2023

Revised: 26 August 2023

Accepted: 27 August 2023

Published: 29 August 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The significance of sentiment analysis has extended across a wide range of fields, finding extensive use in various applications. As digital communication continues to expand, the ability of sentiment analysis to interpret complex human emotions and opinions becomes increasingly important, proving invaluable in fields ranging from social sciences to customer service and beyond. In this era of increasing digitization, leveraging the power of data through sentiment analysis offers unique insights, making significant contributions to sectors such as those previously summarized in various studies, namely, healthcare, social policy, e-commerce, and digital humanities [1]. In the year 2023, sentiment analysis experienced a significant surge in usage, employing advanced techniques to analyze diverse data sources. Twitter feeds concerning global events such as the COVID-19 pandemic [2] and presidential elections [3] have been analyzed, including the identification of harmful comments on social networks [4]. Furthermore, sentiment analysis expanded its reach to encompass various languages, including African languages [5], further extending its impact. The rapid evolution of sentiment analysis is evident in the extensive body of literature dedicated to the field [6]. Numerous reviews have provided comprehensive insights into the current state of sentiment analysis studies [7,8], with some studies even specifically focusing on sentiment analysis of Twitter data [9]. Notably, sentiment analysis is not limited to social media but has also gained extensive application in the evaluation of reviews and comments, underscoring its significance in assessing services or products.

Contemporary studies have explored sentiment analysis within the context of comments or reviews, covering various areas such as online courses [10], Amazon product reviews [11], film reviews [12], hotel online reviews [13], online product reviews [14,15],

and online learning reviews [16]. These studies predominantly employ machine learning and deep learning techniques. This research aligns with the rapid evolution of sentiment analysis discussed earlier, where advanced techniques are utilized to analyze diverse data sources. The application of machine learning and deep learning techniques in contemporary studies signifies the ongoing exploration and refinement of sentiment analysis methods across different domains. Moreover, these studies contribute to the extensive body of literature dedicated to sentiment analysis, further enhancing our understanding of its effectiveness in evaluating and interpreting comments or reviews.

However, one major challenge in sentiment analysis is the prevalence of imbalanced datasets in real-world scenarios, including review datasets. Typically, people tend towards expressing either positive or negative sentiments, resulting in relatively fewer neutral reviews. This imbalance often results in subpar prediction and classification outcomes for minority classes due to a lack of sufficient training data. Several strategies have been proposed to address this issue, with recent trends favoring synthetic data generation for the minority class. One approach involves using Generative Adversarial Networks (GANs) for synthetic text generation [17]. Although the GAN method offers improved performance over imbalanced data, there remains significant potential for further enhancement.

Building on this potential for enhancement, this study proposes the utilization of the state-of-the-art language model, GPT-3, which boasts an enormous 175 billion parameters [18]. With its remarkable capability to generate human-like text, GPT-3 can be employed to generate synthetic reviews, supplementing the sparse minority class and balancing the dataset.

The core goal of this research is to address imbalanced sentiment analysis by generating synthetic reviews using the GPT-3 model. To achieve this, this study focuses on fine-tuning the GPT-3 model to generate synthetic texts that supplement the minority class, aiming to produce synthetic reviews that are contextually relevant to the original data. Fine-tuning the GPT-3 model with our specific dataset is important. This process allows the model to adapt to the unique characteristics and nuances of our data, thus increasing its ability to generate relevant synthetic reviews. With a more specialized model, the quality of synthetic reviews can be significantly improved, yielding more reliable and useful results for sentiment classification. In essence, the fine-tuning process personalizes the robust GPT-3 model to our specific use case, ensuring that it operates optimally within the context of our data and objectives. The findings of this study emphasize the outstanding performance of GPT-3 in generating good-quality synthetic reviews, significantly enhancing sentiment classification results, with an average accuracy increase of 12.76% for all implemented models.

This paper makes the following significant contributions:

- Exploration and Experimentation on Fine-tuning GPT-3
- We conduct detailed research and experiments on fine-tuning the GPT-3 model using the OpenAI API. This contribution focuses on generating synthetic reviews that are contextually relevant to the original data.
- Evaluation of GPT-3 Synthetic Review Generation
- This paper provides a thorough evaluation of the synthetic reviews generated by GPT-3, examining their quality and relevance to the original data context.
- Investigation of Synthetic Reviews' Impact on Sentiment Classification
- We systematically investigate the impact of high-quality synthetic reviews, generated using GPT-3, on the performance of sentiment classification. This demonstrates the model's potential to enhance sentiment analysis accuracy.
- Comparative Analysis of Machine Learning and Deep Learning Models

We conduct a comparative performance analysis of nine machine learning and deep learning models, using both imbalanced and balanced data, specifically in the context of online course reviews. This provides valuable insights into the effectiveness of our approach in handling class imbalance issues.

These contributions significantly advance our understanding of how the state-of-the-art GPT-3 model can be effectively fine-tuned and utilized for synthetic review generation, subsequently enhancing sentiment analysis performance in imbalanced datasets.

The structure of this paper is as follows: In Section 2, we provide a comprehensive analysis of the pertinent literature, focusing specifically on synthetic text generation and imbalanced sentiment analysis. Sections 3 and 4 offer a deep dive into the research methodology, explaining in detail the datasets, preprocessing methods, the design of the text generation models and the sentiment classification, and the metrics used for evaluation. Moving to Section 5, we showcase the outcomes derived from the application of our chosen text generation models as well as the sentiment analysis experiments, performed on both original and balanced datasets. Section 6 presents an extensive discussion of these results, while also suggesting possible directions for future research.

2. Related Work

This section provides a short overview of recent studies in two main areas: text generation models and imbalanced sentiment analysis. We investigate widely used methods for generating text, especially those relevant to sentiment analysis, and we look at different strategies that have been applied to address the issue of imbalanced sentiment analysis.

2.1. Text Generation

A recent systematic literature review by Fatima et al. [19] scrutinized 90 primary studies conducted from 2015 to 2021, which highlighted methods for generating text, quality measures, datasets, and languages, along with their usage in the context of deep learning. This review emphasized the escalating interest in deep learning methodologies for text generation over the studied period. Significantly, it highlighted the potential of GPT-3 in generating text due to its extensive training and substantial generative capabilities. Iqbal and Qureshi [20] furthered this by demonstrating that current deep learning methods applied in the realm of synthetic text creation encompass Variational Auto-Encoders (VAEs) and Generative Adversarial Networks (GANs).

GAN-based text generation has been explored extensively in recent studies. Wang and Wan [21] unveiled a fresh architectural framework—SentiGAN, which encompasses multiple generators and one multi-class discriminator, all architected specifically to concoct a wide range of examples that all carry a specific sentiment label. Building on this, Liu et al. [22] advanced this framework by proposing a GAN that is aware of its category (CatGAN). This was equipped with an efficient model for generating text according to its category, in addition to a hierarchical algorithm for evolutionary learning dedicated to training the model.

The revolutionary “Transformer” model was introduced by Vaswani et al. [23], providing the groundwork for subsequent language generation models, including the GPT and BERT architectures. Following this, Radford et al. [24] presented a seminal paper introducing the GPT-2 model, a noteworthy development in the field of language generation.

Recent studies have employed GPT-2 in various innovative ways for text generation. Anaby-Tavor et al. [25] leveraged GPT-2 in a method called LAMBADA, while Ma et al. [26] proposed the Switch-GPT method. Xu et al. [27] used GPT-2 and T5 to generate table captions, and Bayer et al. [28] also utilized GPT-2, but because of some limitations, they suggested GPT-3 as a viable choice for enhancing results, having utilized GPT-2 in their proposed method for text generation.

The introduction of GPT-3, the successor of GPT-2, marked another milestone in this field [18]. Recently, Zhong et al. [29] investigated the understanding ability of ChatGPT, a GPT model variant, by subjecting it to the well-known GLUE benchmark test and juxtaposing its performance against four emblematic models that had been fine-tuned in the style of BERT. These studies form the backbone of our understanding and application of text generation and sentiment analysis, with this paper intending to contribute further to this growing body of knowledge.

2.2. Imbalanced Sentiment Analysis

The topic of imbalanced sentiment analysis has been a vibrant area of research in recent years, with numerous approaches developed to tackle this problem.

Obiedat et al. [30] introduced a combined method that melds the Support Vector Machine (SVM) algorithm with Particle Swarm Optimization (PSO), along with several oversampling methods to tackle the problem of unbalanced sentiment analysis within a dataset of customer reviews. This tactic proved successful in dealing with data disparity, showcasing the promise of these hybrid methods in this field.

Han Wen and Junfang Zhao [31] introduced an alternate strategy, which suggested a technique for sentiment evaluation of unbalanced comment data utilizing a BiLSTM structure. The approach involved Adaptive Synthetic Sampling in cases where the dataset contained more negative instances than positive ones, deploying a model based on CNN-BiLSTM for classifying the sentiment.

In the same spirit, Tan et al. [32] crafted an innovative hybrid system that amalgamates the advantages of the Transformer model, exemplified by the Robustly Optimized BERT Pretraining Approach (RoBERTa), and the Recurrent Neural Network, embodied by Gated Recurrent Units (GRUs). This hybrid system was engineered to address the issue of unbalanced datasets by applying data augmentation via word embeddings, while oversampling the minority class, thereby boosting the model's ability to represent data and its resilience in executing sentiment classification tasks.

Wu and Huang [33] proposed a different method for handling imbalanced text data. They introduced a hybrid method, which utilizes a generative adversarial network alongside the Shapley algorithm, termed HEGS. This structure could produce a wide range of training phrases to level the textual data and bolster the ability to classify instances belonging to the minority classes.

Almuayqil et al. [34] took an innovative approach by designing a model specifically for imbalanced Twitter datasets. By utilizing an array of text sequencing preprocessing methods combined with random under-sampling of the majority class, they managed to considerably cut down the computational time required for the task.

Further investigating Twitter data, Ghosh et al. [35] assessed the efficacy of varying proportions of synthetic oversampling techniques to manage class imbalance in Twitter sentiment analysis. Concurrently, George [36] introduced a unique synthetic oversampling method, SMOTE, amalgamated with a composite model referred to as the Ensemble Bagging Support Vector Machine (EBSVM), to address the problem of data imbalance.

Cai and Zhang [37] adopted a unique perspective by concentrating on sentiment information extraction from an imbalanced short text review dataset. They introduced a fusion multi-channel BLTCN-BLSTM self-attention sentiment classification strategy, amalgamating focus loss rebalancing and classifier enhancement mechanisms to boost sentiment prediction accuracy.

A recent approach to handling imbalanced sentiment analysis is by generating artificial text for minority classes. Imran et al. [17] utilized a GAN-based model to generate synthetic data for tackling this problem. Similarly, Habbat et al. [38] employed a pretrained AraGPT-2-based model to create synthetic Arabic text, addressing the issue of imbalanced sentiment analysis. Following this, they utilized AraBERT for textual representation and a deep learning model stack for classification. This research illuminates the potential of language-specific models in proficiently managing tasks related to imbalanced sentiment analysis.

Lastly, Ekinici [39] performed a comparative study of imbalanced offensive data classification using an LSTM-based sentence generation method. Various classifiers were trained using TF-IDF and Word2vec for text representation, demonstrating the value of sentence generation methods in handling imbalanced sentiment analysis tasks.

Together, these studies highlight the diverse methods and models available to handle imbalanced sentiment analysis, and they set the foundation for further research in this field.

3. Proposed Approach

3.1. Problem Formulation

In the domain of sentiment analysis, including for online learning platforms like Coursera, dataset imbalance often becomes an important issue. A previous study [17] attempted to tackle this issue using a GAN-based model to generate synthetic data. While this approach showed progress in enhancing the classification performance of the imbalanced dataset, there remains potential for further significant improvements in addressing imbalanced class sentiment analysis using different approaches.

Given the limitations of existing solutions, a pivotal question arises: “How can we more effectively produce synthetic data that can significantly enhance the classification performance of an imbalanced dataset in sentiment analysis, particularly for the Coursera review dataset?”.

3.2. Towards the Proposed Approach

In response to the question, this research proposes another approach, which is GPT-3-based generated synthetic reviews. Recognizing the prowess of GPT-3 as an advanced and expansive language model, we believe that fine-tuning this model can lead to the generation of good-quality and contextually relevant synthetic data. We chose to use GPT-3 due to its well-established expertise in understanding and producing human-like sophisticated text. Thus, by incorporating GPT-3, we expect a more marked improvement in classification performance compared to previous methods.

Figure 1 provides a thorough description of the proposed approach. Initially, our methodology pivots on the original Coursera review dataset, which has an imbalanced distribution among positive, negative, and neutral sentiments. Acknowledging the imbalance, we next focus on the negative and neutral sentiments, which are the minority classes. By focusing on these sentiments, we set the groundwork for creating synthetic reviews with a fine-tuned GPT-3 model.

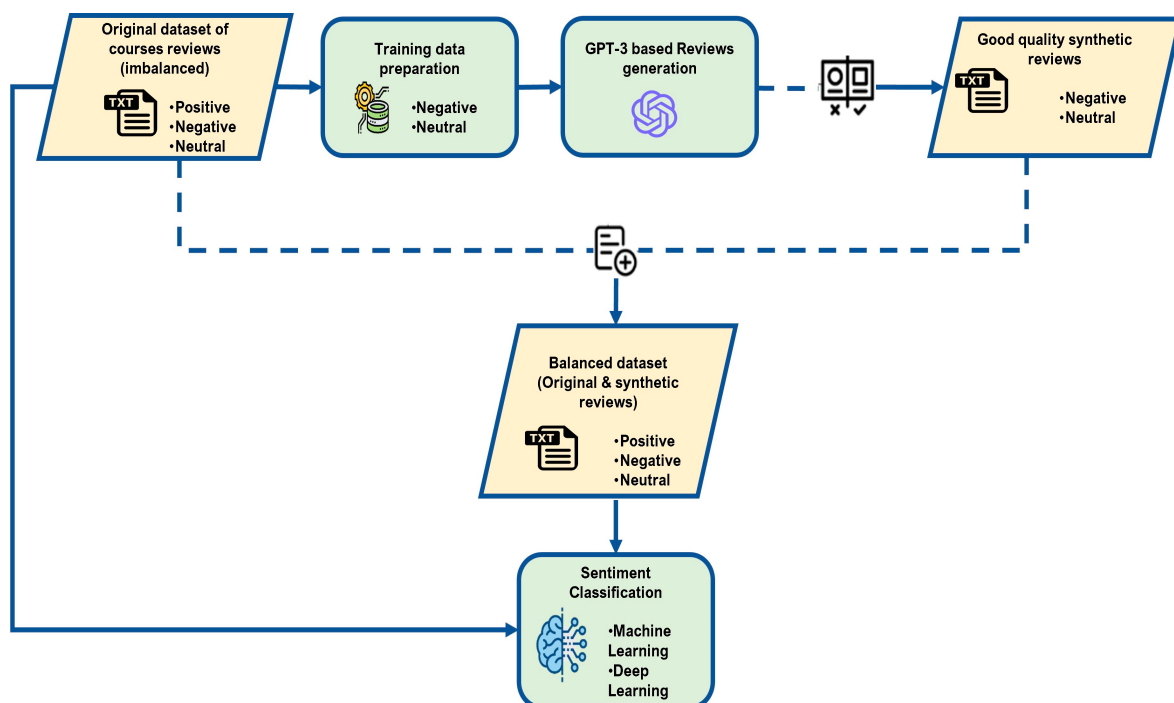


Figure 1. Comprehensive overview of the proposed sentiment classification approach.

Before the GPT-3 model can start being fine-tuned, the preparation of training data is essential. A detailed process of this preparation, alongside the fine-tuning, is illustrated comprehensively in Figure 3. After undergoing a rigorous evaluation process, the culmination of this generation process yields good-quality synthetic reviews. These reviews

are then deemed fit to be added to the original dataset, ensuring a balanced data distribution. After the data preprocessing phase, sentiment classification is conducted utilizing 10 machine learning and deep learning models. To determine the level of improvement made by this balanced data classification, a comparison is made by also classifying the original imbalanced data.

This holistic strategy is our guiding light, leading us to our main goal: good-quality synthetic reviews that significantly elevate the classification performance. By integrating these synthetic reviews into our dataset, we aim to provide a more robust foundation for sentiment analysis.

4. Experimental Detail

4.1. Dataset

For the purpose of this study, our main focus is on a set of reviews that were collected from the Coursera online learning platform [16]. These reviews, written exclusively in the English language, cover a wide range of 15 different courses. These selected courses represent a diverse array of subjects and are facilitated by various instructors, thus providing a comprehensive and wide-ranging dataset for our analysis. The dataset contains a total of 21,937 reviews, each of which has been categorized into one of three sentiment polarity classes: positive, negative, or neutral (as depicted in Figure 2a). These reviews offer insights into five critical aspects of online courses: content, instructor, structure, design, and a general assessment of the course (as shown in Figure 2b). Each aspect carries unique significance in evaluating the overall course quality and learner experience. A detailed definition of each aspect can be found in [16]. The distribution according to these aspects is crucial and serves as a reference for determining prompts during the synthetic data generation process and for guiding the splitting of the dataset into training, validation, and testing subsets.

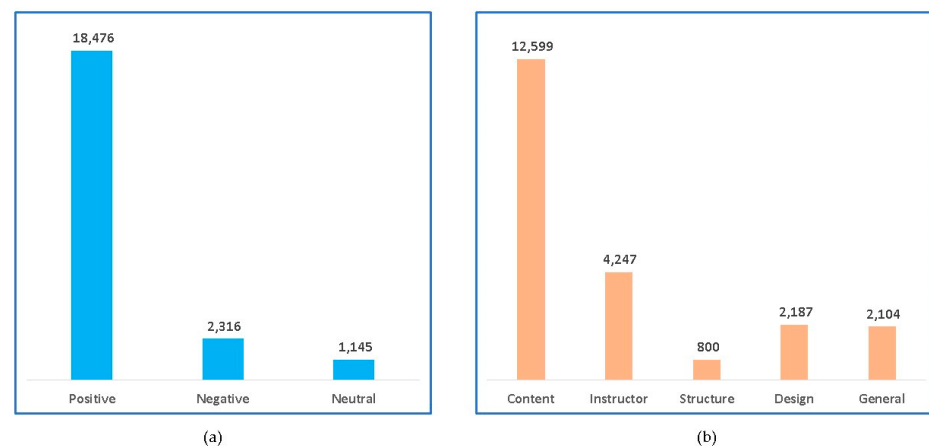


Figure 2. The frequency distribution of the Coursera reviews dataset: (a) based on sentiment. (b) based on aspect.

A notable characteristic of this dataset is its high-class imbalance. Of the total reviews, a significant majority, 18,476 (84.2%) reviews, are positive, while the negative and neutral reviews are comparatively fewer, with 2316 (10.6%) and 1145 (5.2%) reviews, respectively. This skewed distribution presents a considerable challenge for accurate and unbiased sentiment analysis.

4.2. Synthetic Reviews Generation

The generation of synthetic data stands as the foremost endeavor in this study, representing our proposed solution to augment sentiment classification performance for imbalanced data. By integrating the capabilities of the GPT-3 model, we aim to construct

synthetic reviews that not only enhance the data balance but also uphold the authenticity and nuances of real-world feedback. Figure 3 vividly illustrates this intricate process.

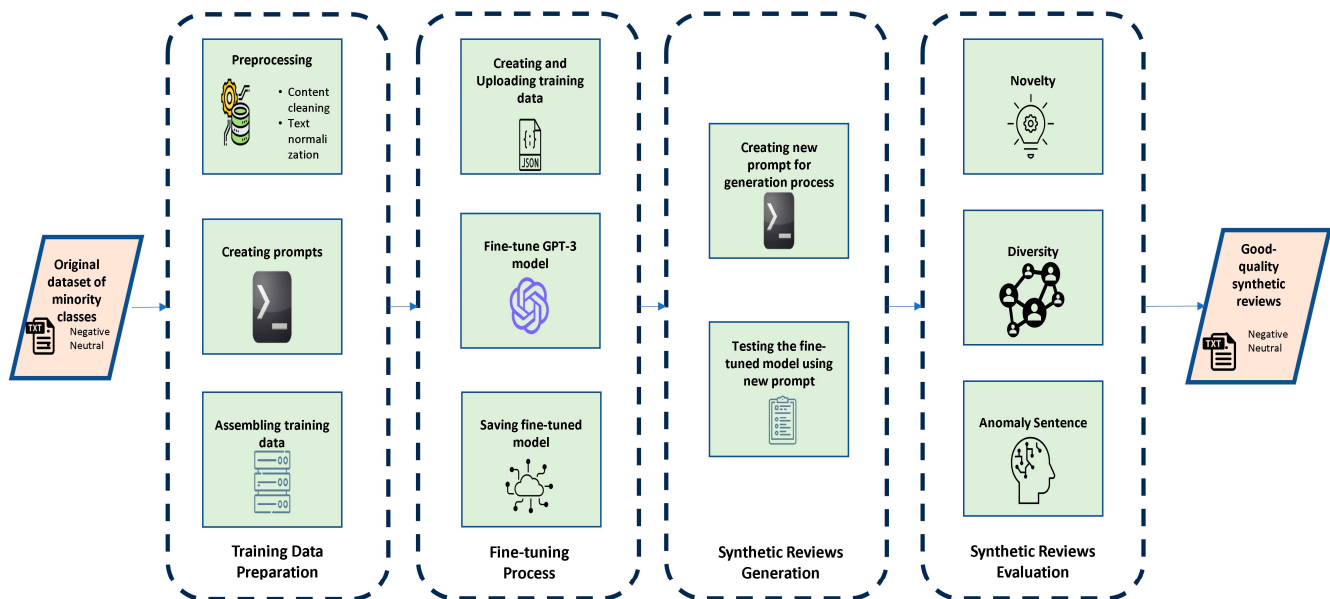


Figure 3. GPT-3-based synthetic reviews generation process.

The process commences with the utilization of original data corresponding to the negative and neutral sentiments, functioning as the primary input. This methodology has been partitioned into four distinct stages. The first three stages are structured based on references from OpenAI documentation [40]. The initial stage emphasizes preparing the training data, which encompasses tasks such as preprocessing, prompt creation, and the assembly of the final training data. Subsequently, the GPT-3 model undergoes a fine-tuning process, ensuring that it is tailor-made to generate reviews mimicking the essence of the original data. Once the model has been fine-tuned, it is ushered into the synthetic review generation phase, where new prompts guide the creation of synthetic reviews. The final and perhaps the most crucial step is the rigorous evaluation of these generated reviews, focusing on novelty, diversity, and manual evaluation metrics. The culmination of this entire procedure results in the procurement of good-quality synthetic reviews, vital for enhancing the robustness of sentiment analysis.

Stage 1: Training Data Preparation

In this initial stage, the foundation for the fine-tuning process is laid out through three critical steps. The explanation of each step is as follows:

Step 1: Preprocessing

In the preprocessing step, we worked on the original reviews that had negative and neutral sentiments as the minority classes. Our goal was to pick the best samples from this dataset. These top-notch reviews would later serve as “completions” for training the model.

We tackled two main tasks: Content Cleaning and Text Normalization. For the cleaning bit, we made sure the language was consistent, got rid of repeated reviews, and made sure the sentiments were clear. We removed non-English sentences, duplicate reviews, and those with vague sentiments. In addition, any email addresses or other unnecessary details were thrown out.

For Text Normalization, we took out things like emojis, which do not add sentiment value. We also simplified punctuation, making the text more readable. For instance, if there were many dots in a row, we changed them to just one. Examples of data training preprocessing are presented in Table 1.

Table 1. Distribution of prompts and completions across sentiments and aspects.

Original Text	After Preprocessing
I'm not finishing it since in the test are missing a lot of sounds	I'm not finishing it since in the test are missing a lot of sounds
the instructor says in the first video: "you do not need to read a book on how to write a script" in my opinion you do! better to read a book or two than waste your time on this course; the videos are extremely short, shallow and a waste of time!	the instructor says in the first video you do not need to read a book on how to write a script in my opinion you do better to read a book or two than waste your time on this course the videos are extremely short shallow and a waste of time
very poor. never expected this from coursera. the person is just reading the slides, there is no learning action taking place	very poor never expected this from coursera. the person is just reading the slides there is no learning action taking place

Step 2: Creating prompts.

The training data for the GPT-3 model guide the desired output, which must be formatted as a JSONL document. Each line of this document presents a prompt-completion pair, serving as an individual training example, as exemplified below:

```
{"prompt": "<prompt text>", "completion": "<ideal generated text>"}
```

The effectiveness of the model improves with the number of these examples, with a recommendation of several hundred examples as a baseline.

For our research, we used a dataset comprised of Coursera course reviews, which were categorized into five aspects: *content*, *instructor*, *structure*, *design*, and *general*. To ensure a balanced representation across synthetic reviews, we proportionally generated reviews following the distribution of aspects within the original data as presented in Figure 2b. Consequently, we prepared 2283 preprocessed negative reviews and 1133 neutral reviews as training data, generating 65 and 54 diverse prompts, respectively, for each category. At its core, our training data consist of prompt and completion pairs. To mitigate the impracticality of creating a unique prompt for each completion, we created an array of diverse prompts. These prompts were arranged in a sequential and repeating pattern within the training data. Table 2 provides a detailed breakdown of the number of prompts and completions across different review aspects.

Table 2. Distribution of prompts and completions across sentiments and aspects.

Sentiment	Aspect	The Number of Prompts	The Number of Completions
Negative	Content	25	1740
	Instructor	13	217
	Structure	10	120
	Design	10	156
	General	7	50
Neutral	Content	20	652
	Instructor	8	71
	Structure	10	100
	Design	10	260
	General	6	50

Step 3: Assembling training data

Assembling training data essentially refers to the process of systematically putting together the "prompt" and its corresponding "completion" to form structured training examples for the model. This arrangement ensures that the GPT-3 model receives precise guidelines on the desired output.

Using our preprocessed reviews (Step 1) and the prompts we created (Step 2), we structured our training data such that for each prompt, there is an appropriate review completion that resonates with the sentiment and aspect the prompt is geared towards. It is

crucial to align the prompt correctly with its completion to ensure that the model captures the desired sentiment and content nuances effectively.

Table 3 provides a snapshot of this assembly, showcasing examples of training data specifically for generating negative sentiment reviews. As can be observed, each prompt is meticulously paired with a completion that is relevant and mirrors the context of the prompt, ensuring that the model learns the intricacies of generating reviews that are coherent and contextually accurate.

Table 3. Example of training data for negative sentiment review generation.

Prompt	Completion
Describe the negative sentiment review of the content of this course	much better than the data science with r concentration provided also by john hopkins. this course has concrete examples and the lecturer doesn't treat his audience like the firstgrade kids. i take this back. the lecturer doesn't know what he's taking about and what he is gonna talk about for at least half of the time.
Please generate negative review about course content in Coursera	a very weak course gives very little knowledge of grammar doesn't really explain quite a lot of things. if you are a first year primary school student this course might be helpful otherwise a complete waste of time.
How did the instructor's performance impact your learning experience negatively?	videos rather slow and don't keep you focus. voice of the professor is rather low key and slow. hard to pay attention at times. even though there are examples it should be more interactive with other videos rather than links of text.
Were there any instances where you felt the instructor lacked sufficient knowledge or experience in the subject matter?	too much version specific details. instructor was reading slides vs. explaining the technologies (telling a story) the quiz part cant be more lame

For example:

The prompt “Describe the negative sentiment review of the content of this course” is paired with a completion that criticizes the course’s content and the lecturer’s manner of teaching.

In essence, this step of assembling training data ensures that the GPT-3 model is exposed to diverse scenarios, enabling it to generate synthetic reviews that are credible and contextually rich.

Stage 2: Fine-tuning Process

The fine-tuning process is pivotal in customizing a pretrained model to address specific tasks or domains. By focusing the model on our curated dataset, we intended to tap into its robust capabilities, tailoring them further to our distinct needs. The main steps for this process are as follows:

Step 1: Creating and Uploading Training Data

After meticulously preparing the training data in the previous stage, the next step was to present it to the OpenAI system. To ensure compatibility with OpenAI’s environment, we transformed our dataset into the JSON Lines (JSONL) format. This format allows each line in the file to be a valid JSON entry, streamlining the processing of large datasets. Once transformed, the dataset was uploaded using the OpenAI File API. The entire data handling process was mediated via the OpenAI API, ensuring a seamless interaction between our data and the platform’s resources. Using the Python command “`openai.File.create()`”, the dataset was uploaded, generating a unique file ID essential for subsequent steps.

Step 2: Fine-Tuning the GPT-3 Model

With the data positioned appropriately within the OpenAI environment, we embarked on the fine-tuning procedure using the GPT-3 “Davinci” as our base model. The rationale behind employing this model is its versatility and extensive training on a wide spectrum of data, making it an ideal candidate for modification. The fine-tuning was initiated using the “`openai.FineTune.create()`” command in Python, where the uploaded training data

informed the model to better align its generation capabilities with our specific sentiment review use-case for Coursera.

Step 3: Saving the Fine-Tuned Model

Upon completion of the fine-tuning phase, the modified GPT-3 model was preserved for future tasks. Saving the fine-tuned model is of paramount importance, as it encapsulates the nuanced learning from our custom training data. The model ID was stored securely, enabling swift recall when generating synthetic reviews in subsequent stages. In essence, this stage marks the model's metamorphosis, transitioning from its generalized capabilities to a specialized tool adept at crafting synthetic reviews with an authentic flair, specifically tailored to Coursera reviews.

To provide a more concrete representation of the processes described in the steps, we have detailed the fine-tuning approach in the form of a pseudocode as presented in Algorithm 1. This algorithm offers a step-by-step breakdown, illustrating the sequence of operations and logic flow essential to fine-tuning the GPT-3 model with our specific dataset.

Algorithm 1: Fine-tuning GPT-3 Model for Review Generation

```

Input: Training dataset path
Output: Fine-tuned GPT-3 model ID
BEGIN
//Initialize System Libraries
Import openai, json, pandas
//Setup API Configuration
SET api_key to YOUR_OPENAI_API_KEY
Configure OpenAI with api_key
//Load Training Data
DataFrame df ← ReadDataFromSource(dataset path)
List training_data ← ConvertDataFrameToListOfDict(df)
Write training_data to JSONL file named "training_data.jsonl"
//Validate Training Data
CALL OpenAITool to validate "training_data.jsonl"
//Upload Training Data to OpenAI
UploadResponse upload_response ← OpenAI.UploadFile("training_data.jsonl", purpose =
'fine-tune')
String file_id ← Extract ID from upload_response
//Execute Fine-tuning Process
FineTuneResponse fine_tune_response ← OpenAI.FineTune(file_id, model = "davinci")
//Monitor Fine-tuning Progress
IF wanting to list events THEN
FineTuneEvents fine_tune_events ← OpenAI.ListFineTuneEvents(fine_tune_response.id)
ELSE
RetrieveResponse retrieve_response ← OpenAI.RetrieveFineTune(fine_tune_response.id)
END IF
//Save the Fine-tuned Model
String fine_tuned_model ← Extract FineTunedModel from fine_tune_response
OR
RetrieveResponse retrieve_response ← OpenAI.RetrieveFineTune(fine_tune_response.id)
String fine_tuned_model ← Extract FineTunedModel from retrieve_response
RETURN fine_tuned_model
END

```

Stage 3: Synthetic Reviews Generation

The creation of synthetic reviews is at the core of this study. Having successfully fine-tuned our model, the stage was set for its most crucial test: generating good-quality, relevant synthetic reviews that would reflect real-world sentiments towards Coursera courses. The fine-tuned model was now poised to take on new prompts, translating

them into a myriad of insightful synthetic reviews. In this study, our main goal was to create synthetic reviews of Coursera courses that covered five different aspects: *content*, *instructor*, *structure*, *design*, and *general*. For each of these aspects, we devised a new prompt. Prompts are the initial pieces of text given to the model, which it then uses as a guide to generate further text. In this case, prompts are designed to guide the model towards generating a review in the context of a particular aspect. Hence, a total of five new prompts were employed, each for one aspect of the reviews. Thus, by leveraging this fine-tuned model, we can generate relevant and good-quality synthetic reviews for Coursera courses across various aspects. The principle was simple: Using a singular prompt, the model was instructed to generate “n” synthetic reviews. Here is an example of how to generate synthetic reviews about an instructor’s quality using a single prompt to produce $n = 3$ reviews:

Input Prompt:

“Please write your negative opinion regarding the instructor in a course you have taken in Coursera”.

Generated reviews:

1. “The instructor is really boring. The content is not challenging, and I don’t think it is worth paying for. Should be included as part of another course but not sold separately. So, again, the content is not challenging, and the instructor is monotone and very boring.”
2. “The worst teacher ever! I have taken a lot of online classes, but this is the worst ever. I stopped the course after 4 weeks. It’s very hard to follow and not clearly presented. Just a lot of reading. Don’t recommend for. . .”
3. “The lecturer is not focused on the topics he is explaining and is missing most of the time from the video. I do not recommend this course to anyone. I am very disappointed in this course. I have no idea why it is part of the CAD. . .”

Moving forward, Algorithm 2 further breaks down the review generation algorithm step-by-step.

Algorithm 2: Generating Synthetic Reviews with a Fine-tuned GPT-3 Model

Input: *fine_tuned_model*, *new_prompt*, *max_tokens*, *temperature*, *n*

Output: Synthetic Reviews

BEGIN

//Generate Reviews using Fine-Tuned Model

Answer $\leftarrow$ openai.Completion.create(

model = *fine_tuned_model*,

prompt = *new_prompt*,

max_tokens = *max_tokens*,

temperature = *temperature*,

n = *n*

)

//Extract Generated Reviews from Answer

List SyntheticReviews $\leftarrow$ EMPTY_LIST

FOR each choice IN Answer[‘choices’]

APPEND choice[‘text’].strip() TO SyntheticReviews

END FOR

RETURN SyntheticReviews

END

Stage 4: Synthetic Reviews Evaluation

The purpose of our evaluation is to assess the quality of the synthetic reviews generated by the GPT-3 model. This evaluation focuses on three main criteria: Novelty, Diversity, and Anomaly sentences detected by manual evaluation.

Criteria 1: Novelty

Novelty pertains to the level of uniqueness of the generated review compared to the training corpus. In simpler terms, it evaluates whether the model generates new review or merely replicates the ones from the corpus [21]. We measure the novelty of each generated review R_i using the formula below:

$$Novelty(R_i) = 1 - \max\{\varphi(R_i, C_j)\}_{j=1}^{j=|C|}$$

where C is the review set of the training corpus, and φ is Jaccard similarity function. A novelty score tending to 0 indicates that the generated review closely resembles the training corpus, while a score approaching 1 signifies that the generated review varies considerably from the corpus.

Criteria 2: Diversity

Diversity, on the other hand, assesses the variety of sentences that the model can produce [21]. Given a collection of generated reviews R , we evaluate the diversity of the generated reviews R_i using a formula below:

$$Diversity(R_i) = 1 - \max\{\varphi(R_i, R_j)\}_{j=1}^{j=|R|, j \neq i}$$

where φ is Jaccard similarity function. A diversity score tending towards zero means that the text is similar to other generated texts, while a score tending towards 1 indicates that the text is different from the other generated texts.

Criteria 3: Anomaly sentences

In this study, anomaly sentences are defined as generated text outputs that exhibit abnormal or nonsensical characteristics. These may include but are not limited to:

- Overly repetitive phrases or sentences, for example, “this course is really really really really really really really really”, where a single word is unnecessarily and illogically repeated.
- Sentences that incorporate non-English words or phrases, or sentences that are entirely in a different language.
- Sentences that, while may be grammatically correct, do not make sense in the context of the review or fail to convey a coherent thought.

The identification of Anomaly sentences within the generated text is achieved both manually and using the diversity score data.

4.3. Sentiment Classification

This part of our study involves three crucial stages: Preprocessing, Sentiment Modeling, and Evaluation.

Stage 1: Preprocessing

The preprocessing stage for sentiment classification is different from preprocessing for review generation. This process includes a few essential actions: first, content cleaning where we removed non-English sentences, duplicates, and unnecessary elements from the content, like URLs, user handles, and hashtags that do not significantly contribute to sentiment analysis; second, text normalization, which involved removing special characters, numbers, and multiple spaces to ensure uniformity in the text; and lastly, language processing that further refines the text data by excluding stop words, emojis, and sentences that only consist of a single word.

Stage 2: Sentiment Modeling

In this study, two sentiment classification scenarios were undertaken: one using the original imbalanced dataset, and the other using a balanced dataset supplemented

with synthetic data generated by GPT-3. Both scenarios employed the same testing data, ensuring a consistent basis for comparison.

To begin with, the original imbalanced dataset was systematically split into two primary subsets: training and testing, following an 80:20 ratio. It is important to highlight that the testing dataset preserved the inherent imbalanced nature of the primary dataset. To guarantee a broad representation, this testing dataset was proportioned according to sentiment categories, namely, positive, negative, and neutral. Moreover, to capture the nuances of each category *aspect* variable, the testing data were proportioned to mirror the distribution of these *aspects*. This detailed strategy ensured that the testing data faithfully represented the original dataset, capturing its intrinsic imbalanced distribution across classes. After this, the creation of training and validation datasets for both imbalanced and balanced datasets took place. For the imbalanced dataset, the training data derived from the initial step were further divided into Train and Validation sets at an 80:20 ratio. For the balanced dataset, the procedure entailed two phases: (1) The training data from the first stage were augmented with synthetic data to achieve a balanced state. (2) Upon reaching this balance, the synthetic data were split into Train and Validation sets, again with the 80:20 ratio. A detailed representation of the data splitting and sentiment classification process can be seen in Figure 4.

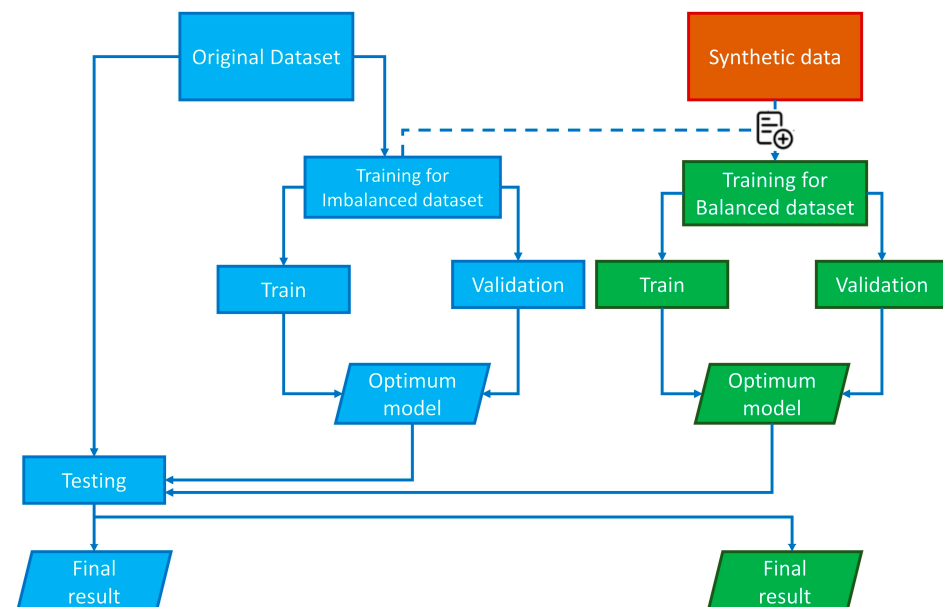


Figure 4. Splitting data and sentiment modeling process.

For both the balanced and imbalanced datasets, we utilized a total of nine models, which spanned both traditional machine learning and deep learning techniques, for sentiment classification. Within traditional machine learning, methods such as the Support Vector Machine, Decision Tree, Naive Bayes, and AdaBoost were used. Conversely, the deep learning domain was explored using architectures like RNN, CNN, and their advanced variants like LSTM, GRU, and BiLSTM. Except for RNN, these deep learning models incorporated GloVe embeddings.

Stage 3: Evaluation

In the assessment of sentiment classification performance, particularly when contending with imbalanced datasets, traditional accuracy can be deceptive. Models might exhibit a misleadingly high accuracy by predominantly predicting the majority class. Therefore, to ensure a more comprehensive and representative evaluation of our models, we opted for Balanced Accuracy and the Macro F1-score as our primary evaluation metrics.

Balanced Accuracy is essentially the arithmetic mean of the recall obtained for each class, capturing the model's effectiveness across all sentiment categories without bias. On

the other hand, the Macro F1-Score, derived from both Macro-Precision and Macro-Recall, averages the precision for every predicted class and the recall for each actual class [41]. This methodological approach guarantees that regardless of the population in the dataset, each sentiment class receives equal weightage. By employing these metrics, we aimed to achieve a nuanced and unbiased insight into the model's performance, especially in the context of diverse and imbalanced sentiment classes.

5. Results

5.1. Synthetic Review Generation

5.1.1. Generated Reviews

In this study, synthetic review generation was carried out by fine-tuning the GPT-3 model using the Davinci base model, which has proven to be more powerful compared to other base models (Curie, Babbage, and Ada). The model was configured with specific generation parameters: “maximum tokens” was set to 20 or 50, and “temperature” was set to 0.9. After the model was fine-tuned under these settings, it was used to generate synthetic reviews. However, in a single generation, it was only capable of producing 128 reviews. Therefore, to obtain the required synthetic reviews, the generation process was repeated multiple times.

At its core, the generated synthetic reviews were employed to balance the training data during the sentiment classification phase. The generation of these reviews was meticulously crafted, mirroring the inherent distribution of the original data. The quantity of these synthetic reviews was determined based on the distribution of the minority classes, namely, negative and neutral sentiments. Furthermore, to encapsulate the true essence and variability of course reviews, the generation was also proportioned according to the *aspect* variable present in the dataset. To anticipate the possibility of getting some low-quality synthetic reviews, we generated more reviews than were needed for training data balancing. In total, 34,186 synthetic reviews were generated, with a breakdown of 16,525 reviews reflecting negative sentiments and 17,661 reviews embodying neutral sentiments. This distribution is visualized in Figure 5a. The number of synthetic reviews based on the *aspect* variable is depicted in Figure 5b. Reflecting the characteristics of the original dataset, most reviews predominantly focused on the *content* aspect for both negative and neutral sentiments.

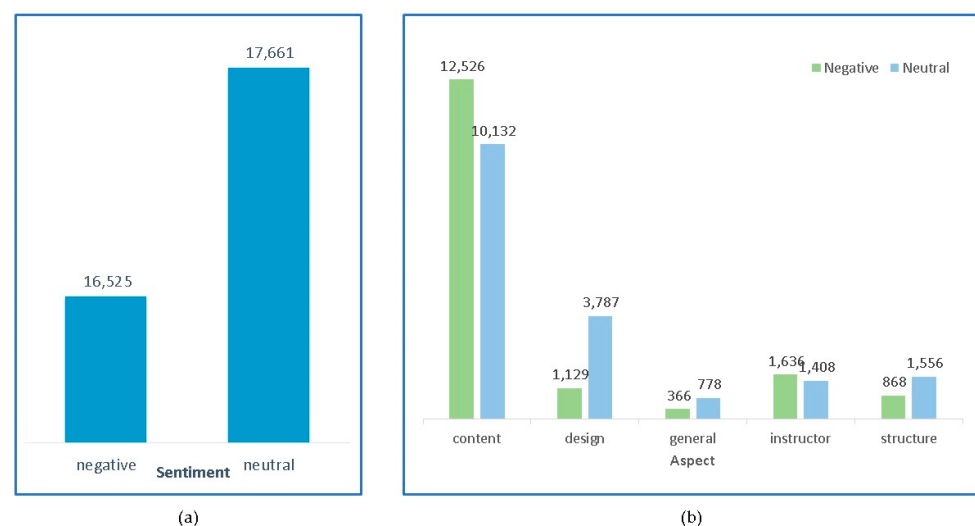


Figure 5. The distribution generated synthetic reviews: (a) based on sentiment and (b) based on aspect.

The distribution of these *aspects* in the synthetic reviews was intentionally modeled to reflect that of the original dataset as shown in Figure 2b. This strategy explains why the

content aspect holds most synthetic reviews in both negative and neutral sentiments, with counts of 12,526 and 10,132, respectively, mirroring its dominance in the original dataset.

This intentional mirroring was performed to ensure the representation of each aspect in the synthetic reviews. This was accomplished by introducing a new prompt for each aspect during the review generation process. Without individual prompts for each aspect, the GPT-3 model could generate reviews randomly, which could potentially lead to an underrepresentation of some aspects. Ensuring representation for all aspects is crucial for creating a diverse set of synthetic reviews that remain relevant to the original data. This method further reinforces the diversity and relevance of the generated reviews, making them an asset for subsequent sentiment analysis tasks.

5.1.2. Evaluation of Generated Reviews

The synthetic reviews generated by GPT-3 were evaluated using three criteria: novelty, diversity, and anomaly. As an illustration, Figure 6 displays the novelty and diversity of the generated synthetic negative reviews.

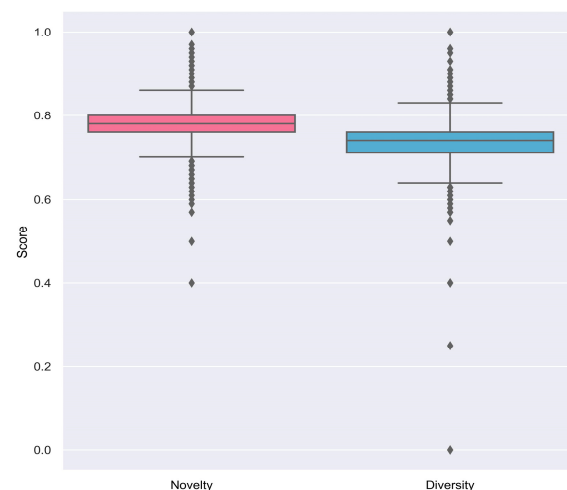


Figure 6. The novelty and diversity score of negative synthetic reviews generated by GPT-3.

The novelty score indicates how novel the synthetic reviews generated by GPT-3 are. A score of 0 implies that the generated review is identical to the reviews present in the training corpus (the original data used as training data during fine-tuning), while a score of 1 denotes that the generated review is entirely novel. Analysis of the results revealed that the minimum novelty score was 0.4 and the maximum was 1. From Figure 6, since no novelty score was 0, this means that all negative reviews generated by GPT-3 are novel and not simply copied from the training corpus. This indicates that GPT-3 successfully generates new reviews that are different from the training corpus.

The diversity score indicates how dissimilar a synthetic review is from the other synthetic reviews. A score of 0 signifies that the review is identical to another review, while a score of 1 suggests that the review is completely different from all other reviews. The analysis showed that the minimum diversity score was 0, and the maximum was 1. This indicates that some generated negative reviews are identical to others, which suggests potential duplication. However, there were also reviews that were entirely distinct from the rest, raising questions about the structure of the sentences within those reviews.

Given these results from the novelty and diversity score evaluations, a manual evaluation was necessary to verify the quality of the generated reviews. This manual evaluation considered the diversity scores, aiming to thoroughly understand and categorize the synthetic reviews. Table 4 presents case examples of anomalous synthetic reviews, detected by manual evaluation, which include nonsensical and non-English sentences. These anomalies are typically associated with reviews exhibiting diversity scores close to 0 and 1.

Table 4. Examples of anomalous sentences detected through manual evaluation.[illegible]

The first two examples (reviews 2332 and 13,376) demonstrate extreme repetition of words, leading to nonsensical sentences. Despite their medium to high novelty scores (indicating that they are considerably different from the reviews in the training corpus), their diversity scores are 0, showing that these reviews are identical to other generated reviews.

In reviews 1441, we still observe extensive repetition. However, the diversity scores have slightly increased. This implies that these reviews have some unique elements, although the overall quality remains poor due to the lack of coherence and meaningful information.

The following two entries (reviews 13,533 and 3737) present non-English sentences, which is unusual as our original training corpus only contained English language reviews. These reviews have high novelty scores (0.95 and 0.96) and similarly high diversity scores, reflecting their distinctiveness both from the original reviews and other synthetic reviews.

Finally, the last three examples (reviews 4570, 13,402, and 16,308) represent a combination of nonsensical phrases, non-spaced words, and undecipherable sequences of characters. Their novelty and diversity scores are at the maximum level of 1, indicating these reviews are entirely different from any existing reviews and from each other. This observation reinforces the need for manual review of the generated data, to filter out such anomalies despite their high novelty and diversity scores.

This examination underlines the importance of manual evaluation in identifying anomalous outputs, which can be missed when relying solely on quantitative measurements, such as novelty and diversity scores.

Figure 7 offers a visual representation of the novelty and diversity scores for negative synthetic reviews once anomalous data have been removed, resulting in what we now refer to as “good-quality synthetic reviews”. Table 5 provides the statistical summary of these scores both for negative and neutral generated reviews. The novelty scores for the good-

quality synthetic negative reviews range from 0.57 to 0.89, with an average of approximately 0.78. These values indicate that these reviews are significantly different from the training corpus, underscoring the model’s success in generating original content. Meanwhile, diversity scores, which measure the uniqueness of each synthetic review compared to others, lie between 0.55 and 0.86, with an average of approximately 0.736. These scores suggest that the good-quality synthetic negative reviews are quite varied, showcasing the model’s ability to produce a diverse set of reviews. Similarly, the good-quality synthetic neutral reviews exhibit a comparable range of novelty and diversity scores.

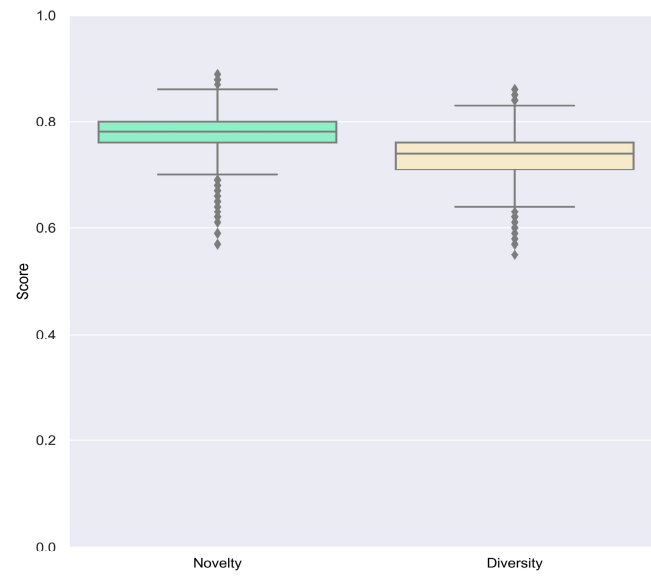


Figure 7. The novelty and diversity scores of good-quality negative synthetic reviews.

Table 5. Descriptive statistics of novelty and diversity scores for good-quality synthetic reviews.

Sentiment	Statistic	Novelty	Diversity
Negative	Min	0.57	0.55
	Max	0.89	0.86
	Average	0.78	0.74
Neutral	Min	0.57	0.54
	Max	0.90	0.88
	Average	0.79	0.74

To better comprehend the impact of the quality control process on the overall dataset, we turn to Figure 8. The figure breaks down the number of reviews classified as “good” and “bad” quality, per sentiment. The data show that out of 16,525 negative sentiment reviews, 163 were classified as bad quality, accounting for approximately 0.99% of the total. The remaining 16,362 reviews (about 99.01%) were deemed to be of good quality. Similarly, for neutral sentiment reviews, out of 17,661 reviews, 330 were labeled as bad quality (about 1.87%), while 17,331 (around 98.13%) were categorized as good quality.

This evaluation highlights the effectiveness of our quality control process in refining the dataset by eliminating anomalies, resulting in good-quality synthetic reviews suitable for downstream sentiment classification tasks.

The good-quality synthetic reviews were subsequently utilized for sentiment classification modeling. For the neutral sentiment, we have 17,331 reviews and for negative sentiment, we have 16,160 reviews as good-quality reviews. From these reviews, we sampled the required number of reviews to balance the training dataset in the sentiment classification phase.

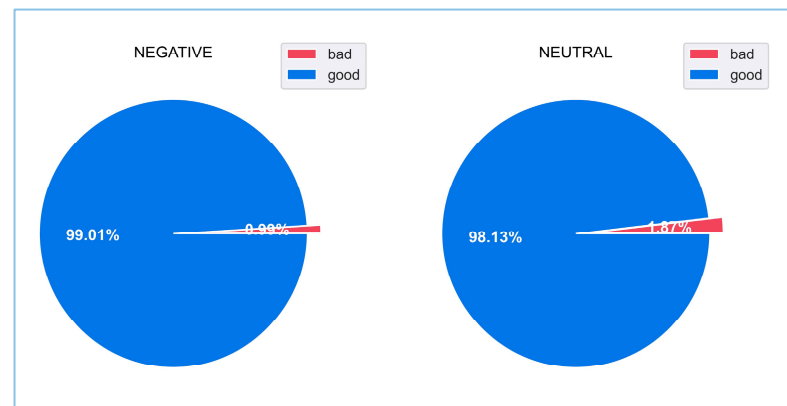


Figure 8. Frequency of good versus bad synthetic reviews for each sentiment.

5.2. Sentiment Classification

The sentiment classification process begins with a meticulous preparation of the data, ensuring that it is ready for the subsequent modeling stages. From identifying and handling imbalanced classes to careful splitting and augmentation, the entire preparation sequence is crucial to the overall success of the classification. The initial dataset comprised a total of 21,937 reviews. Preprocessing, which involved the removal of non-English words and duplicate data, led to a refinement of the dataset, bringing it to a count of 21,726 reviews. This refined dataset was further subjected to additional preprocessing stages and subsequently used for sentiment classification. From this processed data, 17,376 entries were designated as training data, while 4350 were set aside for testing. For the imbalanced dataset, the training data were further partitioned into 13,900 for the train set and 3476 for the validation set.

A distinctive approach was undertaken for the balanced dataset. The initial training data of 17,376 were augmented with 12,765 synthetic reviews for negative sentiment and 13,701 for neutral sentiment, forming a balanced training dataset. This enriched training set was then divided into 13,900 for training and 3476 for validation. Crucial to the integrity of the study, the splitting of the data was carried out proportionally with respect to both sentiment and aspect, thereby ensuring the representation of existing variations within the dataset. The modeling results from both datasets, imbalanced and balanced, were evaluated using the same testing data. Figure 9 provides a visualization of the training and testing data for both the imbalanced and balanced datasets. It illustrates the intricate methodology adopted to prepare the data for the sentiment classification process and offers insight into the proportional distribution across various sentiments and aspects.

For the sentiment classification, both imbalanced and balanced data were trained with nine standard architecture models, encompassing machine learning and deep learning approaches. The deep learning models employed Early Stopping and Model Checkpoint techniques to pinpoint the optimal models during the training process. Balanced Accuracy and Macro F1-score were selected as the most pertinent evaluation metrics, aligning with the unbalanced nature of the data.

The results, as depicted in Table 6 and Figure 10, demonstrate a distinct advantage in using balanced data. All models showcased higher accuracy with the balanced dataset compared to the imbalanced one. A striking observation is the average increase in accuracy of 12.76% across the models when trained on balanced data, highlighting the significance of the balancing approach in improving general classification performance. Among the models, Multinomial Naïve Bayes stands out with the highest accuracy of 75.12%, with the Support Vector Machine realizing the most substantial increase in accuracy at 23.82%, while the lowest accuracy was obtained by the AdaBoost model at 54.08%. A comparison between machine learning and deep learning models reveals that deep learning models consistently exhibited higher and more stable accuracy, above 61%, possibly indicating inherent advantages in handling complex patterns.

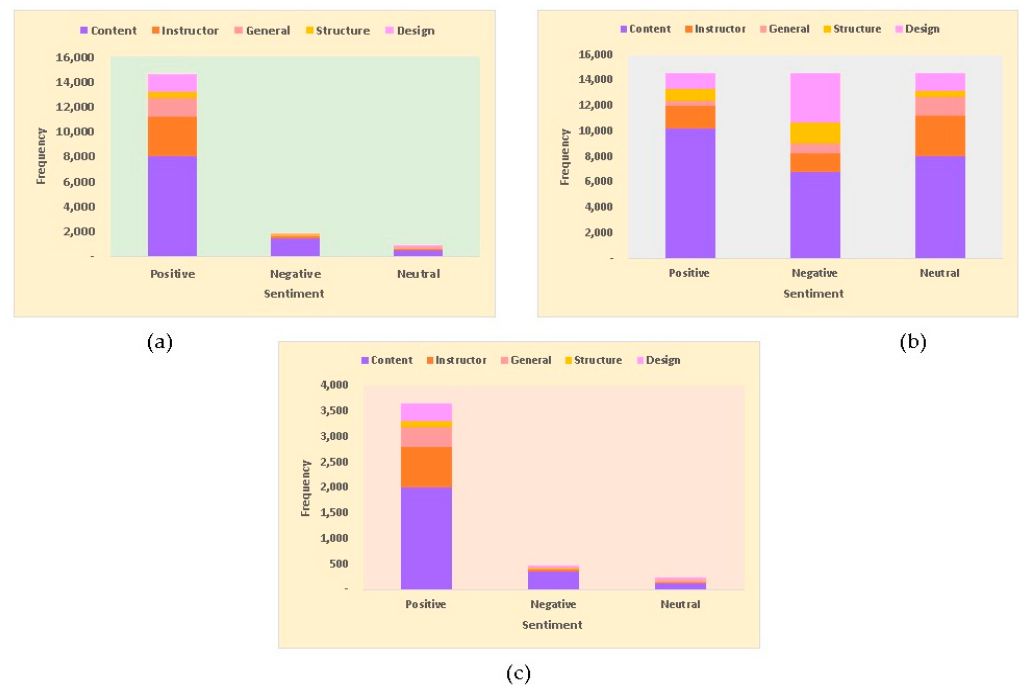


Figure 9. Distribution of dataset for classification: (a) training data of imbalanced dataset, (b) training data of balanced dataset, and (c) testing dataset.

Table 6. Overall sentiment classification results of imbalanced and balanced dataset (in percentage).

Model	Imbalanced Data		Balanced Data		Improvement	
	Accuracy	F1-Score	Accuracy	F1-Score	Accuracy	F1-Score
SVC	42.17	44.64	65.99	62.02	23.82	17.38
Decision Tree	53.83	53.75	57.24	48.87	3.41	−4.88
MultinomialNB	52.12	53.70	75.12	62.82	23.00	9.12
Adaboost	44.71	46.99	54.08	50.96	9.37	3.97
RNN	48.17	50.39	61.46	62.00	13.29	11.61
BiLSTM (Glove)	53.86	54.29	64.12	63.15	10.26	8.86
CNN (Glove)	55.36	53.93	63.16	61.52	7.80	7.59
LSTM (Glove)	53.51	54.07	64.64	63.93	11.13	9.86
GRU (Glove)	52.47	53.48	65.21	64.12	12.74	10.64

The F1-score values in the table offer critical insights into the models' performance, especially considering the imbalanced data. It provides a more nuanced understanding of the models' handling of both False Positives and False Negatives and sensitivity towards the minority class. For instance, the GRU (GloVe) model increased the F1-score from 53.48% to 64.12% with balanced data. Only the Decision Tree showed a decrease, a unique behavior that might require further examination.

Our current study stands in contrast to previous research, which reported accuracy levels of around 30% for machine learning models and 60% for deep learning models using GAN-based synthetic data. Despite differences in methodology, our study's significantly higher performance employing GPT-3-generated synthetic reviews indicates the potential advantages of this approach.

In conclusion, the results substantiate the importance of data balancing, with notable improvements across various models. The unique insights gained through the careful analysis of accuracy and F1-scores provide critical guidance for model selection and refinement. The comparison with previous work highlights the innovative contribution of this study and suggests the need for further comparative research to solidify the understanding of different synthetic review generation techniques, thereby paving the way for more refined and effective models for sentiment classification.

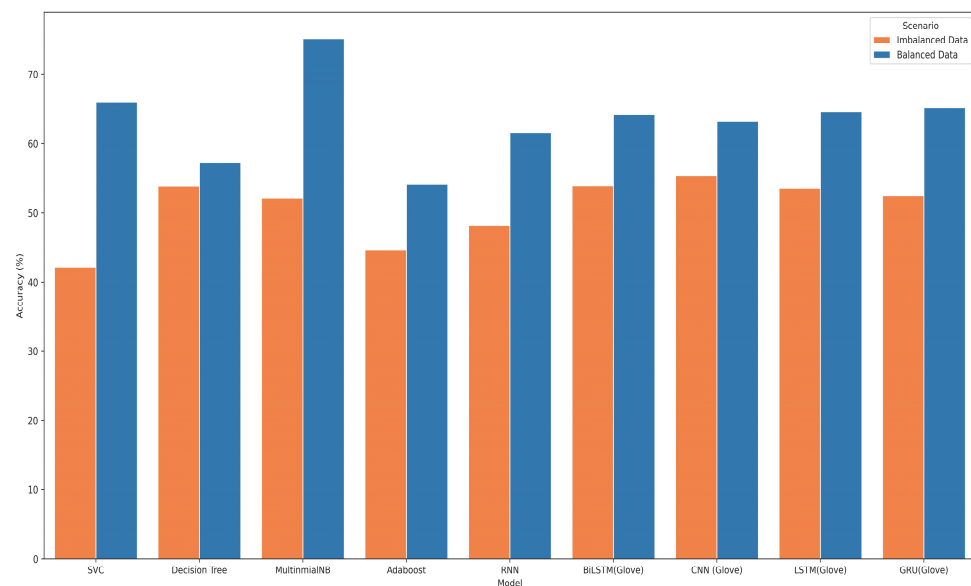


Figure 10. Comparison of accuracy for imbalanced and balanced datasets.

6. Discussion and Future Work

To address our primary research goal of effectively tackling imbalanced sentiment analysis, this study embarked on two pivotal tasks: (1) generating synthetic reviews through the fine-tuning of GPT-3's Davinci base model and (2) employing sentiment classification across nine distinct models on both imbalanced and balanced datasets.

Our findings underscore the remarkable capabilities of GPT-3 in addressing sentiment imbalances, particularly for platforms like Coursera. Through evaluation methods such as novelty and diversity scores, we found evidence that GPT-3 produces genuinely novel synthetic text, it is not just regurgitating content from the training corpus. Moreover, while the generated text is diverse, it remains contextually appropriate for reviews of an online learning platform like Coursera. Our manual checks further solidified these observations: 99% of negative reviews and 98% of positive reviews were of top-notch quality. This not only confirms GPT-3's prowess in generating high-caliber content but also spotlights its unparalleled proficiency in crafting text that echoes human expression. The results firmly place GPT-3 at the forefront of synthetic text generation, offering a bright prospect for future endeavors and applications.

In our classification efforts, the utilization of GPT-3-generated synthetic reviews within the primary dataset has yielded marked improvements. When trained on the balanced dataset, all the models demonstrated an increase in performance compared to the imbalanced data, with an average improvement in accuracy of 12.76%. The obtained accuracies ranged from 57.24% to 75.12%, standing in contrast to the previous study on Coursera that employed GAN-based methods such as CatGAN and SentiGAN for synthetic text generation. While both their study and ours utilized the Coursera review dataset, discrepancies existed in dataset size, chosen methodologies, and preprocessing techniques. Nevertheless, our approach distinctly outperforms, boasting accuracy rates that are significantly higher than the recorded 30% to 60%. This enhancement in classification underscores the potential advantages of using GPT-3-generated synthetic reviews, manifesting not only in the improved accuracy but also in the more robust and consistent performance across various models, regardless of whether machine learning or deep learning methods are applied. However, it must be acknowledged that these results cannot yet be fairly compared, given the differences in dataset size and methodology. Therefore, a future comparison employing the same dataset and experimental setup would be necessary to achieve a more fair and optimal assessment of the differing techniques.

This research's implications suggest that adding GPT-3-generated synthetic reviews to datasets provides a potent solution to the challenges of imbalanced sentiment analysis. However, it is not without limitations. GPT-3 is a non-open-source language model; its paid-access nature might be a deterrent for some researchers. Additionally, the reliance on manual evaluations in this study for gauging the quality of generated reviews, though meticulous, was time-intensive, underscoring the necessity for a swifter automated evaluation process.

Looking ahead, a significant direction for future research is to explore open-source alternatives to GPT-3, such as GPT-Neo and OPT. Transitioning to these alternatives could democratize the technology, making it more accessible to a broader range of researchers and innovators. Such a move would also likely drive further advancements in synthetic data generation. Another pivotal area of exploration is the development of automated techniques for evaluating synthetic reviews. Incorporating such techniques could streamline the evaluation process, greatly enhancing both efficiency and scalability. Furthermore, applying our method to different datasets is essential. This would provide deeper insights into the robustness of our approach and its applicability across diverse contexts, ensuring that the benefits of our method are not limited to specific types of data or subject areas.

In conclusion, while our study has unveiled promising pathways for addressing the class imbalance in sentiment analysis, the journey ahead beckons deeper explorations, optimizations, and a drive to make the process more accessible and efficient.

Author Contributions: Conceptualization, H.-S.Y.; Methodology, C.S.; Software, C.S.; Validation, H.-S.Y.; Formal analysis, C.S.; Investigation, C.S.; Resources, H.-S.Y.; Data curation, C.S.; Writing—original draft, C.S.; Writing—review & editing, H.-S.Y.; Visualization, C.S.; Supervision, H.-S.Y.; Project administration, H.-S.Y.; Funding acquisition, H.-S.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This work is partially supported by the Korea Agency for Infrastructure Technology Advancement (KAIA) grant funded by the Ministry of Land, Infrastructure and Transport (Grant RS-2022-00143782).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data utilized in this study is not publicly available and can be accessed only upon request to Kastrati et al. [16].

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kanojia, D.; Joshi, A. Applications and Challenges of Sentiment Analysis in Real-Life Scenarios. *arXiv* **2023**. [CrossRef]
2. Abiola, O.; Abayomi-Alli, A.; Tale, O.A.; Misra, S.; Abayomi-Alli, O. Sentiment Analysis of COVID-19 Tweets from Selected Hashtags in Nigeria Using VADER and Text Blob Analyser. *J. Electr. Syst. Inf. Technol.* **2023**, *10*, 5. [CrossRef]
3. Hananto, A.L.; Nardilasari, A.P.; Fauzi, A.; Hananto, A.; Priyatna, B.; Rahman, A.Y. Best Algorithm in Sentiment Analysis of Presidential Election in Indonesia on Twitter. *Int. J. Intell. Syst. Appl. Eng.* **2023**, *11*, 473–481.
4. Bonetti, A.; Martínez-Sober, M.; Torres, J.C.; Vega, J.M.; Pellerin, S.; Vila-Francés, J. Comparison between Machine Learning and Deep Learning Approaches for the Detection of Toxic Comments on Social Networks. *Appl. Sci.* **2023**, *13*, 6038. [CrossRef]
5. Muhammad, S.H.; Abdulmumin, I.; Yimam, S.M.; Adelani, D.I.; Ahmad, I.S.; Ousidhoum, N.; Ayele, A.; Mohammad, S.M.; Beloucif, M.; Ruder, S. SemEval-2023 Task 12: Sentiment Analysis for African Languages (AfriSenti-SemEval). *arXiv* **2023**, arXiv:2304.06845.
6. Hartmann, J.; Heitmann, M.; Siebert, C.; Schamp, C. More than a Feeling: Accuracy and Application of Sentiment Analysis. *Int. J. Res. Mark.* **2023**, *40*, 75–87. [CrossRef]
7. Tan, K.L.; Lee, C.P.; Lim, K.M. A Survey of Sentiment Analysis: Approaches, Datasets, and Future Research. *Appl. Sci.* **2023**, *13*, 4550. [CrossRef]
8. Bordoloi, M.; Biswas, S.K. Sentiment Analysis: A Survey on Design Framework, Applications and Future Scopes. *Artif. Intell. Rev.* **2023**, *20*, 1–56. [CrossRef]

9. Singh, S.; Kumar, P. Sentiment Analysis of Twitter Data: A Review. In Proceedings of the 2023 2nd International Conference for Innovation in Technology, INOCON 2023, Bangalore, India, 3–5 March 2023; Institute of Electrical and Electronics Engineers Inc.: Piscataway, NJ, USA, 2023.
10. Su, B.; Peng, J. Sentiment Analysis of Comment Texts on Online Courses Based on Hierarchical Attention Mechanism. *Appl. Sci.* **2023**, *13*, 4204. [CrossRef]
11. Rajat, R.; Jaroli, P.; Kumar, N.; Kaushal, R.K. A Sentiment Analysis of Amazon Review Data Using Machine Learning Model. In Proceedings of the CITISIA 2021—IEEE Conference on Innovative Technologies in Intelligent System and Industrial Application, Proceedings, Sydney, Australia, 24–26 November 2021; Institute of Electrical and Electronics Engineers Inc.: Piscataway, NJ, USA, 2021.
12. Thakkar, G.; Preradovic, N.M.; Tadić, M. Croatian Film Review Dataset (Cro-FiReDa): A Sentiment Annotated Dataset of Film Reviews. In *Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023)*; Association for Computational Linguistics: Dubrovnik, Croatia, 2023; pp. 25–31.
13. Wen, Y.; Liang, Y.; Zhu, X. Sentiment Analysis of Hotel Online Reviews Using the BERT Model and ERNIE Model—Data from China. *PLoS ONE* **2023**, *18*, e0275382. [CrossRef]
14. Sasikala, P.; Mary Immaculate Sheela, L. Sentiment Analysis of Online Product Reviews Using DLMNN and Future Prediction of Online Product Using IANFIS. *J. Big Data* **2020**, *7*, 33. [CrossRef]
15. Iqbal, A.; Amin, R.; Iqbal, J.; Alroobaea, R.; Binmahfoudh, A.; Hussain, M. Sentiment Analysis of Consumer Reviews Using Deep Learning. *Sustainability* **2022**, *14*, 10844. [CrossRef]
16. Kastrati, Z.; Arifaj, B.; Lubishtani, A.; Gashi, F.; Nishliu, E. Aspect-Based Opinion Mining of Students' Reviews on Online Courses. In Proceedings of the 2020 6th International Conference on Computing and Artificial Intelligence, Tianjin, China, 23–26 April 2020; Association for Computing Machinery: New York, NY, USA, 2020; pp. 510–514.
17. Imran, A.S.; Yang, R.; Kastrati, Z.; Daudpota, S.M.; Shaikh, S. The Impact of Synthetic Text Generation for Sentiment Analysis Using GAN Based Models. *Egypt. Inform. J.* **2022**, *23*, 547–557. [CrossRef]
18. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models Are Few-Shot Learners. *arXiv* **2020**. [CrossRef]
19. Fatima, N.; Imran, A.S.; Kastrati, Z.; Daudpota, S.M.; Soomro, A. A Systematic Literature Review on Text Generation Using Deep Neural Network Models. *IEEE Access* **2022**, *10*, 53490–53503. [CrossRef]
20. Iqbal, T.; Qureshi, S. The Survey: Text Generation Models in Deep Learning. *J. King Saud. Univ. Comput. Inf. Sci.* **2022**, *34*, 2515–2528. [CrossRef]
21. Wang, K.; Wan, X. SentiGAN: Generating Sentimental Texts via Mixture Adversarial Networks. In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, Stockholm Sweden, 13–19 July 2018; pp. 4446–4452.
22. Liu, Z.; Wang, J.; Liang, Z. CatGAN: Category-Aware Generative Adversarial Networks with Hierarchical Evolutionary Learning for Category Text Generation. *Proc. AAAI Conf. Artif. Intell.* **2020**, *34*, 8425–8432. [CrossRef]
23. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need 2023. Available online: <https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf> (accessed on 31 July 2023).
24. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language Models Are Unsupervised Multitask Learners. *OpenAI Blog* **2019**, *1*, 9.
25. Anaby-Tavor, A.; Carmeli, B.; Goldbraich, E.; Kantor, A.; Kour, G.; Shlomov, S.; Tepper, N.; Zwerdling, N. Not Enough Data? Deep Learning to the Rescue! *arXiv* **2019**. [CrossRef]
26. Ma, C.; Zhang, S.; Shen, G.; Deng, Z. Switch-GPT: An Effective Method for Constrained Text Generation under Few-Shot Settings (Student Abstract). *Proc. AAAI Conf. Artif. Intell.* **2022**, *36*, 13011–13012. [CrossRef]
27. Xu, J.H.; Shinden, K.; Kato, M.P. Table Caption Generation in Scholarly Documents Leveraging Pre-Trained Language Models. In Proceedings of the 2021 IEEE 10th Global Conference on Consumer Electronics (GCCE), Kyoto, Japan, 12–15 October 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 963–966.
28. Bayer, M.; Kaufhold, M.-A.; Buchhold, B.; Keller, M.; Dallmeyer, J.; Reuter, C. Data Augmentation in Natural Language Processing: A Novel Text Generation Approach for Long and Short Text Classifiers. *Int. J. Mach. Learn. Cybern.* **2023**, *14*, 135–150. [CrossRef] [PubMed]
29. Zhong, Q.; Ding, L.; Liu, J.; Du, B.; Tao, D. Can ChatGPT Understand Too? A Comparative Study on ChatGPT and Fine-Tuned BERT. *arXiv* **2023**. [CrossRef]
30. Obiedat, R.; Qaddoura, R.; Al-Zoubi, A.M.; Al-Qaisi, L.; Harfoushi, O.; Alrefai, M.; Faris, H. Sentiment Analysis of Customers' Reviews Using a Hybrid Evolutionary SVM-Based Approach in an Imbalanced Data Distribution. *IEEE Access* **2022**, *10*, 22260–22273. [CrossRef]
31. Wen, H.; Zhao, J. *Sentiment Analysis Model of Imbalanced Comment Texts Based on BiLSTM*. In Review: 2023. Available online: <https://www.researchsquare.com/article/rs-2434519/v1> (accessed on 31 July 2023).
32. Tan, K.L.; Lee, C.P.; Lim, K.M. RoBERTa-GRU: A Hybrid Deep Learning Model for Enhanced Sentiment Analysis. *Appl. Sci.* **2023**, *13*, 3915. [CrossRef]
33. Wu, J.-L.; Huang, S. Application of Generative Adversarial Networks and Shapley Algorithm Based on Easy Data Augmentation for Imbalanced Text Data. *Appl. Sci.* **2022**, *12*, 10964. [CrossRef]

34. Almuayqil, S.N.; Humayun, M.; Jhanjhi, N.Z.; Almufareh, M.F.; Khan, N.A. Enhancing Sentiment Analysis via Random Majority Under-Sampling with Reduced Time Complexity for Classifying Tweet Reviews. *Electronics* **2022**, *11*, 3624. [CrossRef]
35. Ghosh, K.; Banerjee, A.; Chatterjee, S.; Sen, S. Imbalanced Twitter Sentiment Analysis Using Minority Oversampling. In Proceedings of the 2019 IEEE 10th International Conference on Awareness Science and Technology (iCAST), Morioka, Japan, 23–25 October 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1–5.
36. Department of Computer Science; Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, India; George, S.; Srividhya, V. Performance Evaluation of Sentiment Analysis on Balanced and Imbalanced Dataset Using Ensemble Approach. *Indian J. Sci. Technol.* **2022**, *15*, 790–797. [CrossRef]
37. Cai, T.; Zhang, X. Imbalanced Text Sentiment Classification Based on Multi-Channel BLTCN-BLSTM Self-Attention. *Sensors* **2023**, *23*, 2257. [CrossRef]
38. Habbat, N.; Nouri, H.; Anoun, H.; Hassouni, L. Using AraGPT and Ensemble Deep Learning Model for Sentiment Analysis on Arabic Imbalanced Dataset. *ITM Web Conf.* **2023**, *52*, 02008. [CrossRef]
39. Ekinci, E. Classification of Imbalanced Offensive Dataset—Sentence Generation for Minority Class with LSTM. *Sak. Univ. J. Comput. Inf. Sci.* **2022**, *5*, 121–133. [CrossRef]
40. Fine-Tuning. Available online: <https://platform.openai.com/docs/guides/fine-tuning> (accessed on 1 June 2023).
41. Grandini, M.; Bagli, E.; Visani, G. Metrics for Multi-Class Classification: An Overview. *arXiv* **2020**, arXiv:2008.05756.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.