Egyptian Informatics Journal 23 (2022) 547–557

Contents lists available at ScienceDirect

Egyptian Informatics Journal

journal homepage: www.sciencedirect.com

The impact of synthetic text generation for sentiment analysis using GAN based models

Ali Shariq Imran [a,*], Ru Yang [a], Zenun Kastrati [b], Sher Muhammad Daudpota [c], Sarang Shaikh [d]

[a] Department of Computer Science (IDI), Norwegian University of Science & Technology (NTNU), 2815 Gjøvik, Norway
[b] Department of Informatics, Linnaeus University, 35195 Växjö, Sweden
[c] Department of Computer Science, Sukkur IBA University, Sukkur 65200, Pakistan
[d] Department of Information Security and Communication Technology (IIK), Norwegian University of Science & Technology (NTNU), 2815 Gjøvik, Norway

ARTICLE INFO

Article history:
Received 15 March 2022
Revised 14 May 2022
Accepted 24 May 2022
Available online 3 June 2022

Keywords:
Text generation
Sentiment analysis
SentiGAN
CatGAN
Deep learning
Language modeling
Machine learning
GANs
Generative adversarial networks
Data imbalance

ABSTRACT

Data imbalance in datasets is a common issue where the number of instances in one or more categories far exceeds the others, so is the case with the educational domain. Collecting feedback on a course on a large scale and the lack of publicly available datasets in this domain limits models' performance, especially for deep neural network based models which are data hungry. A model trained on such an imbalanced dataset would naturally favor the majority class. However, the minority class could be critical for decision-making in prediction systems, and therefore it is usually desirable to train a model with equally high class-level accuracy. This paper addresses the data imbalance issue for the sentiment analysis of users' opinions task on two educational feedback datasets utilizing synthetic text generation deep learning models. Two state-of-the-art text generation GAN models namely CatGAN and SentiGAN, are employed for synthesizing text used to balance the highly imbalanced datasets in this study. Particular emphasis is given to the diversity of synthetically generated samples for populating minority classes. Experimental results on highly imbalanced datasets show significant improvement in models' performance on CR23K and CR100K after balancing with synthetic data for the sentiment classification task.
© 2022 THE AUTHORS. Published by Elsevier BV on behalf of Faculty of Computers and Artificial Intelligence, Cairo University. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

Many universities throughout the globe have pivoted to online courses from brick-and-mortar settings due to the COVID-19 pandemic. This smooth transition, though not common everywhere globally, is made possible thanks to the recent technological development and adaption of various tools including different Learning Management Systems such as open source Moodle, and Teams and zoom for online learning, both by students and the teachers. Coupled with the popularity of massive open online courses (MOOCs), many students in developing countries and remote areas could register and access courses offered by top universities from the comfort of their homes. The success of distance and blended education relies heavily on students' feedback.

Feedback assessment is an essential element of any quality enhancement cell within an institute. It allows the faculty and teaching staff to reflect on the teaching and other course aspects and provides students with an opportunity to speak up. Students' feedback is often unstructured, allowing them to express their thoughts to open-ended questions related to various aspects of teaching, course, and content. The institutes examine the feedback provided by the students to alter and improve upon the process, teaching, courses, tools and platforms used to deliver the lectures and educational resources. For instance, the authors in [1] indicated that students' feedback helped to implement the co-creation process.

It is impractical to monitor and analyze the textual feedback and open-ended responses obtained from students manually over a longer period. Moreover, for a large student representation in an institute, it becomes a tedious and time-consuming task [2]. Therefore, processing students' responses requires employing

* Corresponding author.
E-mail addresses: ali.imran@ntnu.no (A.S. Imran), ru.yang@ntnu.no (R. Yang), zenun.kastrati@lnu.se (Z. Kastrati), sher@iba-suk.edu.pk (S.M. Daudpota), sarang.shaikh@ntnu.no (S. Shaikh).

Peer review under responsibility of Faculty of Computers and Information, Cairo University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.eij.2022.05.006
1110-8665/© 2022 THE AUTHORS. Published by Elsevier BV on behalf of Faculty of Computers and Artificial Intelligence, Cairo University.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

state-of-the-art automated text analysis techniques that fall under opinion mining and text classification categories, i.e., to perform aspect extraction and sentiment analysis on students' responses via machine learning tasks. Although this is a widely studied domain in the past decade and has achieved remarkable results [3], the lack of enough and an equal number of samples in each class category often limits models' performance, particularly for deep learning-based techniques, which are data-hungry. Unlike other domains such as movies [4], social media platform analysis [5,6] and e-commerce [7], the eLearning domain still is in infancy state when it comes to utilizing sentiment analysis techniques. It is largely attributed to the data imbalance and scarcity of publicly available datasets in the education domain, as is advocated by the authors in [1].

The impact of uncommonly seen samples is often more significant than commonly seen ones as they normally contribute towards suggestions and recommendations. The model trained with an imbalanced dataset often has an abundance of training samples from majority classes, resulting in poor prediction and classification performance for minority classes due to inadequate training samples. Researchers address this issue by applying data balancing techniques. The idea is to produce similar samples belonging to a class with fewer examples and bring it at par with the rest of the classes. Resultantly, deep learning models will have enough training samples to attain a higher classification accuracy. Traditionally, synthetic minority oversampling technique (SMOTE) and adaptive synthetic (AdaSYN) are widely employed to generate new samples. These, however, are not suitable for generating grammatically as well as semantically meaningful text data.

With the recent advent of deep learning and the success of long short-term memory (LSTM) in processing sequential data, text generation techniques are gaining popularity. Table 1 highlights some of the recent state-of-the-art techniques to generate synthetic text. This study utilizes SentiGAN and CatGAN to address the data-imbalance issues on sentiment polarity assessment tasks in the education domain. These models are chosen due to their task-specific nature of generating diversified high-quality sentimental texts for different polarity categories. The significant contributions of this article are as follows:

- Investigation of the impact of synthetic text on the sentiment polarity assessment task in education domain for highly imbalanced datasets.
- Training and utilization of two state-of-the-art deep learning generative adversarial network models on two large students' feedback datasets for generating diverse synthetic samples.
- Validation of SentiGAN and CatGAN generated text via NLL and BLUE performance evaluation metrics for diversity and text generation quality.
- Detailed comparison of ten conventional and deep learning models on two sentiment classification settings including original and balanced datasets.

The rest of the paper is organised as below. Section 2 presents relevant literature work focused on text generation and sentiment analysis. Section 3 depicts methodology along with information regarding datasets, preprocessing techniques, text generation model architectures and evaluation metrics. Section 4 discusses the major results from selected text generation models and sentiment analysis experiments on original as well as balanced datasets. Finally, Section 5 shows conclusion and future work.

2. Related works

This section highlights some of the key research in recent years on the topic of sentiment analysis in education domain and text generation models.

2.1. Text generation

In the recent years, deep neural networks including RNN [18,19], LSTM[20], and CNN [21,22] have shown remarkable ability to generate synthetic text.

Vu & Le in [18] proposed the use of text topic along with input text to generate text through RNN. They achieved an accuracy improvement of 23% on Vietnamese news dataset in comparison to baseline model without text generation. Masum et al. in [19] used bi-directional RNN to generate news head lines in Bengali language. They trained the network on Bengali newspapers' text and generated synthetic news headlines. Similarly, Li & Zhang [20] conducted an exploratory research to compare text generation performance of LSTM with its different variants focusing on long text sequences. They conclude that performance of LSTM is better than many of its variants on different evaluation metrics including BERT, BLEURT etc. Akhtar et al. [23] used an optimized version of GRU to generate Bengla sentences.

Although, RNN, LSTM and CNN have been used extensively in the literature for text generation, however the generation quality decreases for long sentences' dependencies. In recent years, text generation has shifted from these traditional approaches to more advanced approached based on attention mechanism [24] including transformers [25,26], BERT [27] and GPTs[28]. Readers are advised to refer to the systematic mapping study on text generation techniques [29] for more information.

2.2. GAN-based text generation models

This section provides a brief overview of the available text generation models in the literature, especially those relying on GAN models.

Yu et al. in [8] proposed the SeqGAN model to generate text sequences by using GANs. The authors focused on one of the limitations of GANs which is the generating sequences of discrete tokens. Several set of experiments consisted of synthetic data as well as evaluation with the real data are performed. The authors evaluated the performance of SeqGAN by generating synthetic data in three real-world domains including Chinese poems, Obama political speech, and music.

Guo et al. in [9] proposed the LeakGAN model to generate long texts. In general, the text generated by GANs is always evaluated as real or fake once it is generated but the authors in this study proposed to use intermediate evaluation during the text generation process. The authors claimed that the proposed approach is good for generating long texts. Two evaluation metrics namely, NLL and BLEU scores are used to evaluate the performance of LeakGAN. For the synthetic data, the study compared the performance of LeakGAN using NLL metric with MLE, SeqGAN [8] and RankGAN [30]. Among all three models, the LeakGAN outperformed. Furthermore, the authors also compared the performance using BLEU scores metric with same above three models using EMNLP2017 WMT[1] dataset. Again, the LeakGAN model outperformed all other models.

[1] http://statmt.org/wmt17/translation-task.html.

Table 1
10 GAN-based text generation models.

Type	Name	Author	Year	Ref.
General	SeqGAN	Yu et al.	2017	[8]
	LeakGAN	Guo et al.	2018	[9]
	MaliGAN	Che et al.	2017	[10]
	JSDGAN	Li et al.	2019	[11]
	RelGAN	Nie et al.	2018	[12]
	DPGAN	Xu et al.	2018	[13]
	DGSAN	Montahaei et al.	2021	[14]
	CoT	Lu et al.	2019	[15]
Category	SentiGAN	Wang et al.	2018	[16]
	CatGAN	Liu et al.	2020	[17]

Che et al. proposed the MaliGAN text generation model in their study [10]. The model aimed to enhance the instability of GANs in back propagation during the training process. The authors evaluated the performance of proposed model using the BLEU scores and PPL evaluation metrics on poetry generation task. An evaluation performance of MaliGAN with MLE and SeqGAN text generation models is performed, and findings showed that MaliGan outperformed both of them.

Li et al. in [11] proposed the JSDGAN model for the text generation. The authors compared the performance of proposed model with MLE, SeqGAN, RankGAN and LeakGAN models using NLL and BLEU score metrics. The findings reported in the study showed that for the synthetic data the JSDGAN model outperformed among all models with overall NLL score of 5.50. Also, the authors evaluated the proposed model using BLEU scores with above text generation models on Chinese poems, COCO image captions and Obama political speech datasets.

Sarang et al. in [31] addressed the issue of data imbalance using text sequence generation algorithms. They used an LSTM based text generation model along with the GPT-2 for generating synthetic data. Three highly imbalanced datasets from different domains were studied in their work. A 17% improvement on the results for similar deep learning models were observed in their case. Another conclusion that the authors drew from their experiments is that LSTM performs well in generating synthetic text at sentence level, whereas at paragraph or document level, the performance of GPT-2 is much better.

This study is an extension of the work presented in [31]. Instead of using generic text generation models, in this study we utilized pre-trained category GANs for generating sentiment-specific synthetic samples in the education domain.

2.3. Polarity assessment in educational settings

With the rapid emergence of educational resources and an abundance of digital tools for delivering online courses over the last decade, fueled by the recent COVID-19 outbreak, many researchers proposed solutions to effectively incorporate automatic assessment models for quality enhancement cells within an institute. These solutions fall in two categories based upon the approaches: lexicon-based sentiment analysis and those employing machine/deep learning techniques.

Sindhu et al. [32], for example, developed a two layers LSTM model for analyzing sentiments from students' reviews. The authors implemented two levels of categorization. First is identifying aspects such as teacher, course, content, and the second is the classification of reviews belonging to these aspects into positive, negative, and neutral categories. They reported 93% sentiment classification accuracy on the manually labeled students' reviews. Kastrati et al. compared the conventional machine learning and deep learning algorithms on sentiment analysis based on 21,940 students' reviews scrapped from the Coursera MOOC platform [33]. They performed experiments implementing Naive Bayes, Decision Tree, SVM, and Boosting. They further developed the 1D-CNN model for extracting aspects and to predicts sentiments on them. As per the authors, 1D-CNN achieved 88.2% F1-score, better than the rest of the classifiers. However, they claim that the conventional models performed better at aspect extraction.

In another study, Anna et al. [34] conducted a survey-based questionnaire containing multiple open ended questions. Out of 204 feedback, 161 were classified as positive, while 42 were classified as negative. The authors used *K*-nearest neighbor and Naive Bayes to predict sentiment from the students' reviews. The authors compared the results with the Recursive Neural Tensor Network (RNTN) [35] method showing that despite better precision, RNN has poor recall and accuracy. Katragadda et al. [36] employed several supervised algorithms and a DNN model on 30,000 feedback reviews into positive, negative, and neutral classes. Naive Bayes showed 50% accuracy on their dataset, SVM reported 60.8%, and their proposed DNN model achieved 88.2% classification accuracy.

Lwin et al. [37] in their study utilized a dataset containing text reviews having rating scores along with an open-ended textual feedback question. The authors implemented a K-means clustering algorithm to pre-label the feedback data into five rating scores, i.e., worse, bad, neutral, good, and excellent. The last question was categorized into positive and negative sentiment polarity. They applied conventional machine learning algorithms, including logistic regression, multilayer perceptron, SVM, and Random forest, to train and classify sentiments. Sadriu et al. performed sentiment classification of students' feedback on the Albanian language using Monkey learn API and Textblob [38]. They reported 72% accuracy on 114 descriptive feedback.

The authors in [3] generated two corpus: SentiTEXT and eduSERE. The former contains positive and negative polarity statements, while the latter was categorized into learning-centered emotions like engaged, excited, bored, and frustrated. The dataset was obtained from multiple sources, including YouTube videos and other educational platforms. In their work, the authors maintained an emotional dictionary by building connections between words and their polarity. This dictionary was used as a lexical resource to generate the ground truth based on word frequency count. The author then used BERT and EvoMSA models for the classification task achieving 93%-94% on SentiTEXT and 83%-84% on EduSERE, respectively.

Despite numerous research in this domain, there is no benchmark dataset to report and test the best performing model for sentiment classification on students' feedback. Furthermore, Kastrati et al. [39] advocated that the high accuracy reported in most research on sentiment classification tasks in the education domain is on their own (private) datasets, favoring the majority classes. And most of these models fail when applied to a real-world, highly imbalanced dataset on a large scale. They also presented a

systematic mapping study of sentiment analysis of students' feedback with NLP and Deep Learning. They stated that the sentiment analysis is still infancy development stage, especially when it comes to the lack of structured, manually-labeled publicly available datasets. Most datasets used in recent studies favour neutral or positive classes. Moreover, the authors also indicated that structured datasets, standard solutions, and sentiment expression and detection need further attention.

3. Methodology

This section provides an overall methodology of research study. In particular, it presents information about datasets, preprocessing techniques, models and algorithms used for the experiments as well as the evaluation metrics. A high-level conceptual view of the methodology is given in Fig. 1.

3.1. Dataset

To validate the benefit of text generation for sentiment analysis, we used two datasets, namely CR23K [33] and CR100K.[2] Both datasets are from education domain and contain manually labeled course reviews for three classes (e.g. positive, negative & neutral). Both datasets are highly imbalanced with maximum number of reviews for the positive label. The first dataset, CR23K, contains 21,940 course reviews collected from the online learning platform Coursera. The dataset is in English language and is manually labeled with three sentiment labels. More specifically, 84.2% of the reviews are labeled as positive, 10.6% as negative, and 5.2% of them are labeled as neutral. The second dataset, CR100K, includes 107,016 reviews obtained from Kaggle. The dataset contains rating scores ranging from 1 to 5 for each of the review as a sentiment label. A conversion strategy is used to convert ratings into sentiment labels. Ratings 4 and 5 are assigned a positive sentiment, rating 3 a neutral sentiment, and ratings less than 3 are labeled as negative. After conversion, 90.9% of the reviews are labeled as positive, 4.7% as negative, and 4.4% of them are labeled as neutral. The sentiment distribution shows the highly imbalanced nature of both datasets with major inclined towards positive sentiment label.

3.2. Preprocessing

We applied few preprocessing steps to the dataset before feeding it to the classifiers. Initially, we removed all the non-English texts from both datasets. Next, we removed all the stop words, converted text to lower case and then performed lemmatization of the text. The advantage of pre-processing is depicted in Fig. 2, where the part (a) and (b) show the length distribution of the texts before and after preprocessing, respectively. Although, after preprocessing the length is shortened but the overall datasets are now in normal distribution in terms of text length. Furthermore, after applying the above preprocessing steps the CR100K dataset reduced up to 25%. However, the CR23K dataset remains same after the preprocessing steps.

3.3. Architecture of text generation models

As the focus of this study is to investigate the impact of text generation on sentiment analysis on datasets comprising multiple categories/labels, we therefore employed two text generation models, namely SentiGAN and CatGAN.

SentiGAN is proposed by Wang et al. in [16]. The model consists of multiple LSTM-based generators and a classifier. The classifier predicts synthetic text from the real ones. In SentiGAN, multiple generators are used to generate text of each category in an unsupervised manner. These generators can work independently and do not rely on each other. Similar to the research study [40], the researchers used the sequence generation process as the sequence decision process as well. They applied random initialization strategy to parameters of each generator model and then Monte Carlo search is used to search for the appropriate parameters. Then, the classifier is used to evaluate the generated text, which contributes towards the learning of the generator. The main advantage of this model is the use of penalty mechanism, which force each generator to generate sentiment-specific text along with its polarity. The basic model structure of SentiGAN is shown in Fig. 3.

CatGAN is proposed by Liu et al. in [17]. The architecture of the model is shown in Fig. 4. The model consists of two main parts. One is the category aware model. It finds the error between the generated text and the original for each category to reduce the error. The generator is based on relational memory core to generate text with a specific category. The second part is the hierarchical evolutionary algorithm. It is used for training the model and for classifying the generated samples from the original ones for each category. It further tries to maintain the quality and generated text diversity of the CatGAN.

3.4. Evaluation metrics – text generation

This section provides various evaluation metrics used for evaluating the quality of the generated text in reference to the original text.

BLEU, which stands for Bilingual Evaluation Understudy, is a metric for evaluating the model with multiple correct output results [41]. For example, comparing overlapping degree of generated text and original text. The higher the overlapping degree, the higher the quality of generated text. In real life, usually $N = 1 - 4$is used where N refers to n-gram in the text. Then, we calculate the weighted average score of all the n-grams using Eq. 1.

$$BLEU = BP \times \exp(\sum_{n=1}^{N} W_n \times \log P_n) \tag{1}$$

where BP is defined in Eq. 2.

$$BP = \begin{cases} 1 & l_c > l_r \\ \exp(1 - l_r/l_c) & l_c \leqslant l_r \end{cases} \tag{2}$$

The score range of BLEU is from 0 to 1. Generated text gets a score of 1 if it's exactly the same with the reference text.

$\mathrm{NLL}_{gen}, \mathrm{NLL}_{div}$ – to evaluate diversity of the generated text, we used two metrics called NLL_{gen} [17] and NLL_{div} [42]. Both metrics are mathematically defined in Eq. 3 and 4.

$$\mathrm{NLL}_{gen} = -\mathbb{E}_{Y_\gamma \sim P_\gamma}[\log P_\theta(\gamma_1, \ldots, \gamma_T)] \tag{3}$$

$$\mathrm{NLL}_{div} = -\mathbb{E}_{Y_\theta \sim P_\theta}[\log P_\theta(y_1, \ldots, y_T)] \tag{4}$$

where, P_θ denotes the samples distribution of the generated text and $(y_1, \ldots, y_T)$ is the generated text by generator, $y_t \in \upsilon$, where υ is vocabulary of all inputted tokens, P_γ is the samples distribution of the real generated text.

3.5. Evaluation metrics – sentiment analysis

To evaluate the performance of different models applied for sentiment analysis, we used information retrieval based evaluation metrics such as accuracy, precision, recall and F1-score, defined in Eq. (5)–(8).

[2] https://www.kaggle.com/septa97/100k-courseras-course-reviews-dataset.

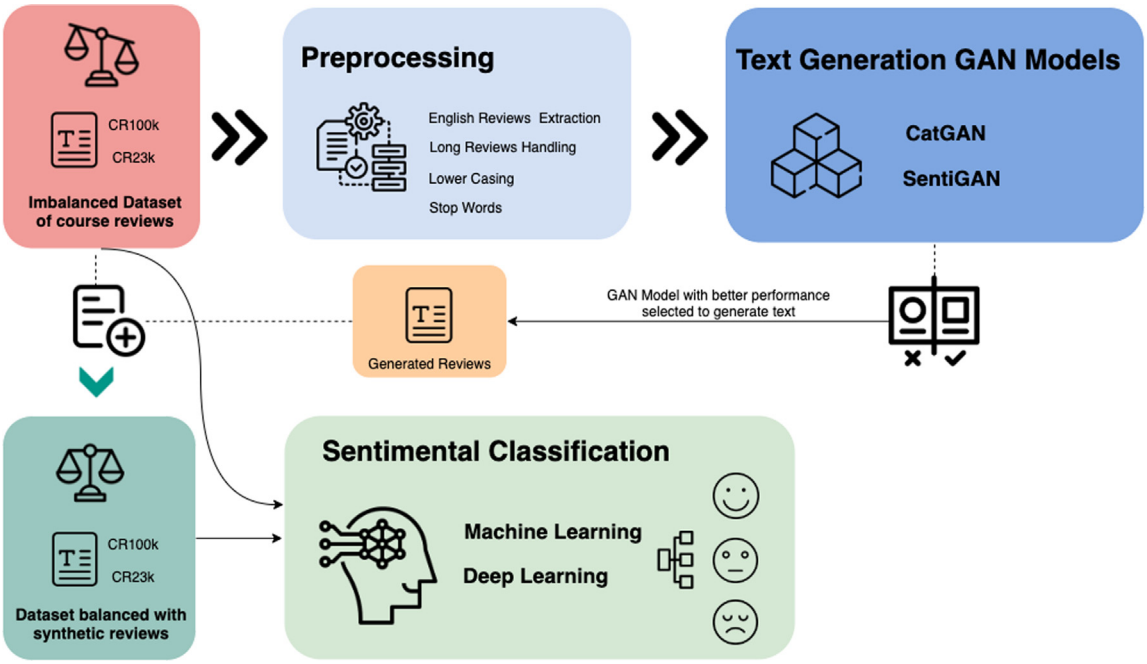
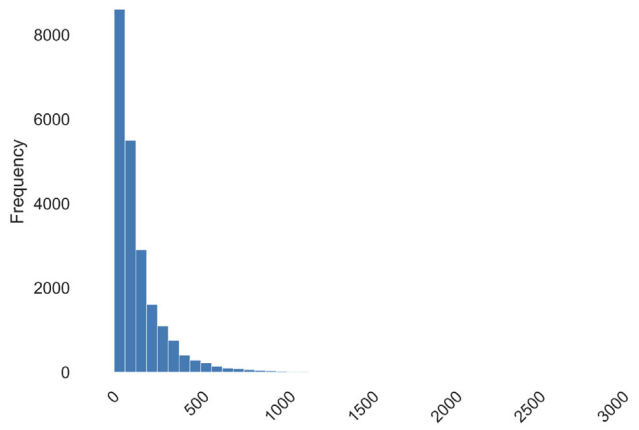
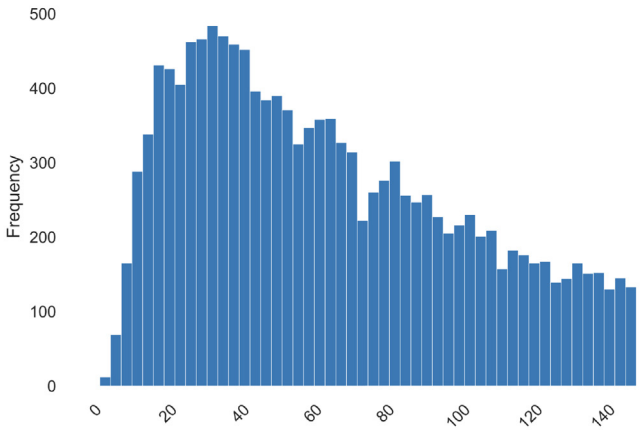


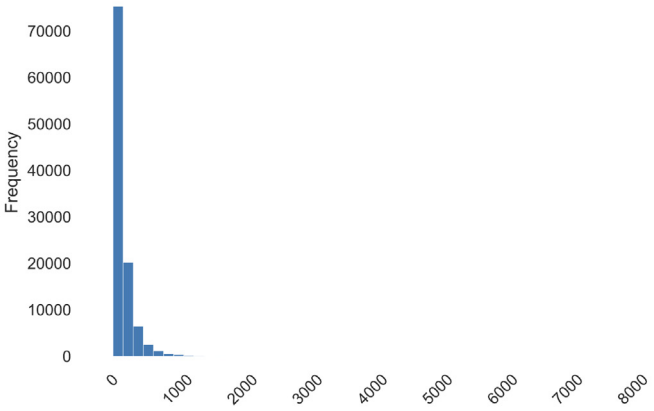
Fig. 1. High-level overview of the sentiment classification approach.



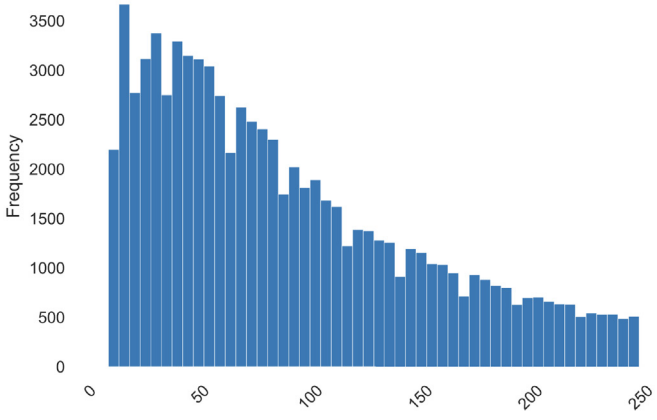
(a) Length distribution before preprocessing - CR23K



(b) Length distribution after preprocessing - CR23K



(c) Length distribution before preprocessing - CR100K



(d) Length distribution after preprocessing - CR100K

Fig. 2. Length distribution of CR23K and CR100K datasets before and after preprocessing.

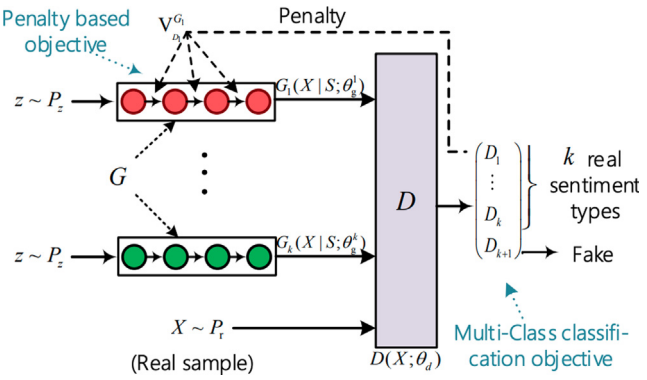


Fig. 3. SentiGAN [16].

$$Accuracy = \frac{N_{pre}}{N_{total}} \tag{5}$$

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

$$F1 - score = \frac{2TP}{2TP + FN + FP} \tag{8}$$

where, N_{pre} represents samples that are predicted correctly, N_{total} indicates total samples in the test dataset, TP is true positive, FP is false positive and FN is false negative.

4. Results and discussion

4.1. Text generation

We trained two different types of text generation models on two different datasets. In order to compare the performance between SentiGAN and CatGAN, two different evaluation metrics are used as depicted in Section 3.4. First, we will discuss the results with respect to BLEU scores with $n = 2, 3, 4, 5$ of both SentiGAN and CatGAN for CR23K and CR100K datasets, respectively. The higher the BLEU score, the generated text is more similar to the original text. Fig. 5 shows some of the important statistics regarding BLEU score calculations. The black dotted line is the division line marking the pre-training and adversarial training process. Before the black dotted line, it is the pre-training part where the initial generator was trained. The field after the black dot line is the adversarial training part where the pre-trained generator gets strengthened training aiming to improve the quality and diversity of generated text again.

As can be seen from the Fig. 5, BLEU scores of SentiGAN for CR23K and CR100K have a sharp increase to a quite high score for the positive and negative category when in the adversarial training process and then remain constant. For the neutral category, the curve is a bit different. The score increases sharply at first but then jump down instantly to almost zero and remains constant. However, BLEU scores of CatGAN for CR23K and CR100K increase slowly but stably for the positive and negative category. The sharp increase after the adversarial training in SentiGAN is due to the gradient vanishing problem in adversarial training, causing SentiGAN to lose the ability to generate diverse text and would only output the same text every time. On the contrary, CatGAN can keep the diversity of generated reviews with high text quality. This is in line with the empirical findings which shows a better of CatGAN over SentiGAN.

Next, we have used NLL_{div} and NLL_{gen} evaluation metrics to evaluate the text diversity of the generated text for both models CatGAN and SentiGAN for CR23K and CR100K datasets. Fig. 6 and 7 depict NLL_{div} and NLL_{gen} scores that are reciprocal of each other. The smaller the NLL_{div} value, the more diverse the generated text and the higher the NLL_{gen} value, the less diverse the generated text.

The NLL_{div} decreases a lot to a stable value in pre-training part and remains overall stable afterwards. When in adversarial training, NLL_{div} remains stable at the first few epochs but then decreases dramatically to 0 and keeps 0 later. NLL_{div} becomes 0 meaning the generated text every time is the same without any diversity. The same decreasing trend applies to all the three sentiment labels (i.e. positive, negative and neutral). The NLL_{gen} metric results also show the same trend but on the opposite direction because the higher NLL_{gen}, the more diverse the generated text. NLL_{gen} shows different results at the pre-training process that it decreases sharply at first and then remain constant. This means the diversity of generated text increases firstly and then remain unstable. This happened because those two metrics are specially designed for text generated from the adversarial training part and may not accurately display the generated text in the pre-training part. However, our study focuses on the adversarial part, so the slight

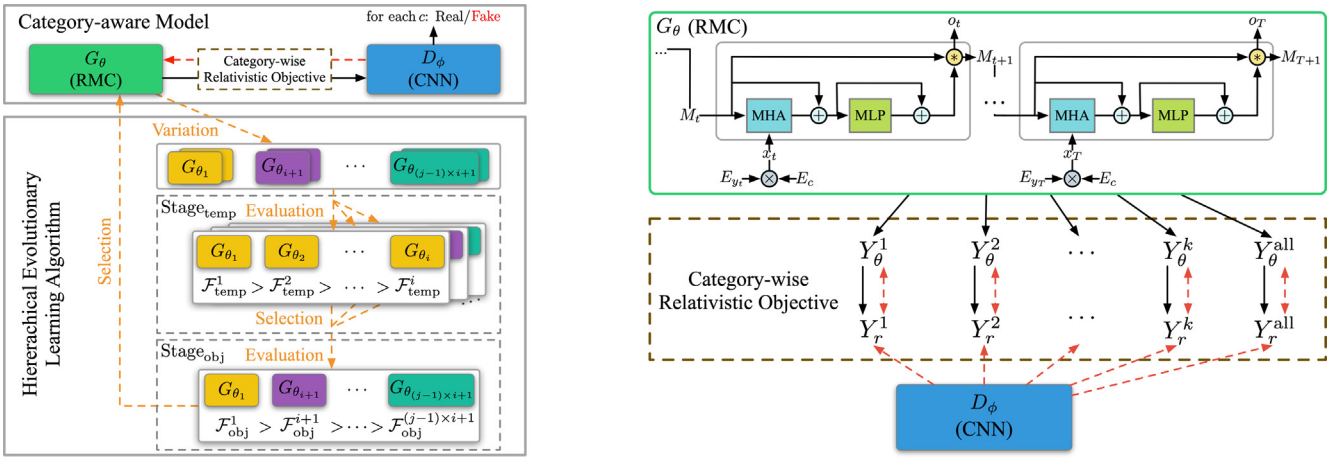
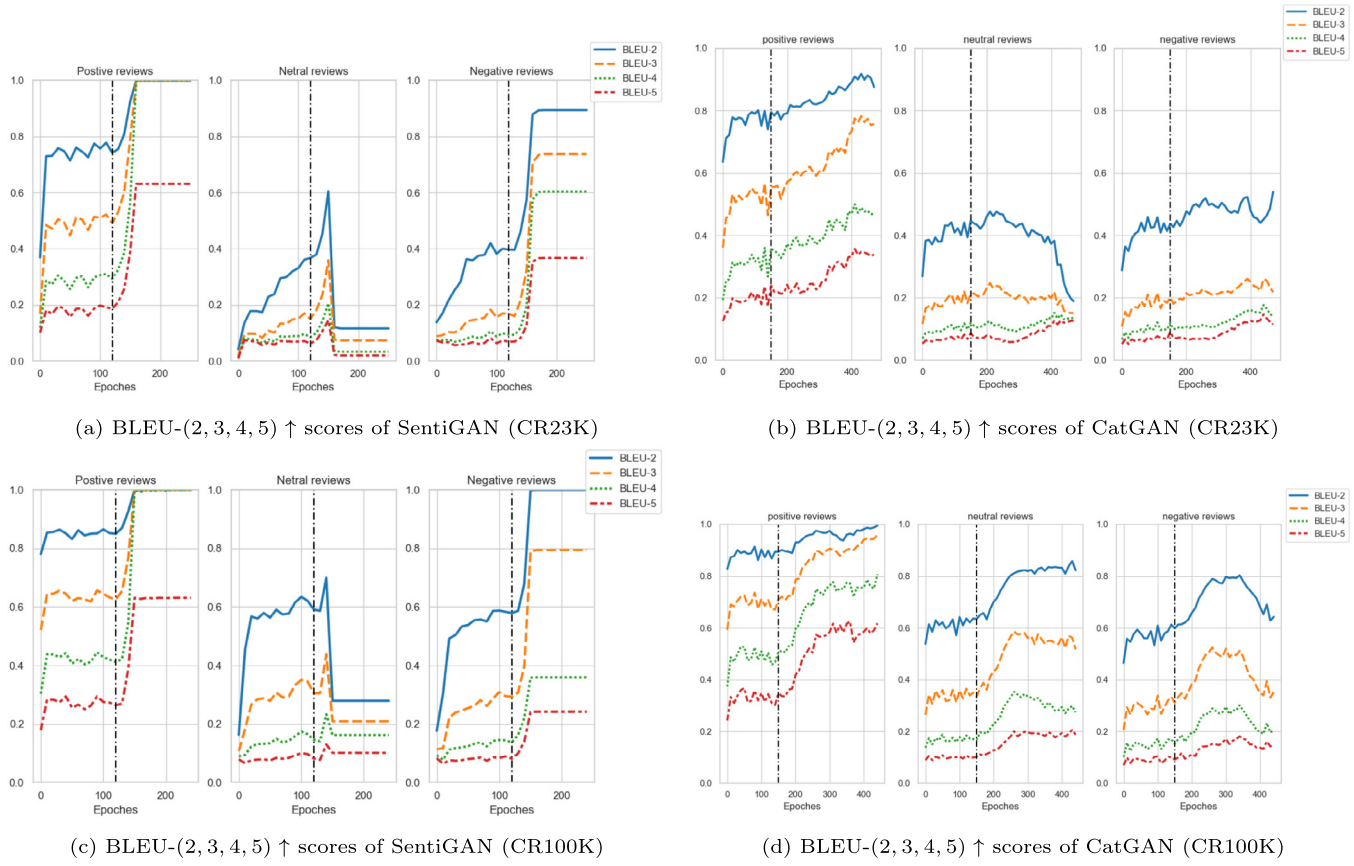


Fig. 4. CatGAN with hierarchical evolutionary learning [17].



(a) BLEU-$(2, 3, 4, 5)$ ↑ scores of SentiGAN (CR23K)

(b) BLEU-$(2, 3, 4, 5)$ ↑ scores of CatGAN (CR23K)

(c) BLEU-$(2, 3, 4, 5)$ ↑ scores of SentiGAN (CR100K)

(d) BLEU-$(2, 3, 4, 5)$ ↑ scores of CatGAN (CR100K)

Fig. 5. BLEU-$(2, 3, 4, 5)$ ↑ scores of SentiGAN and CatGAN on each sentimental polarity for CR23K and CR100K datasets. The black dot line is the division line between pre-training and adversarial training process that are before and after black dot line part, respectively.

difference of metrics results in the pre-training part is acceptable. NLL_{gen} score proves the assumption from NLL_{div} score that generated text every time is the same without any diversity in most adversarial training process parts. For CatGAN scores, when training comes to the adversarial part, the NLL_{div} would increase slightly to a local best value for all three sentimental polarities and then starts decreasing down to 0. The decreasing trend of NLL_{div} for CatGAN is mild and slow compared with SentiGAN that would experience again a fast gradient vanishing problem.

These analysis again indicate that SentiGAN has a gradient vanishing problem on both datasets CR23K and CR100K. This causes SentiGAN to lose the ability of generating diverse text and can only generate the same sentence every time. Based on the above results and analysis on both datasets, CatGAN is selected to synthesize/generate reviews, which are then used to balance the original datasets.

4.2. Balanced dataset

Based on the quality and diversity of generated text, we trained two CatGAN models to generate the text. One model was trained for 90 epochs to generated reviews for the CR23K dataset. Second model was trained for 50 epochs to generate reviews for the CR100K dataset. The total number of reviews for each sentiment label, before and after balancing for both datasets, is shown in Fig. 8.

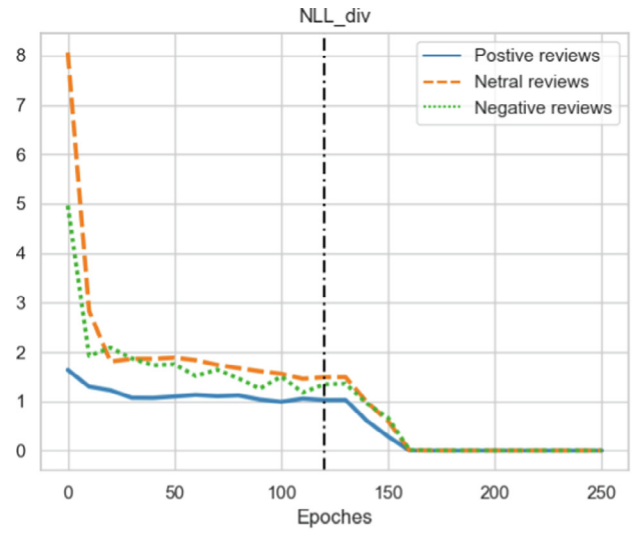
4.3. Sentiment classification

To check the impact of text generation on sentiment analysis, we selected 10 baseline machine and deep learning models. For this purpose a standard test dataset is needed. If we just split the original dataset into two parts and use one part as a test dataset, the models' actual performance will not be well tested. For example, for the original imbalanced dataset, 90% of the reviews in the dataset were positive. The model can get 90% or more accuracy but this cannot be used in real life since it cannot predict negative and neutral categories accurately. At the same time, the balanced dataset with synthetic data will be only part of the training dataset to analyze if this improves the training of the model. Therefore, a test dataset from original imbalanced dataset is needed.

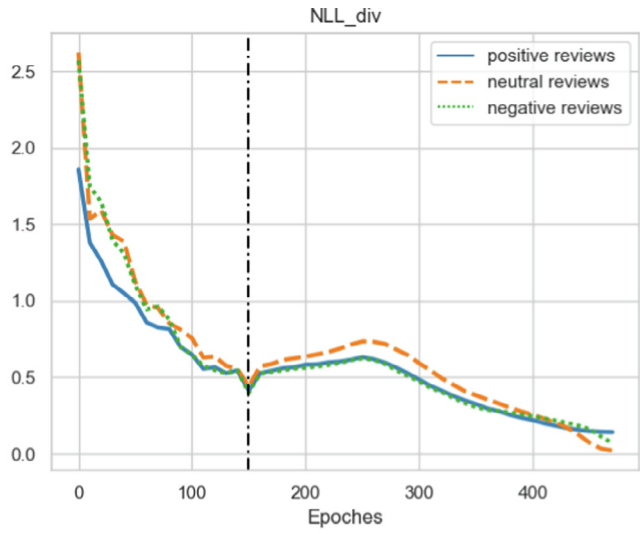
We extracted a total of 870 original course reviews from the Coursera online learning platform and labeled them manually into three sentiments (i-e: positive, negative, and neutral). The sentiment distribution of test dataset has been shown in Fig. 9a. Besides this, to reflect the sentiment classification model's performance accurately on real course reviews in the test dataset as much as possible, we do not have any long length reviews. The length-frequency distribution graph has been shown in Fig. 9b.

Table 2 shows the accuracy and F1-score for test dataset regarding 10 different baseline models trained on imbalanced as well as balanced datasets for both CR23K and CR100K datasets. It is evident that the models which were tested on balanced dataset have better accuracy and F1-score as compared to the models which were only trained on the original (imbalanced) dataset. The third column depicting the difference between the accuracy and F1-score between results obtained on imbalanced and balanced dataset clearly shows significant performance improvement on sentiment classification task. Moreover, the results support the selection of CatGAN over the SentiGAN text generation model for balancing the dataset.

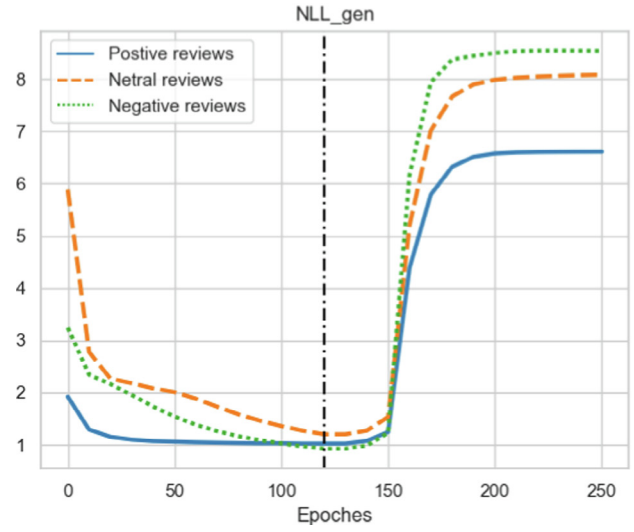
It is interesting to note from results shown in Table 2 that more complex and deep model architectures perform worse on the



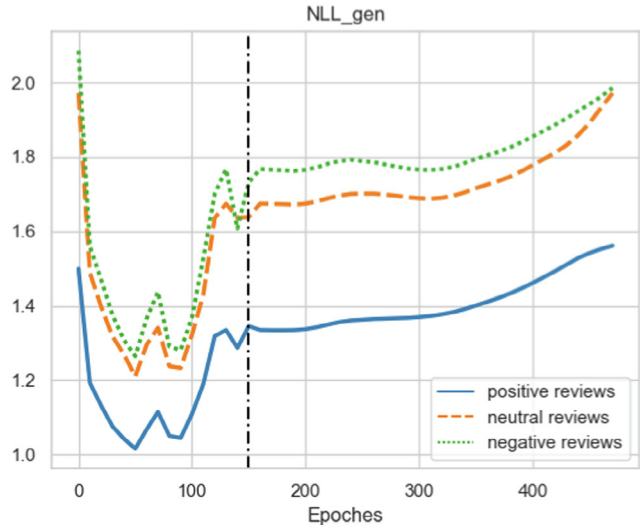
(a) NLL_{div} scores of SentiGAN (CR23K)



(b) NLL_{div} score of CatGAN (CR23K)



(c) NLL_{gen} scores of SentiGAN (CR23K)



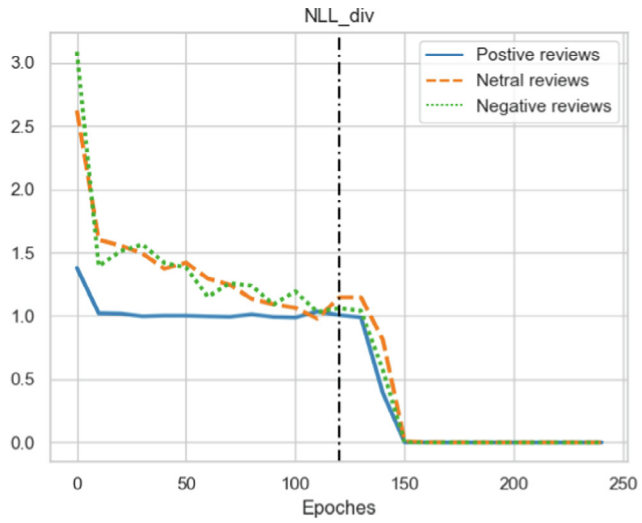
(d) NLL_{gen} scores of CatGAN (CR23K)

Fig. 6. NLL_{div} ↑ and NLL_{gen} ↓ scores of SentiGAN and CatGAN on each sentimental polarity for CR23K and CR100K datasets.

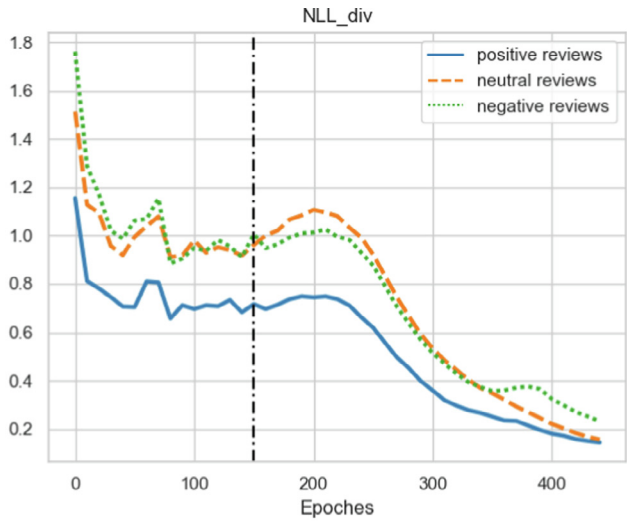
highly imbalanced dataset. For instance, as in the case of BERT (Bi-LSTM and GRU) as compared to other DNN variants and conventional machine learning algorithms. It is highly likely due to a greater number of network parameters to train. Thus the network may need more data fed in order to achieve outstanding performance. Therefore, when the dataset is small or highly imbalanced, the models are more likely to be influenced by major classes. The results of conventional machine learning techniques (SVM, Naive Bayes, Decision Tree, AdaBoosting) and RNN also indicate the same, as their network structure is relatively simpler than other deep learning models. Therefore, the performance is only slightly improved on CR23K dataset even after balancing, as the ratio of negative and positive class samples is small. Additionally, significant performance improvement is observed in case of LSTM and GRU models that have BERT transformers. The BERT model utilized in this study has eleven layers. These consist of self-output layer, attention layer, intermediate and output layers. In case of imbalanced dataset, LSTM and GRU with BERT transformers are more likely to focus on the majority category (i.e., positive class in our experiment) but ignore the other class with fewer data. Balancing the dataset with equal number of classes therefore improves the accuracy and F1-score of these models. Similar results pattern can be observed for CR100K dataset.

Another interesting fact to note is that while the gap between positive and negative or neutral classes increases in the dataset, the performance of sentiment classification models improves. The proportion of instances in positive and neutral category is $18746 \div 2316 \approx 8.094$for the CR23K and $74191 \div 2602 \approx 28.51$ for the CR100K. The accuracy of all baseline models, including conventional machine learning algorithms, is improved for 2.04 percentage points for CR23K dataset and 4.82 percentage points for CR100K dataset. The performance of models trained on CR100K dataset is more than two times greater than the average improvement of models trained on CR23k and an explanation for this is the size CR23k dataset which contains a greater number of instance. The average improvement of models trained on both datasets is shown in Table 3.

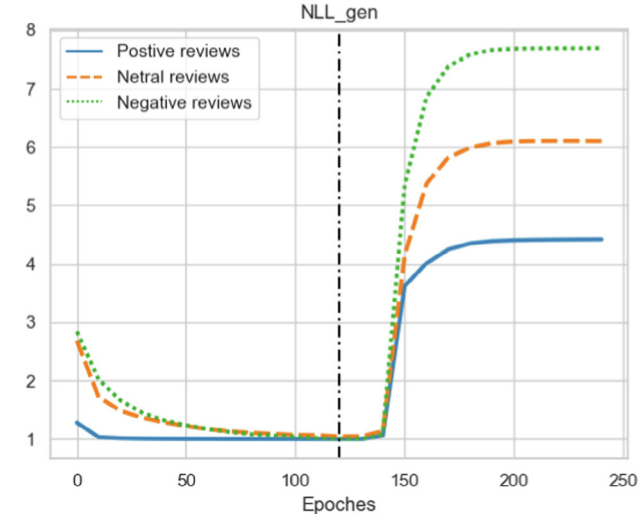
We can easily conclude that the CatGAN's performance is improved on a balanced dataset as compared to the original imbalanced one. It shows that the synthetic text samples can add value



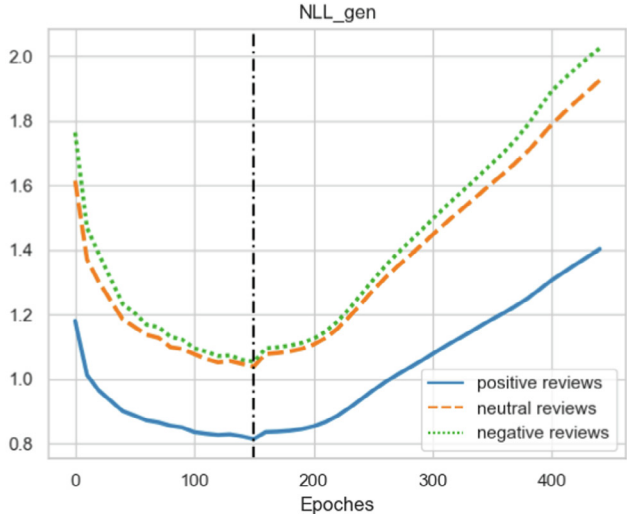
(a) NLL_{div} scores of SentiGAN (CR100K)



(b) NLL_{div} score of CatGAN (CR100K)

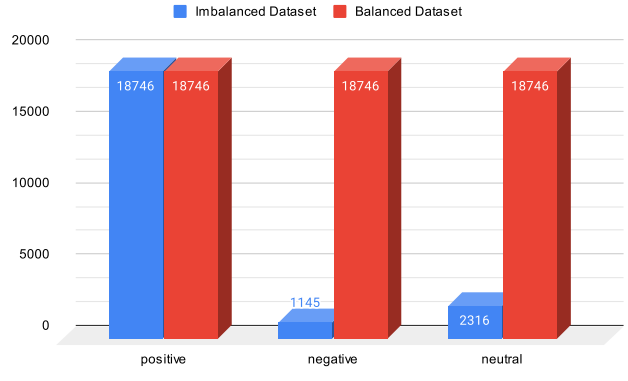


(c) NLL_{gen} scores of SentiGAN (CR100K)

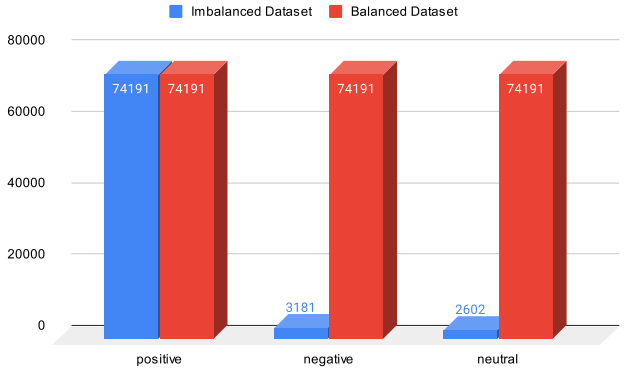


(d) NLL_{gen} scores of CatGAN (CR100K)

Fig. 7. NLL_{div} ↑ and NLL_{gen} ↓ scores of SentiGAN and CatGAN on each sentimental polarity for CR23K and CR100K datasets.

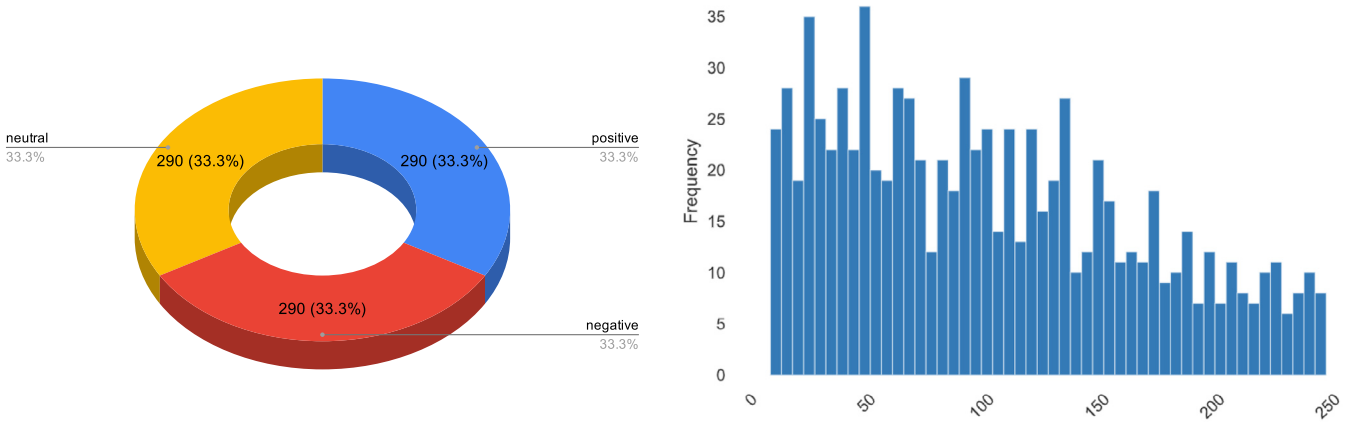


(a) CR23k dataset before and after balancing



(b) CR100k dataset before and after balancing

Fig. 8. CR23k and CR100k dataset before and after balancing.



(a) Distribution of sentimental polarity (b) Histogram of lengths of sentimental polarity

Fig. 9. Distribution of sentiment polarity and length frequency histogram for test dataset.

Table 2
Overall results of different algorithms and models for imbalanced dataset and dataset balanced with generated reviews from CatGAN model for CR23k and CR100k dataset.

Dataset	Model	Imbalanced Dataset		Balanced Dataset		Difference	
		Accuracy (%)	F1-score (%)	Accuracy (%)	F1 score (%)	Accuracy	F1-score
CR23K	SVM (RBF kernel)	32.07	**31.35**	**33.21**	27.62	1.15	−3.73
	Decision Tree	**34.25**	34.12	34.14	**32.83**	−0.11	−1.11
	Naive Bayes	32.41	**27.12**	**32.53**	26.10	0.12	−1.02
	AdaBoosting	31.83	**24.13**	**32.53**	21.54	0.69	−2.59
	RNN	33.97	**17.27**	**34.10**	17.14	0.13	−0.13
	Bi-LSTM (GloVe)	62.31	58.49	**63.45**	**60.04**	1.14	1.55
	Bi-LSTM (FastText)	61.85	56.62	**62.19**	**58.14**	0.34	1.52
	CNN (GloVe)	56.64	**61.80**	**62.07**	61.72	5.43	−0.08
	Bi-LSTM (BERT)	33.99	16.61	**39.47**	**31.77**	5.48	15.16
	GRU (BERT)	33.65	17.11	**39.67**	**35.44**	6.02	18.33
CR100K	SVM (RBF kernel)	30.00	21.09	**32.64**	**29.48**	2.64	8.39
	Decision Tree	33.21	31.41	**34.48**	**34.41**	1.27	3.00
	Naive Bayes	31.14	20.34	**34.60**	**31.67**	3.46	11.33
	AdaBoosting	34.82	24.27	**35.86**	**25.85**	1.04	1.58
	RNN	33.52	16.53	**35.96**	**21.87**	2.44	5.34
	Bi-LSTM (GloVe)	59.89	53.73	**63.71**	**62.31**	3.82	8.58
	Bi-LSTM (FastText)	61.84	57.50	**63.48**	**61.64**	1.64	4.14
	CNN (GloVe)	55.63	58.31	**61.58**	**62.10**	5.95	3.79
	Bi-LSTM (BERT)	33.37	16.61	**44.55**	**35.36**	11.18	18.75
	GRU (BERT)	33.43	16.64	**46.21**	**43.82**	12.78	27.18

to a minority class and in overall improving the model's performance. The major analysis are summarized below:

For machine learning, algorithms with the balanced dataset for both datasets have slightly higher accuracy than algorithms with an unbalanced dataset but not very much.
For deep learning models without transformers, the bidirectional LSTM with GloVe and FastText embedding and CNN with GloVe embedding, their accuracy increases $0.34\% \sim 5.43\%$ for CR23K as well as $1.64\% \sim 5.95\%$ for CR100K. The reason why compare these three model together is that they are with the same type of embedding.
For deep learning models with transformer BERT, the bidirectional LSTM and GRU models are tested. We can see that their accuracy increases from $1.64\% \sim 5.95\%$ and F1-score increases $18.33\% \sim 27.18\%$ after balancing the imbalanced dataset with synthetic text generated by CatGAN model.

5. Conclusion & future work

The importance of mining students' opinions in online courses has gained significance as many educational institutes switch to online and digital forms of teaching amidst the COVID-19 pandemic. Such feedback helps teachers and institute better understand student needs. However, automating sentiment analysis models on student feedback is often challenging due to a low response turnaround and the lack of a publically available dataset. Data imbalance is another problem in educational settings. A highly imbalanced dataset will adversely affect the model's performance on the sentiment classification task. This article explored CatGAN and SentiGAN text generation models to generate new samples for minority classes. We further analyzed the impact of synthetic text generation on the sentiment classification task for the highly imbalanced dataset using deep learning and machine learning models. The CatGAN model was able to generate higher-quality text without losing text diversity compared to SentiGAN model and was selected to generate text to balance the highly imbalanced datasets of CR23K and CR100K. We trained several machine learning and deep learning models on synthetic balanced and imbalanced version of both datasets. The results conclude that comparing with the original imbalanced dataset, the accuracy and F1-score of the model trained on synthetic balanced dataset is improved. The accuracy increases to 2.039% and 4.822% for CR23K and CR100K dataset. Moreover, the F1-score increases to

Table 3
Summary statistics of the degree of difference of sentiment classification models performance on the balanced datasets

			CR23k	CR100k
Difference	**Accuracy (%)**	Deep learning	3.09	6.64
		Machine Learning	0.46	2.10
		Overall average	2.04	4.82
	F1-score (%)	Deep learning	6.06	11.30
		Machine Learning	−2.12	6.08
		Overall average	2.79	9.21

2.79% and 9.208% for CR23K and CR100K dataset, respectively. In the future, the ongoing work can be extended by exploiting different types of complicated text generation models like GPT-3 and more complex sentiment analysis models in order to have better and more generalized models.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Kastrati Z, Imran AS, Kurti A. Weakly supervised framework for aspect-based sentiment analysis on students' reviews of MOOCs. IEEE Access 2020;8:106799–810.
[2] Edalati M, Imran AS, Kastrati Z, Daudpota SM. The potential of machine learning algorithms for sentiment classification of students' feedback on mooc. Proceedings of SAI Intelligent Systems Conference, Springer 2021:11–22.
[3] Estrada MLB, Cabada RZ, Bustillos RO, Graff M. Opinion mining and emotion recognition applied to learning environments. Expert Syst Appl 2020;150:113265.
[4] Dos Santos C, Gatti M. Deep convolutional neural networks for sentiment analysis of short texts. In: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. p. 69–78.
[5] Imran AS, Daudpota SM, Kastrati Z, Batra R. Cross-cultural polarity and emotion detection using sentiment analysis and deep learning on covid-19 related tweets. IEEE Access 2020;8:181074–90.
[6] Kastrati Z, Ahmedi L, Kurti A, Kadriu F, Murtezaj D, Gashi F. A deep learning sentiment analyser for social media comments in low-resource languages. Electronics 2021;10(10):1133.
[7] Vanaja S, Belwal M. Aspect-level sentiment analysis on e-commerce data. In: 2018 International Conference on Inventive Research in Computing Applications (ICIRCA). p. 1275–9.
[8] Yu L, Zhang W, Wang J, Yu Y. Seqgan: Sequence generative adversarial nets with policy gradient. In Proceedings of the AAAI conference on artificial intelligence, vol. 31; 2017..
[9] Guo J, Lu S, Cai H, Zhang W, Yu Y, Wang J. Long text generation via adversarial training with leaked information. In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32; 2018..
[10] Che T, Li Y, Zhang R, Hjelm RD, Li W, Song Y, Bengio Y. Maximum-likelihood augmented discrete generative adversarial networks, arXiv preprint arXiv:1702.07983..
[11] Li Z, Xia T, Lou X, Xu K, Wang S, Xiao J. Adversarial discrete sequence generation without explicit neuralnetworks as discriminators. In The 22nd International Conference on Artificial Intelligence and Statistics, PMLR; 2019. pp. 3089–3098..
[12] Nie W, Narodytska N, Patel A. Relgan: Relational generative adversarial networks for text generation. In International conference on learning representations; 2018..
[13] Xu J, Ren X, Lin J, Sun X. Dp-gan: diversity-promoting generative adversarial network for generating informative and diversified text, arXiv preprint arXiv:1802.01345..
[14] Montahaei E, Alihosseini D, Baghshah MS. Dgsan: Discrete generative self-adversarial network. Neurocomputing 2021;448:364–79.
[15] Lu S, Yu L, Feng S, Zhu Y, Zhang W. Cot: Cooperative training for generative modeling of discrete data. In: International Conference on Machine Learning, PMLR. p. 4164–72.
[16] Wang K, Wan X. Sentigan: Generating sentimental texts via mixture adversarial networks. In IJCAI; 2018. pp. 4446–4452..
[17] Liu Z, Wang J, Liang Z. Catgan: Category-aware generative adversarial networks with hierarchical evolutionary learning for category text generation. In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34; 2020. pp. 8425–8432..
[18] Vu D-H, Le A-C. Topic-guided rnn model for vietnamese text generation. In: Research in Intelligent and Computing in Engineering. Springer; 2021. p. 827–34.
[19] Masum AKM, Islam MM, Abujar S, Sorker AK, Hossain SA. Bengali news headline generation on the basis of sequence to sequence learning using bi-directional rnn. Soft Computing Techniques and Applications, Springer 2021:491–501.
[20] Li L, Zhang T. Research on text generation based on lstm. Int Core J Eng 2021;7(5):525–35.
[21] Song L. Structural information preserving for graph-to-text generation, uS Patent App. 16/883,475 (Dec. 2 2021)..
[22] Logeswaran L, Lee H, Bengio S. Content preserving text generation with attribute controls. Advances in Neural Information Processing Systems 31..
[23] Akhtar NI, Shazol KMI, Rahman R, Yousuf MA. Bangla text generation using bidirectional optimized gated recurrent unit network. In: Proceedings of International Conference on Trends in Computational and Cognitive Engineering, Springer. p. 103–12.
[24] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In Advances in neural information processing systems; 2017. pp. 5998–6008..
[25] Guo B, Wang H, Ding Y, Wu W, Hao S, Sun Y, Yu Z. Conditional text generation for harmonious human-machine interaction. ACM Transactions on Intelligent Systems and Technology (TIST) 2021;12(2):1–50.
[26] Nadakuduti SS, Enciso-Rodríguez F. Advances in genome editing with crispr systems and transformation technologies for plant dna manipulation. Front Plant Sci 2021;11:2267.
[27] Xu JH, Shinden K, Kato MP. Table caption generation in scholarly documents leveraging pre-trained language models. In: 2021 IEEE 10th Global Conference on Consumer Electronics (GCCE), IEEE. p. 963–6.
[28] Rebuffel C, Soulier L, Scoutheeten G, Gallinari P. A hierarchical model for data-to-text generation. Advances in Information Retrieval 2020;12035:65.
[29] Fatima N, Imran AS, Kastrati Z, Daudpota SM, Soomro A. A systematic literature review on text generation using deep neural network models. IEEE Access 2022;10:53490–503.
[30] Lin K, Li D, He X, Zhang Z, Sun M-T. Adversarial ranking for language generation, arXiv preprint arXiv:1705.11001..
[31] Shaikh S, Daudpota SM, Imran AS, Kastrati Z. Towards improved classification accuracy on highly imbalanced text dataset using deep neural language models. Appl Sci 2021;11(2):869.
[32] Sindhu I, Daudpota SM, Badar K, Bakhtyar M, Baber J, Nurunnabi M. Aspect-based opinion mining on student's feedback for faculty teaching performance evaluation. IEEE Access 2019;7:108729–41.
[33] 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..
[34] Koufakou A, Gosselin J, Guo D. Using data mining to extract knowledge from student evaluation comments in undergraduate courses. In: 2016 International Joint Conference on Neural Networks (IJCNN), IEEE. p. 3138–42.
[35] Socher R, Perelygin A, Wu J, Chuang J, Manning CD, Ng AY, Potts C. Recursive deep models for semantic compositionality over a sentiment treebank. In: Proceedings of the 2013 conference on empirical methods in natural language processing. p. 1631–42.
[36] Katragadda S, Ravi V, Kumar P, Lakshmi GJ. Performance analysis on student feedback using machine learning algorithms. In: 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS). p. 1161–3.
[37] Lwin HH, Oo S, Ye KZ, Lin KK, Aung WP, Ko PP. Feedback analysis in outcome base education using machine learning. In 2020 17th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON). IEEE; 2020. pp. 767–770..
[38] Sadriu S, Nuci KP, Imran AS, Uddin I, Sajjad M. An automated approach for analysing students feedback using sentiment analysis techniques. Mediterranean Conference on Pattern Recognition and Artificial Intelligence, Springer 2022:228–39.
[39] Kastrati Z, Dalipi F, Imran AS, Pireva Nuci K, Wani MA. Sentiment analysis of students' feedback with nlp and deep learning: A systematic mapping study. Appl Sci 2021;11(9):3986.
[40] Yu L, Zhang W, Wang J, Yu YS. Sequence generative adversarial nets with policy gradient. 489 in. In AAAI conference on artificial intelligence. vol. 490; 2017..
[41] Papineni K, Roukos S, Ward T, Zhu W-J. Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. p. 311–8.
[42] Zhu Y, Lu S, Zheng L, Guo J, Zhang W, Wang J, et al. Texygen: A benchmarking platform for text generation models. In The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval; 2018. pp. 1097–1100..