Evaluating Gender Bias in Textual Data Using NBias: A Natural Language Processing Perspective

*

Protiva Das
Department of CSE
BRAC University
Dhaka, Bangladesh
protva.das@g.bracu.ac.bd

Annajiat Alim Rasel
Department of CSE
BRAC University
Dhaka, Bangladesh
annajiat@gmai

Tamima Binte Wahab
Department of CSE
BRAC University
Dhaka, Bangladesh
tamimabw@gmail.com

***Abstract*—This study tackles the challenge of detecting gender bias in textual comments, a crucial issue given the global concern over gender-related discrimination. Utilizing a secondary dataset from Kaggle, we explored the effectiveness of various natural language processing models, including LSTM, Bi-LSTM, GRU, Bi-GRU, BERT, and DistilBERT, to identify and analyze gender biases. Our approach began with preprocessing the data, followed by splitting it into training and test sets and then applying the models to determine which best detects biases at an early stage. The findings reveal that certain models are more adept at highlighting subtle biases, contributing significantly to the efforts of making online and offline communication spaces more inclusive. Looking forward, this research opens avenues for addressing other forms of bias such as political or religious, thereby broadening the impact of our work on society.**

***Index Terms*—component, formatting, style, styling, insert**

I. Introduction

In today's world, what we say online can shape how people see and treat each other. Unfortunately, comments on the internet often show bias against certain genders, which can lead to unfair stereotypes and hurt feelings. This kind of bias isn't just mean; it can also make people feel unsafe or unwelcome. Given how big these issues have become globally, it's more important than ever to address them. Our research aims to spot these gender biases in in-text comments. We decided to tackle this because we want to help reduce the harm caused by biased comments and make online spaces nicer for everyone. While there has been plenty of work on spotting hate speech, our project is different because it looks specifically at how biases against all genders show up in what people write, not just on social platforms but anywhere. To do this, we got many comments from a dataset on Kaggle, a website where people share data. We cleaned up this data and organized it, then used some smart tech tools like LSTM, Bi-LSTM, GRU, Bi-GRU, BERT, and DistilBERT to see which could best find the bias. Our goal is to catch these biases early on so we can do something about them and make the internet a better place for talking and sharing ideas. The contributions of the paper are summarized below:

Identify applicable funding agency here. If none, delete this.

1) Detecting online sexism based on gender bias with proper accuracy and precision.
2) Providing a platform automatically detects the bias and instantly hides it. The proposed research work will also apply to shifting the socioeconomic behaviour of the people.
3) Analyzing the performance of multiple RNN based models and Transformers.

II. Literature Review

These studies utilizing transformer-based token classification models have demonstrated capability in detecting biased language across various domains. However, these models struggle with misspellings, changes in severity, embedded biases in training data, and performance inconsistencies across languages and cultural contexts. These limitations of [1] underscore the need for refined approaches to enhance accuracy and generalizability in bias detection.

Evaluated bias detection in narratives generated by large language models like GPT-3.5, using it both to produce and analyze stories for [2] inherent biases, and employing a fine-tuned GPT-2 for dedicated bias classification. Despite achieving a high accuracy of 97.5

[3] Introduces the Contextualized Bi-Directional Dual Transformer (CBDT) Classifier, employing two synergistic transformers for enhanced bias detection. However, its focus remains on English textual data, omitting biases in images and audio and not incorporating newer technologies like LLAMA 2 and GPT-4. These limitations highlight opportunities for broader and technologically updated investigations.

Researchers worked on the [4] Contextual-Dual Bias Reduction Recommendation System (C-DBRRS) for news, tested on two real-world datasets. Challenges include limited scalability due to computational complexity, sensitivity to hyperparameter settings, and potentially inaccurate assumptions about the importance of temporal dynamics in user-item interactions. These limitations suggest critical areas for further model optimization and broader applicability testing.

Multilabel classification of toxic comments using word embeddings (GloVe, Word2vec, FastText) and deep neural networks (NN, CNN, RNN, LSTM, GRU), with the BiGRU model achieving the highest performance metrics were worked on [5]. Despite these successes, the research highlighted a need for improved strategies to address imbalanced data effectively.

Introduced an LSTM-LM transfer learning model designed to automatically identify bias in enterprise content using manually tagged documents, achieving an accuracy of 0.89. It [6] compared this approach against 10 baseline models, effectively demonstrating the LSTM-LM's robustness in performance. This underscores the model's potential for enhancing bias detection processes within organizational settings. Headline Attention Network for bias detection in news articles, uniquely designed to mimic human reading patterns and incorporate an attention mechanism focused on headlines. This model in [7] achieved 89.54

The study [8] introduced DA-RoBERTa, DA-BERT, and DA-BART for sentence-level media bias detection, achieving an F1 score of 0.814. While effective, the models are limited to sentence-level bias and evaluations on a single dataset. Future improvements should extend capabilities to detect biases across different textual levels and develop more nuanced evaluation metrics to enhance detection accuracy comprehensively.

BERT-BiGRU model, combining BERT's contextual word representations with BiGRU's bidirectional text feature extraction for text classification in [9]. Achieving high metrics (accuracy, recall, F1 score all above 0.9), the model is however resource-intensive due to BERT's complexity, relying on Google's pre-trained models. Optimization is needed to reduce resource consumption and training time.

The research project examined LSTM and GRU models and used a Bi-GRU-LSTM-CNN classifier for deep learning, with an F1-score of 70.576

For stance identification, a two-channel CNN-GRU fusion network outperformed SVM, CNN, GRU, and an earlier hybrid model in [11] and got accuracy and F1 scores. To improve the method and use it for public opinion analysis, this network includes convolutional layers for local feature extraction and GRU for time characteristics. In the future, it is intended to add an attention mechanism and expand the dataset.

The study in [12] presents a bidirectional LSTM model for fake news detection, demonstrating superior accuracy over CNN, RNN, and unidirectional LSTM using two unstructured news datasets. It achieved a 4.18

A. Methodologies

At first, the data are gathered properly for performing bias detection from numerous social media comments. Most of the comments are sexism-based. The comments are subcategorized into multiple factors. For, evaluating the model properly, the model has been split into three categories namely, training, testing and validation. At first, the model is trained on 70% data from the gathered dataset. Later, the remaining 30% data are preserved for the validation set. Finally, the test

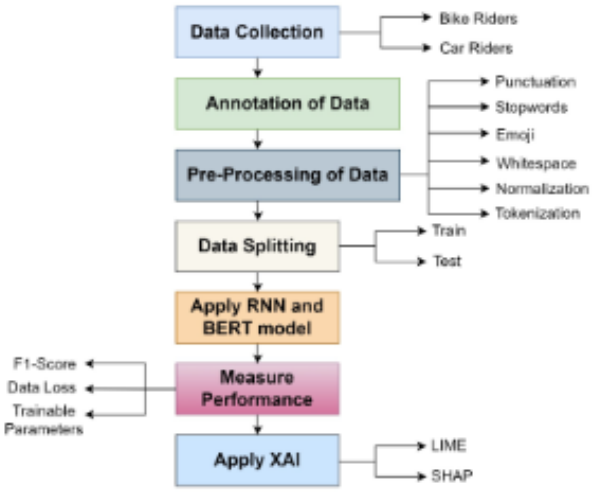

Fig. 1. Proposed Methodology

data is supplied from outside comments. The dataset needs to be annotated properly in order to understand the model precisely. More than, 20,000 comments are available in this dataset where there are multiple columns representing different kinds of biasness towards people.The main categories of the comments are:

1) **ID Rewire**
2) **Comment**
3) **Sexiest label**
4) **Category label**
5) **Vetor label**

Here, Rewire id represents two things. The first thins is the year in which the comment was made. Secondly, what type of person has made the comment. The biased comment related to sexism denote what the actual comment is. After that, the sexiest label category represents whether the comment is sexiest or not. Be specificly, what type of sexiest comment this is. In the dataset, there is also a sub sector that is known as label vector. The subcategory of the comment is denoted by the Label vector. All the necessary attributes are available inside this dataset. That is the cardinal reason for selecting this dataset.

1) **Animosity**
2) **Online threats**
3) **SPrejudiced Discussion**
4) **Derogation**

From data exploration, we can see that there are 20,000 comments in total. Twelve thousand are related to sexism, and others are not related to sexism. The comments subcategories are also important for processing data. Figure 2 shows the detailed procedure.

B. Preprocessing of the dataset

Data preprocessing is vital for providing data to numerous architectures to understand semantics precisely. It includes a good number of operations, which depend on the language and

pattern of the dataset. Another important part is ensuring the model is not overfitted because of the anomaly in the dataset. For Bengali, numerous preprocessing techniques are available. Hence, the authors have used only the vital operations required to provide the DL and ML architectures. The adopted preprocessing techniques are stated below:

1) **Dropping Null Values:** The very thing is to drop the null values from the dataset. We have utilized the Pyhton nltk library to operate. Other than that, all other columns are eradicated from the dataset except for the comment attribute.
2) **Removing Stopwords and Keywords:** Bengali language consists of many stopwords that do not provide any special meaning to the models. On the contrary, they introduce ambiguity while capturing the inner meaning of these sentences. That is why all the stopwords have been removed. Another vital factor is to eradicate the punctuation marks from the sentences. Punctuation marks also do not provide any special meaning in these sectors.
3) **Special Character Removal:** All the special characters from the comments have been removed. To fulfill this purpose, the RE module of Python is used.
4) **Tokenization:** Tokenization is the process of dividing a written document into smaller pieces so that machine learning methods can be used. We have converted all the big comments into smaller tokens. That allows the models to understand and capture the words that are helpful to understand the semantics. The Bangla BERT tokenizer has been utilized to provide the model to the Bangla BERT model.
5) **Creating Dictionary:** A dictionary has been created to identify unique words. The word definition is also created using this dictionary. As we are mining semantics from the texts, we have created a dictionary to extract the inside information with precision. The word confusion can be resolved with the aid of this dictionary. Dataset represents the most occurred words encountered approximately 309 times. Apart from this, there are 7404 words available in the dataset, where 507 words are unique. After closely understanding the words, it can be observed that most of the words are related to finance.
6) **Stemming:** After observing closely, the third and fourth words disclose the same meaning but differ in spelling. The primary purpose of stemming is to convert the data into its base form. As a result, it becomes easier for the models to understand the meaning. As a result, the number of variations reduces to a greater number. In this research, stemming also plays a vital part in handling the sparsity of the feature space.
7) **Word Embedding:** The cardinal purpose of using Word Embedding is to represent data in a dense vector in a vector space. Word embedding is a proper solution for a one-hot encoder that produces high-dimensional vectors. Learning distributed representations of words based on their context in a sizable corpus of text data is the fundamental notion behind word embedding. In this research, authors have utilized neural network word embedding techniques such as Word2Vec and Glove vectorizer. For the Bangla BERT model, no embedding technique has been adopted as this model incorporates word embedding into its architecture.

C. Experimental Models

Several ML and DL architectures were utilized to validate the dataset. This section provides a brief summary of these models. While running these architectures, the train-to-test size was 70% % to 30%.

a) **LSTM:** LSTM is one of the most significant architectures based on Recurrent Neural Networks (RNNs). The main goal of utilizing this is the ability to solve the vanishing gradient problem. This architecture's memory cells can comprehend the meaning behind a lengthy string of words. To regulate the information flow, three gates are used: the forget gate, the input gate, and the output gate. These gates enable us to update and forget information according to semantics regularly. Two LSTM layers with a hidden layer size '45 are employed for every embedding layer. To produce the probability distribution, the model has an internal Softmax layer. The benefits of LSTM are stated in the below portion:

i) **Handling the Dependencies:** The introduction of LSTM ensures the proper address of the vanishing gradient problem, one of the major issues with the Recurrent Neural Network (RNN). The LSTM model properly understands the long-term dependencies. The Cell state available in the LSTM allows for the preservation of information. The long sequence is understood properly with the availability of these cells.

ii) **Preservation of Memory:** The most important factor is that LSTM can capture long-sequence information. In the proposed research, the dataset has long sentences, and sequence information is the most important factor. The presence of a conveyor belt can update, delete and add important information throughout three gates.

The parametric details are stated below Fig 3.

8) **Gated Recurrent Unit:** The Gated Recurrent Unit (GRU), an alternative to the LSTM network, is another popular RNN architecture. The Reset and Update gates are the only two gates in this architecture, but they allow much semantic capturing. While the update gate includes fresh information, the reset gate typically handles previously collected data. The primary benefit of GRU lies in its computational speed. Table **??** provides the GRU's parametric details.

Hyperparameter Name	Value
Number of epoch	30
Activation function	Softmax
LSTM layers	16
Embedding vector length	64
Recurrent Dropout	0.15
Amount of Train data	0.70

Fig. 2. LSTM Hyperparameter

Hyperparameter Name	Value
Number of epoch	30
Activation function	Softmax
GRU layers	8
Embedding vector length	64
Recurrent Dropout	0.18
Percentage of train data	0.70

Fig. 3. GRU Hyperparameters

The main properties of GRU are stated below:

a) **Generality and Simplicity:** Unlike LSTM, GRU has a simple architecture with fewer trainable parameters than the mentioned architecture. As a result, the required training time is less than LSTM. Another important property is that GRU has a tiny reduced computational overhead when the amount of resources is limited.

b) **Efficiency in smaller Datasets:** GRU can understand data properly even if the dataset is smaller. It has good sequence-capturing ability and can optimize sequences properly for certain applications.

c) **Adaptability:** Like LSTM, GRU can also handle sentences of variable length. This architecture can be trained on any dataset and can properly handle sequence information. Sometimes, preprocessing techniques such as padding and truncation are required for flexibility.

The parametric details are stated in Fig 4.

9) **Bidirectional Long Short-Term Memory:**
The Bidirectional Long Short-Term Memory, or BiLSTM, functions similarly to an intelligent assistant by evaluating sounds or words from both the forward and backward directions. This approach is useful for jobs requiring thorough context comprehension, like text generation, language translation, and even sentence-level word prediction. It may be set up to go into data sequences and detect the tiny clues that reveal a more complete tale using tools like TensorFlow or PyTorch. BiLSTM's ability to retain specifics and recent information from far back in the sequence makes it an effective tool for situations where every nuance of context matters. It's similar to talking to someone who hears the final few words and retains the entire conversation. Because of this feature, BiLSTM is the preferred choice for projects where precise forecasting or analysis depends on knowing the details of the situation.
The BiLSTM architecture provides advantages in many cases. Some of the reasons are stated above:

a) **Ability to capture Bi-directional Context:** The cardinal advantage of BiLSTM is that it can capture sequence in both forward and backward ways. This mechanism properly understands not only past information but also future information.

b) **Improved Memory Access:** BiLSTM can access more contextual information from both sides, which improves memory retention. This approach mostly benefits machine translation and sentiment analysis.

c) **Improved Feature Representation:** BiLSTM can understand and learn features properly considering the future and past information. The sequential patterns help the model to understand features in numerous tasks.

10) **Bidirectional Gated Recurrent Unit:** The Bidirectional Gated Recurrent Unit, or BiGRU for short, enables computers to comprehend data sequences, such as sounds or sentences, by examining them from start to finish. This method works well for jobs where context is crucial, such as summarizing lengthy articles, identifying speech, or determining the mood of a document.

a) **Difference in Memory Management:** While BiLSTM has more complex mechanism of managing memory, BiGRU manages this operation properly. The main difference is that,BiGRU has fewer parameters with only two gates. The gates are known as Reset gate and Update gate.

b) **Complexity in Computation:** BiLSTM has more trainable parameters, so it requires more computations. Because of its complex architecture, BiLSTM shows such patterns. On the contrary, BiGRU is efficient in such cases. Deployment of this model is easier due to the low number of parameters. They reflect great performance in resource-constrained environments.

c) **Handling Dependancy:** Although both BiLSTM and BiGRU has almost same capability to capture long sequence by processing input in both ways. Here, both ways refer to forward and backward direction. Experimental results exhibit that BiLSTM has slightly better sequence-capturing capability. The primary reason is that, BiLSTM keeps separate

Hyperparameter Name	Value
Number of epoch	20
Activation function	Softmax
Learning rate	0.03
Optimizer	Adam
Recurrent Dropout	0.25

Fig. 4. BiLSTM and BiGRU Hyperparameters

memory cells for longer periods. Fig 5 shows the parametric details of BiLSTM and BiGRU.

11) **BERT:** Transformer models are widely popular for multiple tasks. One of the successful transformer models is BERT. The elaboration of BERT stands for Bidirectional Encoder Representation from Transformers. Primarily, this model is successfully utilized for NLP tasks. The main advantage of these models is they can perform sequence-to-sequence tasks very well.
    a) **Understanding the Contextual Representation:** For most of the symbolic language models, word embeddings are performed in a unidirectional way. Examples of this type of model are Word2Vec and GloVe. GloVe is typically used for deep learning architectures. The first bi-directional approach was taken by BERT, where sequence information was captured from both directions. The task is performed simultaneously from both sides. As a result, the context is understood precisely and properly, hence providing an accurate representation.
    b) **Availability of Transformer Architecture:** A transformer is basically a Deep learning architecture that was introduced by the researchers based on the self-attention mechanism. This model is highly effective for sequential data. BERT understands the contextual information with the aid of self-attention mechanism. Here, contextual information refers to the embeddings represented by the words.
    c) **Pretrained Architecture:** The main advantage of BERT is that it is pretrained on a large-scale corpus. Unsupervised learning is one of two popular learning techniques used here. During the training phase, BERT usually predicts masked words, which are referred to as hidden words within a sentence. For downstream tasks, BERT is usually fine-tuned properly.
    The description of the BERT model is depicted in Fig 6.

D. Performance Metrics:

Although numerous performance metrics have been adopted, including sustainability, authors have focused on

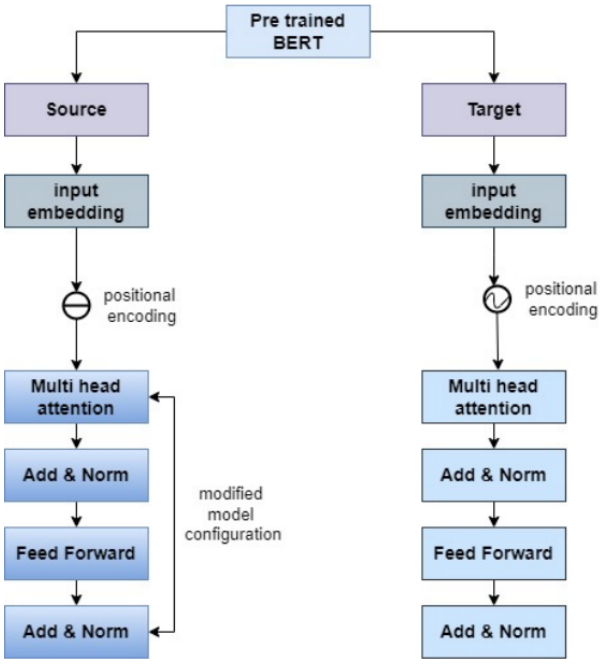

Fig. 5. Description of BERT

precision, recall, accuracy, F1-score, data loss, and trainable parameters. As DL and transformer-based architectures can not be judged based on the first four criteria, other techniques have also been considered. The formulas for precision, recall, accuracy, and F1-score are below.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{1}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{2}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{3}$$

$$\text{F1 score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

E. Result Analysis

The authors are focused primarily on the above performance metrics to observe the models' performance. Table 7 demonstrates the performance shown by the models, where it is found that BiGRU shows better precision, recall, and accuracy than other models. As the dataset has a rating of 1 to 5 available and a comment associated with the rating, precision, recall, and accuracy were found for each class. Table II-E mentions the average result of these performance metrics.

Furthermore, the authors are focused on the F1-score exhibited by the RNN architectures. It has been found that BiGRU surpasses other models on the slightest margin. Figure **??** exhibits an F1 score of 92.45%, the zenith value, whereas ANN shows only 81.25%, the lowest value in the same case. Fig 7 shows the result of the BiGRU.

TABLE I
PERFORMANCE OF THE MODELS OF THE JJJ DATASET

Name of the Model	Precision	Recall	F1-Score
LSTM	88.07%	88.19%	87.35%
GRU	83.44%	84.25%	82.16%
BiLSTM	90.17%	89.49%	90.14%
BiGRU	92.45%	91.78%	92.67%

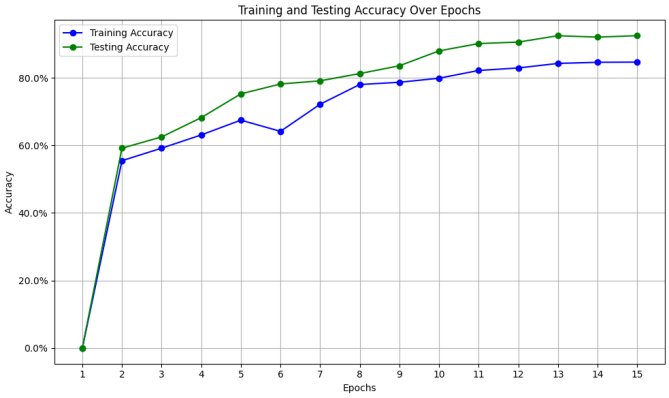

Fig. 6. BiGRU result epochwise

Epoch	Avg Precision	Avg Recall	Avg Accuracy	Avg F1-Score
1	79.05	62.19	54.51	68.45
2	82.95	66.49	67.47	71.22
3	83.57	68.81	72.48	73.32
4	85.61	69.57	73.88	74.56
5	85.40	68.66	78.47	76.82
6	85.83	71.44	77.42	77.45
7	84.80	72.57	82.47	78.23
8	86.87	73.74	80.53	78.45
9	87.25	73.29	81.66	80.25
10	87.21	73.83	81.37	80.63

Fig. 7. Result Shown by BERT Architecture

The best result has not been by the transformers as the dataset is not sufficient enough to train the models properly. Even hypertuning of parameters didn't allow the model to be precise enough.

III. CONCLUSION AND FUTURE WORK

In this work, authors are focused on detecting gender bias based on sexism that are leading people towards depression. Now the factor is that, the dataset was a balanced one where several subcategories were considered. Before feeding the dataset to models necessary preprocessing had been adopted. Five models were considered, with BERT as a transformer model. Due the insufficient amount of data, transformer model did not work precisely. RNN-based architectures have shown the capability of capturing sequence information with this limited amount of data. BiGRU had outperformed other models with more than 92% f1-score. In the future, we will extend the dataset and fine-tuning the transformer architecture. The necessary hardware constraints will also be resolved.

REFERENCES

[1] International Air Transport Association, "The economic impact of the aviation industry," *IATA Economics*, 2023. [Online]. Available: https://www.iata.org/economics.

[2] International Air Transport Association. (2021). Annual review 2020: 76th Annual General Meeting. IATA. Retrieved from https://www.iata.org/annual-review-2020

[3] Air Transport Action Group. (2021). The impact of COVID-19 on the aviation industry. ATAG. Retrieved from https://www.atag.org/impact-of-covid19.html

[4] Boeing. (2023). Boeing forecasts 763,000 new pilot jobs worldwide by 2035 as demand for air travel recovers. Boeing Newsroom. Retrieved from https://boeing.mediaroom.com/2023-03-09-Boeing-Forecasts-763000-New-Pilot-Jobs

[5] Vulturius, S., Budd, L., Ison, S., & Quddus, M. (2024). Commercial airline pilots' job satisfaction before and during the COVID-19 pandemic: A comparative study. Research in Transportation Business & Management, 53, 101108. DOI: 10.1016/j.rtbm.2023.101108.

[6] P. Kioulepoglou, S. Chazapis, and J. Blundell, "A comparative analysis of job satisfaction among military and airline pilots: During, and post COVID-19," Research in Transportation Business & Management, vol. 53. Elsevier BV, p. 101103, Mar. 2024. doi: 10.1016/j.rtbm.2024.101103.

[7] Athanasiadou, C., Theriou, G., & Chatzoudes, D. (2024). Flying responsibly: effects of perceived corporate social responsibility on attitudes and behaviors of employees in the European aviation industry. International Journal of Organization Theory & Behavior. DOI: 10.1108/IJOTB-06-2023-0156

[8] Han, T. Y., Bi, J. W., & Yao, Y. (2024). Exploring the antecedents of airline employee job satisfaction and dissatisfaction through employee-generated data. Journal of Air Transport Management, 115, 102545. DOI: 10.1016/j.jairtraman.2023.102545

[9] Samosir, J., Purba, O., Ricardianto, P., Triani, D., Adi, E., Wibisono& Endri, E. (2024). The role of service quality, facilities, and prices on customer satisfaction in Indonesia aviation in the COVID-19 pandemic. Uncertain Supply Chain Management, 12(1), 91-100.

[10] Elshawi, R., Wahab, A., Barnawi, A., & Sakr, S. (2021). DLBench: a comprehensive experimental evaluation of deep learning frameworks. Cluster Computing, 24, 2017-2038.

[11] Zeynel, E. (2023). An Investigation on Hope and Life Satisfaction of Employees in the Aviation Sector in New Normal Era. Journal of Aviation, 7(1), 133-140. DOI: 10.22161/ija.73.12

[12] T. Jayawardana and K. Abdul-Cader, "The Impact of the COVID-19 Pandemic on Airline Cabin Crew Motivation," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 59, no. 3, pp. 1089-1095, Mar. 2023