IELTS Writing Essay Score Prediction

Andrea Lolli, Ana Suarez, Valentina Bitetto 508285

INTRODUCTION

This report presents an in-depth analysis of the 'IELTS Writing Scored Essays Dataset,' a valuable repository of sample essays designed for the writing component of the International English Language Testing System (IELTS) examination. The IELTS is internationally recognized as a pivotal tool for evaluating language proficiency, serving the dual purpose of academic and general language assessment. The writing segment, in particular, holds a special significance as it assesses candidates based on specific and standardized criteria.
Within the confines of the IELTS Essay Dataset, we embark on an exploration guided by text mining techniques, aiming to process knowledge and recognize emerging patterns from this extensive collection of essays.
Our analysis will traverse multiple dimensions of the dataset, focusing on the structure and essence of these essays. We will focus on aspects such as vocabulary utilization, grammatical complexity, coherence, and the potential interactions between linguistic features and essay scores.

OVERVIEW OF THE DATASET

The IELTS Essay Dataset contains a large quantity of information, providing a comprehensive view of candidates' written linguistic abilities across a diverse spectrum of topics. To understand the dataset's context, it's essential to grasp how the IELTS essay evaluation works. This evaluation includes the following key elements:

- **Test Format**: The IELTS Writing test consists of two tasks, each designed to assess different aspects of written communication.
    - **Task 1**: usually a descriptive or informational report, a letter, or a summary of information conveyed through diagrams or charts.
    - **Task 2**: the IELTS essay (focal point of this dataset), requiring test-takers to construct an essay in response to a specific prompt.
- **Task 2 Prompt**: The prompts in Task 2 will present a question or statement on a specific topic, spanning from technology and education to health, society, culture, and the environment.
- **Word Limit and Time Constraints**: essays must be at least 250 words in length within a strict 40-minute time frame. Deviation from this word limit may result in a scoring penalty.
- **Essay Structure**: Importance of a well-structured introduction, organized body paragraphs, and a concise conclusion is emphasized.
- **Coherence and Cohesion**: The dataset underscores the necessity of coherence and cohesion in essays, with logical progression and the effective use of linking words.
- **Vocabulary and Grammar**: Attention to vocabulary diversity and grammatical accuracy is crucial to ensuring the clarity and quality of essays.
- **Task Response**: Effective essays directly address the given prompt, avoiding deviations or irrelevant content.

- **Scoring**: IELTS essays are scored on a scale from 0 to 9, with evaluations based on criteria including task achievement, coherence and cohesion, vocabulary usage, and grammatical proficiency.

The IELTS Essay Dataset provides a special chance to explore language proficiency assessment deeply. It can uncover trends and insights that improve our understanding of global language assessment and enhance English proficiency assessment methods.

PROJECT OBJECTIVE

Our project objectives involves a variety of steps, beginning with the preprocessing of the IELTS Essay Dataset. During this phase, we clean and prepare the textual data, removing noise and irrelevant elements, and ensuring that it's prepared for the further analysis.
Next, we proceed into vector representation of words, commonly known as embeddings. This step involves transforming the textual content into numerical representations that capture the semantic meaning and contextual relationships between words. These embeddings serve as the foundation for our subsequent analyses.
The heart of our project lies in the creation of a robust and insightful model. Using natural language processing techniques, we design a model that can find the patterns and structures within the essays. This model is created to explore the essays, discern linguistic differences, and gain a deeper understanding of the writing proficiency displayed by candidates.
A pivotal component of our model architecture involves the utilization of a two-layered Recurrent Neural Network (RNN). This architectural choice allows us to capture temporal dependencies within the essays, identifying how ideas evolve and connect throughout the text. Furthermore, we employ bidirectional RNNs to enhance our model's ability to grasp context, ensuring that it considers both past and future words when making predictions. This bidirectional approach is useful in achieving a comprehensive understanding of the essays' textual flow and coherence.
In conclusion, our goal is to use the power of text mining models to analyze the set of features and predict CEFR scores for the essays with high accuracy and minimal loss. Through data preprocessing, embeddings, and the construction of a robust neural network architecture, we aim to find patterns within the essays, using language proficiency as an evaluation.

DATA EXPLORATION

Before diving into the preprocessing of this dataset, we initiated the data preparation process by strategically dropping columns with null values. Specifically, we removed the columns 'Examiner_Comment,' 'Task_Response,' 'Coherence_Cohesion,' 'Lexical_Resource,' and 'Range_Accuracy.' These columns, while potentially valuable, contained missing data that could interfere with our initial analyses. In this way, we ensured that the subsequent preprocessing steps would be carried out on a more focused set of data.

We explored a range of only essential features to gain valuable insights into the data. These features are task type, questions, essays, and overall scores. Task type distribution analysis reveals that Task 1 and Task 2 are equally distributed components of the dataset, with Task 1 constituting 44.7% and Task 2 accounting for 55.3%.

Additionally, we analyse the distribution of overall scores, which are critical in the context of language assessment. Our findings indicate that scores of 5.5, 6.0, 6.5, and 7.0 are the most prevalent, with these scores representing the majority of the dataset. This allowed us to assess the prevalence of different proficiency levels among the essay submissions, providing valuable context for our next analyses.
Furthermore, we present a conversion table that facilitates the translation of IELTS scores to CEFR (Common European Framework of Reference for Languages) scores. This conversion is fundamental for understanding the alignment between IELTS and CEFR frameworks, allowing for more standardized language proficiency evaluations.
Lastly, we examine the distribution of CEFR levels among the dataset, revealing that the majority of submissions fall into the B2 and C1 levels. This observation highlights the proficiency levels of the test-takers, providing information for language assessment.

Additionally, we have decided to include extra features, such as the number of missing words to meet the length requirement (where essays must be at least 250 words), mean words per sentence, vocabulary richness, readability scores (using Flesch-Kincaid Grade Level and Gunning Fog Index), usage of transitional words, and evaluations of grammar and spelling errors. These additions significantly enhance the dataset's value. Mean sentence length provides insights into essay complexity and structure, revealing candidates' writing styles and proficiency levels. Vocabulary richness quantifies linguistic diversity, offering deeper insights into language capabilities. Readability scores assess essay accessibility and comprehensibility, extending language assessment beyond vocabulary and grammar. The analysis of transitional words reveals essay coherence and logical flow, enriching our understanding of writing skills. Lastly, grammar and spelling error assessments offer valuable indicators of language proficiency and an additional dimension for evaluation. These feature enhancements provide the dataset with greater depth, enabling more comprehensive text mining and analysis, ultimately improving the precision and interpretability of our results.

The results of our analysis provide valuable insights into the diverse characteristics of the essays within our dataset. Firstly, we observed that some essays are notably short of the required 250-word minimum, leading to varying numbers of missing words across the collection. Moreover, we found variations in mean sentence lengths, indicating differences in the complexity and structure of the essays, with values ranging between 20 and 25.
Additionally, our examination of vocabulary richness revealed a wide range, with unique word counts spanning from 52 to 138, highlighting the diversity in vocabulary usage among the essay writers.
Furthermore, our assessment of readability through the Flesch-Kincaid Grade Level and Gunning Fog Index scores identified reading difficulty levels within the essays. The Flesch-Kincaid scores ranged from approximately 6.8 to 11.9, reflecting varying linguistic complexity, while the Gunning Fog Index scores clustered mostly between 8.02 and 12.94, demonstrating diversity in the essays' readability levels.
In our exploration of transitional words, we discovered distinct usage patterns. For additive transitions, counts varied from 0 to 4, indicating differences in how writers incorporated additive elements to enhance argument flow. Adversative transitions exhibited counts from 0 to 2, suggesting variations in employing contrasting elements to present opposing viewpoints. Causal transitions spanned from 0 to 5, indicating differing degrees of causal relationship exploration. Lastly, sequential transitions ranged from 0 to 2, revealing variations in the use of sequential elements.

Overall, these results underscore the dataset's richness and diversity, laying a strong foundation for text mining, analysis, and the development of machine learning models for essay evaluation.

DATA PREPROCESSING

Preprocessing is the initial phase of data preparation in which raw data is cleaned, transformed, and organized to make it suitable for analysis. This involves tasks like handling missing values, removing noise, standardizing formats, and extracting relevant features, all with the goal of improving data quality and usability for subsequent analytical tasks.

In this preprocessing phase, we are preparing the IELTS Essay Dataset for advanced analysis. The process involves several key steps:

- **Downloading Resources:** downloading essential resources for NLP (stopwords, tokenization, and lemmatization tools, using the NLTK library).
- **Stopword Removal and Lemmatization**: remove common English stopwords and perform lemmatization on the dataset (eliminate noise and standardize words to base form).
- **Text Cleaning**: remove special characters and punctuation marks, and all text converted to lowercase (uniformity and consistency in the dataset).
- **Tokenization**: breaking it into individual words or tokens (easy to work with).
- **Statistical Analysis**: assess the dataset's characteristics, such as, calculating the maximum and mean lengths of questions and essays before and after preprocessing providing insights into the dataset's structure and distribution.
- **Visualization**: visualize the statistics in a bar chart comparing the maximum and mean lengths of questions and essays both before and after preprocessing.
- **Dataset standardization & vocabulary construction**: collect unique words in a list from the preprocessed essays and transform it into a NumPy array excluding duplicate entries. Next, establish a word-to-index mapping, creating a dictionary. The resulting vocabulary size, including a special *<UNK>* token, to reach minimum of 250 words.

The preprocessing steps are vital in ensuring that the dataset is cleaned, standardized, and ready for advanced text mining and analysis techniques. This initial data preparation sets the stage for extracting meaningful insights from the IELTS Essay Dataset.

VECTOR REPRESENTATION

In this section, we will explore the techniques of Positive Pointwise Mutual Information (PPMI), Term Frequency-Inverse Document Frequency (TF-IDF), and Cosine Similarity. These methods play an essential role in transforming the textual data within the dataset into numerical vectors, enabling quantitative analysis and modeling. Additionally, we will introduce GloVe (Global Vectors for Word Representation) embeddings, a sophisticated approach for capturing semantic relationships between words.

Positive Pointwise Mutual Information (PPMI)

In the process of analyzing this dataset, we aim to extract meaningful information about word associations and semantic relationships among the words used in the essays. One powerful technique for this is the Positive Pointwise Mutual Information (PPMI) matrix. The PPMI matrix helps us understand how likely two words are to co-occur together in the same context, providing insights into word semantics.

A table is created where rows represent words in our vocabulary, and columns represent documents (in this case, essays) storing co-occurrence counts of words within a specified window size. For each word, it identifies its context words within a defined window size. Co-occurrence counts are accumulated in the term-document matrix, indicating how many times words appear together in the same context within the specified window. In order to save memory and computation time we have converted the co-occurrence matrix in a Compressed Sparse matrix. Subsequently, we have calculated PMI scores for each word pair in the co-occurrence matrix measuring the statistical dependency between two words. The final output consists of a PPMI matrix (negative values mapped to 0) reflecting the strength of association between words in the dataset. Higher values signify stronger associations, while lower or negative values suggest weaker or no associations.

Term Frequency-Inverse Document Frequency (TF-IDF)

TF-IDF is a fundamental method that allows us to understand the importance of words within a collection of essays. This technique uncovers valuable insights from our dataset, shedding light on the significance of words in assessing language proficiency and essay scoring.

We have initialized a TF-IDF vectorizer and applied it to our preprocessed corpus of essays. Next, the TF-IDF matrix is generated, where each row represents an essay, and each column corresponds to a unique word from the essays' vocabulary. This matrix quantifies how relevant each word is to each essay, providing valuable insights into word significance within the dataset. Additionally, we retrieved the feature names, which represent the unique words considered during the analysis. By executing this code, we gain a numerical representation of word importance that contributes significantly to our understanding of language proficiency assessment and the factors influencing essay scoring in the IELTS exam.

The resulting matrix is (1435, 11750), indicating that there are 1,435 essays in the dataset, and the TF-IDF analysis considers 11,750 unique words from these essays. It is possible to observe that many scores of 0 appear, meaning that the word doesn't appear in the corresponding essay. Also, the "Feature Names (Vocabulary)" is printed showing examples of unique words (vocabulary) considered during the analysis, such as 'abandon,' 'abandoned,' 'abandoning,' 'zoom,' 'zouns,' and ‘zte.'

GLOVE EMBEDDINGS

GloVe (Global Vectors for Word Representation) embeddings are pre-trained word vectors used to understand word semantics. They capture how words relate to each other in a text corpus and are very important for various natural language processing tasks.
We have loaded and used GloVe word embeddings where by iterating through each line in the file performs a splitting in word and vector components. These word vectors are stored in a dictionary where words serve as keys, and their corresponding vectors as values. Subsequently, we construct an embedding matrix taking three parameters: the loaded GloVe word embeddings, the embedding dimension (100, 200, or 300), and the vocabulary size (the number of unique words in the dataset). The embedding dimension impacts the output level of accuracy, a higher dimension captures more information but requires more memory. This function initializes an empty matrix with dimensions corresponding to the vocabulary size and embedding dimension. It then iterates through each word in the dataset's vocabulary, checks if its GloVe embedding exists, and if so, assigns the corresponding vector to the matrix. This process creates an embedding matrix that can be used to initialize word embeddings in neural network models. In this way, three different embedding matrices are created with dimensions of 100, 200, and 300, respectively, corresponding to different GloVe embedding dimensions.

DATA TRANSFORMATION AND LABEL ENCODING

At this point, we have transformed our dataset for machine learning analysis by converting the textual data into numerical form. We iterate through essays with uniform lengths, and assigns a unique integer to each word, using a mapping established during preprocessing. At this point, the resulting list represents the entire dataset in numerical format.
The next step is to arrange the dataset for model training. It combines various features, including task type, missing words, unique words, grade level, linguistic features, and grammar and spelling errors, into a comprehensive numerical representation. This array of features, along with label encoding for the overall scores, sets the stage for machine learning. The resulting dataset, with 1435 samples and 324 features, prepares the data for model building and evaluation. This data transformation and encoding are essential steps in our analysis, enabling us to apply machine learning techniques to predict and analyze IELTS scores effectively.

Train, Test and Validation Split

In the context of preparing this dataset for machine learning analysis, a valuable step involves splitting the dataset into distinct subsets for training, testing, and validation. This process assesses the performance of machine learning models accurately. Initially, the dataset is divided into training and test sets, with 90% of the data allocated for training and 10% for testing. This split allows for model training and preliminary evaluation. Subsequently, the training data is further divided into training and validation sets, again with a 90%-10% ratio. The validation set is crucial for fine-tuning models and optimizing, ensuring that the final trained model generalizes well to unseen data. This division of the dataset sets the foundation for robust machine learning experiments, enabling the assessment of model performance with a clear distinction between training, validation, and testing phases.

NEURAL NETWORK ARCHITECTURE CREATION

Neural networks have demonstrated remarkable capabilities in tasks such as image recognition, natural language processing, and decision making. In this context, we explore the design of a specific neural network architecture adjusted to the task of essay rating. This architecture leverages bidirectional recurrent layers to capture text data's sequential information and combines it with task-specific features to predict essay ratings.
We have designed a neural network architecture creation function, taking several parameters to configure it, such as input shape, output shape, embedding matrix, and various activation functions.
We have initially separated with a Lambda layer the input data into two parts, representing text data and numerical data. Next after that the text has passed through the embedding layer and words are mapped to dense vectors, the embedding matrix is initialized with pre-trained GloVe word embeddings and remains non-trainable during model training. The text data then enters two bidirectional GRU (Gated Recurrent Unit) layers. These recurrent layers capture sequential information from the text data in both forward and backward directions. The output sequences from these layers are returned for further processing. These are flattened and concatenated with numerical input data allowing the model to consider simultaneously textual content and the numerical information. By now two dense layers are involved, each of which is followed by batch normalization to stabilize training and dropout layers to prevent overfitting. These layers apply non-linear transformations to the concatenated data. The final output layer consists of five units, corresponding to the potential ratings for the essays. The activation function used is typically softmax for multi-class classification tasks.
Lastly, the model is created and compiled with a specified loss function, optimizer, and accuracy as a metric.

HYPERPARAMETER TUNING

In order to build a robust and high-performing neural network for essay rating, hyperparameter tuning plays a key role. It involves systematically optimizing the various settings and configurations of the model to achieve the best possible performance.

RNN and NN Activation Functions

To fine-tune our neural network we utilized GridSearchCV technique by systematically exploring a range of hyperparameter values, helping to identify the optimal combination that maximizes performance, ensuring the most accurate and efficient essay rating. Our model was designed to operate for 3 epochs, and we initiated the model with specific hyperparameters, including the embedding matrix and the choice of the tanh activation function for the GRU (Gated Recurrent Unit) layer.
To effectively explore the impact of different activation functions on the performance of the model, we used a list of three commonly used ones: 'relu,' 'sigmoid,' and 'tanh.' The GridSearchCV procedure was then executed, evaluating the efficiency for each activation function for a specific hidden layer within the model. This rigorous grid search was initialized with the training data, with the primary objective of discovering the most optimal combination of hyperparameters that would yield the highest accuracy on the training dataset.
Following the completion of the grid search, we identified the configuration that achieved the highest accuracy as the best result. Specifically, the 'relu' activation function was found to be the most effective, delivering an accuracy of approximately 44.36% on the training data. In

contrast, the 'sigmoid' activation function exhibited significantly lower performance, with an accuracy of approximately 19.48%, while 'tanh' displayed intermediate performance at around 41.27%.
During the second epoch of our hyperparameter tuning process, we observed significant improvements in model performance. The loss decreased notably from its initial value of 44.3780 in the first epoch to 21.2596. Simultaneously, the model's accuracy increased substantially from 30.92% to 49.61%. These metrics demonstrate the model's progressive learning and improvement as it undergoes training.

RNN and NN Units
Also for the RNN units and NN units we could observe an increasing trend in accuracy and decreasing over loss with the proceeding of the two epochs, indicating that the model is learning and improving its performance over time. In the context of our RNN units, like GRU, an increase in accuracy and a decrease in loss indicate that the model is effectively capturing sequential dependencies and patterns in the data. As the training progresses, the RNN units learn to remember and utilize information from previous time steps to make better predictions. Whereas considering NN Units, an increase in accuracy and a decrease in loss indicate that the model is learning to extract meaningful features from the input data and is making better decisions based on those features.

Learning Rate and Dropout Rate
In order to optimize hyperparameters for our neural network model designed for essay rating, we focused on the learning rate and dropout rate. The learning rate is a critical hyperparameter that controls the step size during gradient descent, impacting the speed and stability of training. We explored three learning rates: 0.001, 0.01, and 0.1. Notably, the learning rate of 0.1 emerged as the most effective choice, with the highest accuracy at approximately 49.70%. This result suggests that a higher learning rate accelerated convergence and enabled the model to find a more favorable configuration faster, thus enhancing its predictive accuracy. Furthermore, dropout rate, a regularization technique to prevent overfitting, was investigated using dropout rates of 0.0, 0.1, and 0.2. The dropout rate of 0.2 exhibited the best performance, achieving an accuracy of approximately 48.75%. This finding highlights the importance of dropout regularization in promoting model generalization and preventing overfitting, ultimately leading to improved accuracy.

Kernel Regularizer and Initializer
In our exploration of hyperparameters, we explored the effects of kernel regularizers and kernel initializers on our neural network model's performance. In the case of kernel regularizers, we tested various types, including 'l1,' 'l2,' 'l1_l2,' and 'None.' Interestingly, the absence of kernel regularization ('None') yielded the highest accuracy of approximately 49.35%. This suggests that, for our specific essay rating task, avoiding additional weight penalties associated with regularization contributed to better model performance. Conversely, the addition of 'l1' or 'l2' regularization, or a combination of both ('l1_l2'), led to slightly reduced accuracies, indicating that in this context, they did not provide significant benefits. For kernel initializers, we compared 'glorot_uniform' and 'he_normal.' Here, 'he_normal' initialization achieved the highest accuracy at approximately 45.82%, outperforming 'glorot_uniform.' These results highlight the influence of kernel initialization techniques in setting the initial weights of the neural network, with 'he_normal' proving to be more effective in our essay rating task.

Optimization

Through a grid search optimization process, we systematically explored the impact of two different optimizers, 'adam' and 'sgd,' on the model's performance. The results reveal that the 'adam' optimizer outperformed 'sgd,' achieving a better accuracy score of approximately 34.38% on the training data compared to 26.44%. This highlights the importance of selecting the right optimizer for training our neural network, with 'adam' proving to be the more suitable choice in this context.

Batch Size

Next, we focus on the influence of batch size, a key hyperparameter, on our neural network model designed for essay rating. Our model configuration includes various hyperparameters such as activation functions, units in different layers, learning rate, dropout rate, kernel initializer, and optimizer. Through a grid search optimization process, we systematically explored the impact of different batch sizes (64, None, 16, and 32) on the model's performance. The results reveal that a batch size of 32 led to the highest accuracy, approximately 47.89%, on the training data, outperforming other batch sizes significantly. This finding emphasizes the importance of selecting an appropriate batch size, as it can significantly impact training efficiency and model effectiveness.

ANALYZING ESSAY CLASSIFICATION MODELS

Model evaluation is a crucial step in assessing the performance of different natural language processing models. In our analysis, we compared the performance of three distinct models: GloVe, TF-IDF, PPMI, and BERT, each designed for the task of essay rating.

The GloVe Model

Firstly, we will discuss the GloVe model, which underwent extensive training over 15 epochs. The GloVe model achieved notable results in terms of accuracy and loss. The training accuracy steadily increased, reaching approximately 96.38% by the final epoch, demonstrating the model's ability to effectively learn from the training data. Conversely, the validation accuracy exhibited fluctuations but settled around 51.54%, indicating that the model maintained a reasonable level of generalization.

The loss metrics followed a similar pattern. The training loss decreased consistently throughout the training process, reflecting the model's capacity to minimize errors during training. The validation loss, however, displayed fluctuations but eventually stabilized, indicating a degree of generalization. While the GloVe model demonstrated a strong ability to learn from the data, its performance on the validation set suggested a moderate level of generalization.

TF-IDF Model

The TF-IDF model was also trained and evaluated over 15 epochs to assess its performance comprehensively. The training process exhibited a notable improvement, with the training loss steadily decreasing and the accuracy increasing.

On the other hand, the model's performance on the validation data showed a different trend. While the validation loss started at 3.4452 with an accuracy of 50%, it fluctuated throughout training and eventually increased to 4.1833 with an accuracy of 47.69% in the final epoch. This discrepancy between training and validation results suggests a potential issue of

overfitting, where the model was fitting the training data too closely but struggled to generalize to unseen data. Overall, these results highlight the need for further model adjustment to address overfitting and improve the model's ability to make accurate predictions on new essays.

The PPMI Model

The PPMI model was created with specific hyperparameters, including GRU activation functions, units, learning rate, dropout rate, kernel regularizer, and optimizer settings. It was trained over 15 epochs. During training, the model's performance showed steady improvement, with the training loss consistently decreasing and accuracy increasing. However, when evaluating the model on the validation data, a different trend emerged. The validation loss started at 4.3845 with an accuracy of 36.92% in the first epoch but fluctuated throughout training. In the final epoch, the validation loss increased to 3.8045 with a validation accuracy of 45.38%. This discrepancy between training and validation results suggests potential overfitting, where the model may have been fitting the training data too closely, making it less effective at generalizing to unseen data.

BERT Model

To evaluate and analyze the neural network models for essay rating, we focus on the Bert model, a transformer based model. This model has a high potential in this context especially for its impressive language understanding capabilities.
The results of the BERT model over 15 epochs show a general trend where the model gradually learns and improves its ability to predict CEFR levels based on essay content. The accuracy increases indicating the model's capacity to make more precise predictions, while the declining loss suggesting that it's minimizing errors effectively. We can see that the training process converges over time as indicated by the decreasing training loss. This convergence is a positive sign, suggesting that the model is learning from the training data. The accuracy increased from an initial 37.98% to a final accuracy of 49.61%. In parallel, the loss decreased from an initial 1.7295 to a final loss of 1.1783. This trend indicates great learning and adaptation by the model, improving its language proficiency assessment. While the fluctuations observed during training are normal, the overall trajectory demonstrates the model's capacity to enhance its predictive accuracy and reduce errors effectively.

MODEL EVALUATION

We have decided to visualize and compare the results provided from the performance of the different classifiers on our dataset. Three classifiers were evaluated in our analysis: GloVe, TF-IDF and PPMI. The objective was to assess their effectiveness in classifying essays into proficiency levels and to identify potential trends and challenges in their performance.
By comparing with bar charts the loss values and accuracy scores for each classifier showed that GloVe and PPMI exhibited higher accuracy than others.
Next, we examined the training and validation accuracy for each classifier over the course of training. The line plots displayed the evolution of accuracy as the number of training epochs increased showing how well each classifier learned from the training data and its ability to generalize to unseen validation data. We observed diverging trends between training and validation accuracy for some classifiers, indicating potential overfitting issues especially in the TF-IDF.
Lastly, we presented a confusion matrix that evaluated the performance of the GloVe model, the one that appeared more suitable for the analysis of our dataset. The confusion matrix

provided a detailed breakdown of the classifier's ability to correctly classify essays into different proficiency levels (A, B1, B2, C1, C2).
In summary, this analysis represents a comprehensive view of the performance of various classifiers on the essay dataset. In conclusion, based on our dataset and evaluation results, the GloVe model emerges as the more suitable choice for our specific text processing needs.

CONCLUSION

In conclusion, this report has presented the analysis and assessment of essays using natural language processing techniques and machine learning models.
Our initial data preparation involved column selection and data cleaning to ensure that the dataset was ready for analysis. We examined important features, including task type, questions, essays, and overall scores, shedding light on the distribution of these critical attributes. Moreover, we integrated our dataset with additional features, such as the number of missing words, mean sentence length, vocabulary richness, readability scores, usage of transitional words, and evaluations of grammar and spelling errors. These features enabled more comprehensive text mining and analysis.
Our model selection included the implementation of a GloVe model, a TF-IDF model, PPMI model and a BERT model, discussing the model architectures and optimization techniques, along with training progress and validation accuracy. The model that appeared more suitable for this type of task is the GloVe model, a word embedding technique that captures semantic relationships between words in a continuous vector space.
Overall, our model development process resulted in positive outcomes. We observed improvements in accuracy and reductions in loss, indicating that our models are learning and predicting more effectively.