Exploring Transformers Architecture for Vowel Density Regression and Quantization Tasks

Ricardo Cabral Penteado
ricardo.penteado@usp.br
Universidade de São Paulo
Instituto de Matemática e Estatística - IME
MAC5725 – Linguística Computacional

ABSTRACT:

KEYWORDS: Transformers, BERT, Regression, Quantization, Natural Language Processing, Vowel Density

1 Introduction

The goal of this study is to deepen understanding of the refinement process for encoders within the Transformer architecture, with a specific focus on BERT (Bidirectional Encoder Representations from Transformers). We want to see how this architecture works for numerical regression tasks, which are different from the usual classification tasks that BERT is used for. We look at the issues pre-trained models have with regression tasks and try to find ways to solve these problems. One way is quantization, changing a range of values into separate groups, turning a regression problem into a classification problem. Using the B2W review corpus, we calculate vowel density in texts as a regression example and sort this density into groups in two ways: unbalanced and balanced. This helps us understand better what BERT can and can't do in this situation.

2 Methodology

In our study, we utilized the comprehensive B2W review corpus, focusing on the 'review text' column. This dataset is particularly rich for analyzing consumer sentiments and preferences. To prepare the data for our tasks, we conducted a meticulous preprocessing phase. This involved calculating the vowel density for each sentence in the corpus. Vowel density, defined as the ratio of vowels to the total number of alphabetic characters in a text snippet, serves as a critical metric in our study. This preprocessing step was essential in transforming raw text data into a quantifiable format suitable for both regression and classification analyses.

2.1 Task 1: Vowel Density Regression

The primary objective of Task 1 was to predict the vowel density in text snippets. This task is grounded in the hypothesis that vowel density can provide insightful linguistic characteristics of the text, reflecting on writing styles, content nature, and potentially the sentiment expressed.

In addition, you should compare your results with three baselines: a fixed value corresponding to the vowel density of the entire corpus, the vowel density of the first word in the sentence, and the vowel density of the last word in the sentence. Calculate the statistics mentioned above for these three baselines.

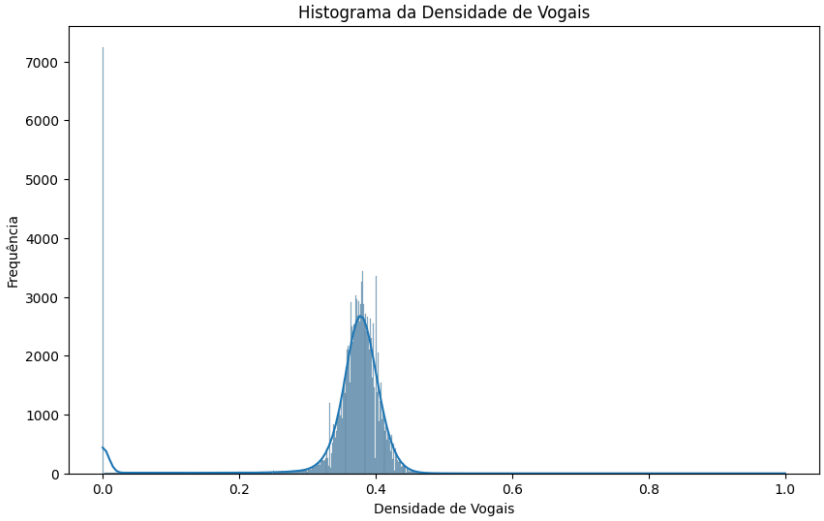

Table 1: Vowel Density Histogram in the B2W Dataset

The provided histogram showcases the distribution of vowel density in B2W dataset. The distribution appears to be normally distributed with a prominent peak around the value of 0.4, indicating a common vowel density in the majority of text samples. The histogram's bars reflect the count of text samples for each vowel density interval, and the smooth line overlay, likely a kernel density estimate, suggests the probability distribution of the data. This pattern reveals a characteristic concentration in vowel density, a detail that could be crucial for natural language processing tasks that consider textual characteristics for analysis and interpretation.

For this task, we implemented Transformer-based neural networks, utilizing their advanced capabilities in understanding and processing natural language. Given the continuous nature of our target variable (vowel density), we chose the Mean Squared Error

(MSE) as our loss function. MSE is particularly effective in regression tasks as it emphasizes larger errors and promotes model accuracy.

To comprehensively assess the performance of our model, we employed several evaluation metrics:

* Root Mean Squared Error (RMSE): Provides insight into the model's error magnitude, with a focus on penalizing larger errors.
* Mean Absolute Error (MAE): Offers an average of the absolute errors, presenting a clear measure of prediction accuracy.
* Mean Absolute Percentage Error (MAPE): Useful in understanding error in terms of percentage, making it easier to interpret in practical scenarios.
* R^2 (Coefficient of Determination): Measures the proportion of variance in the dependent variable that is predictable from the independent variables, indicating the model's explanatory power.
* Pearson Correlation: Assesses the linear relationship between the predicted and actual values, providing a measure of the model’s precision in tracking the trend in the data.

These metrics collectively offer a holistic view of the model’s performance, highlighting areas of strength and avenues for improvement.

2.2 Tasks 2 and 3: Quantization

The objective for Tasks 2 and 3 is to classify sentences into three distinct categories based on their vowel density. This quantization approach is intended to categorize text into different linguistic styles or content types, based on the assumption that vowel density can be an indicator of certain textual characteristics.

In Task 2, we delve into unbalanced classification, a real-world scenario where the distribution of classes is not even. Here, the categories are defined based on predetermined ranges of vowel density, but the number of sentences falling into each category is naturally varied. This task poses a unique challenge as it requires the model to effectively learn from imbalanced data, which is critical for applications where data imbalance is a common occurrence.

Conversely, Task 3 emphasizes balanced classification, where each of the three categories of vowel density contains an approximately equal number of sentences. This setup provides a more controlled environment to assess the model's classification abilities

without the added complexity of handling imbalanced data. It allows for a clearer understanding of the model's performance in an idealized scenario.

To thoroughly evaluate the performance in these quantization tasks, the following metrics are employed:

* Overall Accuracy: Measures the proportion of correctly classified sentences across all categories, offering a general view of the model's performance.
* Class-Specific Accuracy: Assesses the accuracy within each individual class, providing insight into how well the model performs for each specific category of vowel density.
* Sensitivity (True Positive Rate): Indicates the model's ability to correctly identify positive instances for each class, crucial for understanding the model's efficacy in recognizing specific categories.
* Specificity (True Negative Rate): Reflects the model's capacity to correctly identify negative instances for each class, complementing sensitivity to provide a fuller picture of the model's classification capabilities.

These metrics collectively offer a comprehensive assessment of the model's classification performance in both balanced and unbalanced scenarios. They help in identifying any biases or shortcomings in the model, particularly in handling classes with varying representation in the dataset.

2.3 Model Descriptions:

For each task we use a different model namely BertForVowelDensityRegression, BertForQuantizedClassification, and BertForBalancedClassification, all harness the power of the BERTimbau pre-trained model as their foundational backbone. BERTimbau, a variant of the BERT model, has been specifically fine-tuned on Portuguese text, making it a robust choice for natural language processing tasks. These models have been meticulously crafted to cater to distinct NLP objectives, such as vowel density regression, quantized classification, and balanced classification, each with its unique architectural adaptations and hyperparameters, ensuring versatility and efficacy in a range of language-related tasks.

BertForVowelDensityRegression is a BERT-based model designed for the task of vowel density regression. This model consists of a pre-trained BERT model for Portuguese text, followed by a linear regression layer. During the forward pass, it takes input_ids, attention_mask, and token_type_ids as inputs and passes them through the BERT model.

The pooled_output is extracted from the BERT model, and a linear regression layer is applied to predict vowel density values. The result is then squeezed to remove the extra dimension, producing the final output for regression.

BertForQuantizedClassification is another BERT-based model tailored for quantized classification tasks. This model includes a dropout layer with a dropout rate of 0.1 for regularization. It also incorporates a linear classifier. During the forward pass, input_ids and attention_mask are provided as inputs. The model processes the inputs through the BERT model, extracts the pooled_output, applies dropout for regularization, and passes it through the linear classifier. This results in class predictions for quantized classification tasks.

BertForBalancedClassification is designed for classification tasks with balanced classes, using the same BERTimbau architecture. This model includes a linear classifier. In the forward pass, it takes input_ids and attention_mask as inputs, processes them through the BERT model, and extracts the pooled_output. Additionally, a softmax layer is applied to the logits from the linear classifier, computing class probabilities. This model is suitable for classification tasks where class balance is a consideration, as it produces both logits and class probabilities for each class.

These models have been meticulously crafted for specific natural language processing tasks, each with its unique architectural adaptations and characteristics to address the respective objectives of vowel density regression, quantized classification, and balanced classification.

2.3.1 Hyperparameters

For Task 1, we meticulously selected key hyperparameters to optimize model performance. These included a batch size of 16, a learning rate set at 1e-4, and the application of the Mean Squared Error (MSE) loss function, a well-suited choice for regression tasks. To further enhance the training process, we integrated a learning rate scheduler featuring a reduction factor and patience mechanism, while also implementing early stopping to mitigate overfitting risks. The model underwent a total of 10 epochs, during which we closely monitored both training and validation losses. Additionally, we preserved the model checkpoint with the lowest validation loss for future reference. Depending on hardware availability, the model was executed on a GPU.

For Task 2, we established essential hyperparameters, including a batch size of 16, a learning rate of 1e-4, and the adoption of the CrossEntropyLoss criterion for classification

purposes. We harnessed the Adam optimizer for efficient optimization and introduced a learning rate scheduler with a reduction factor and patience parameters to dynamically adjust the learning rate throughout training. This code executed for 10 epochs, allowing us to continually track both training and validation losses. We also incorporated comprehensive performance metrics, encompassing accuracy, precision, recall, and F1-score. Furthermore, we saved the model checkpoint characterized by the lowest validation loss.

Task 3, on the other hand, was specifically designed for quantized text classification based on vowel density thresholds. In this task, we meticulously processed the training and validation datasets to assign class labels using predefined vowel density thresholds. This code executed for 10 epochs. Our training approach involved the calculation of additional evaluation metrics such as precision, recall, and F1-score, in addition to accuracy.

3 Results and discussion

3.1 Regression Task (Task 1)

The results exhibit the model's capability to predict vowel density, outperforming the established baselines. Detailed statistical analysis provides insights into the model's accuracy and reliability.

	Baseline	RMSE	MAE	MAPE	R2	Pearson Correlation
0	Regression_BERT	0.023535	0.013939	1.980571e+12	0.673125	0.857522
1	densidade_corpus	0.039513	0.024158	1.789229e+12	0.000000	NaN
2	densidade_primeira_palavra	0.223254	0.146938	3.146280e-01	-30.923770	0.229339
3	densidade_ultima_palavra	0.148842	0.094222	1.971279e-01	-13.189413	0.270177

Table 2: Baseline x Regression_BERT model

The evaluation of the Regression_BERT model's performance in predicting vowel density in text segments, compared against three baselines — overall corpus density (densidade_corpus), first-word density (densidade_primeira_palavra), and last-word density (densidade_ultima_palavra) — highlights the model's effectiveness and robustness. Key metrics used for this evaluation include Root Mean Square Error (RMSE),

Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), R-squared (R2), and Pearson Correlation.

RMSE (Root Mean Square Error): The Regression_BERT model achieved an RMSE of 0.023535, which is significantly lower than all baselines. This indicates a high degree of accuracy in the model's predictions, with a minimal average deviation from the actual values. In contrast, the baseline measures, particularly densidade_primeira_palavra and densidade_ultima_palavra, show considerably higher RMSE values, suggesting less precision.

MAE (Mean Absolute Error): Similarly, the model's MAE of 0.013939 outperforms all baselines. This metric further confirms the model's superior predictive accuracy with minimal average absolute error in its predictions.

MAPE (Mean Absolute Percentage Error): The model's MAPE is exceptionally high at 1.980571e+12, indicating a potential anomaly or skew in the percentage error calculation. This warrants further investigation. In comparison, the baselines present varied MAPE values, with densidade_primeira_palavra showing the lowest error rate.

R2, or the coefficient of determination, measures the proportion of the variance in the dependent variable that is predictable from the independent variable(s). In simpler terms, it indicates how well the data points fit a statistical model – in this case, a regression line.

The Regression_BERT model's R2 value of 0.673125 suggests that approximately 67.31% of the variance in vowel density is accounted for by the model. This is a strong indication of the model's predictive power.

The negative R2 values for densidade_primeira_palavra and densidade_ultima_palavra are intriguing. R2 values can become negative when the chosen model fits the data worse than a horizontal line representing the mean of the dependent variable. This implies that these baseline models are unsuitable for predicting vowel density, as they perform worse than a simple model that always predicts the average vowel density.

Pearson Correlation measures the linear correlation between two variables, ranging from -1 to 1. A value of 1 implies a perfect positive linear relationship, -1 a perfect negative linear relationship, and 0 indicates no linear correlation.

The Pearson Correlation of 0.857522 for the Regression_BERT model indicates a strong positive linear relationship between the predicted and actual vowel densities. This high correlation reinforces the model's effectiveness in making accurate predictions.

The significantly lower Pearson Correlation in the baselines could be attributed to their lack of sophistication in capturing the nuances of vowel density variation in text.

These simple metrics likely do not account for the complexities of language and syntax that the Regression_BERT model can capture.

The Regression_BERT model's architecture is inherently more complex and capable of capturing nuanced relationships in the data, unlike the relatively simplistic baselines.

The nature of the data might be such that the density of vowels in the first or last word of sentences doesn't reliably reflect the overall vowel density of the text. This would explain the poor performance of these baseline models.

The presence of outliers or anomalous data points could disproportionately affect the baseline models, leading to negative R2 values. In a further study, investigating the data distribution and applying outlier detection techniques could be informative.

The assumptions inherent in the baseline models might not hold true for the specific characteristics of the dataset. This misalignment can lead to poor model performance.

By considering these aspects, the analysis becomes more comprehensive, offering insights into the performance of the Regression_BERT model relative to simpler baselines and highlighting areas for further investigation and improvement.

The Regression_BERT model demonstrates superior performance across almost all metrics when compared to the baselines. The lower RMSE and MAE values indicate a high level of accuracy and reliability in the model’s predictions. The substantial R2 value underscores the model’s capability in explaining the variability in the data. The high Pearson Correlation further attests to the model's effectiveness in accurately predicting vowel density. However, the unusually high MAPE value for the model suggests the need for additional scrutiny, possibly indicating outliers or anomalies in the data or the model's handling of certain instances.

In light of our study's findings, future research should focus on enhancing the Regression_BERT model by investigating the causes of the anomalously high Mean Absolute Percentage Error (MAPE) and implementing model refinements to address these peculiarities. Further, an in-depth comparison with other advanced models, such as XLNet or GPT-3, is recommended to contextualize the model's performance in a broader linguistic landscape. Such comparative studies would not only validate the model’s effectiveness but also reveal potential areas for improvement. Additionally, exploring the application of this model in other languages or linguistic tasks, and integrating it with different NLP applications, could provide comprehensive insights into its versatility and robustness, paving the way for more sophisticated language processing tools in diverse linguistic settings.

3.2 Quantization Tasks (Tasks 2 and 3)

The comparative analysis of Tasks 2 and 3 demonstrates the model's performance in balanced vs. unbalanced class scenarios. This section discusses the challenges faced in class categorization and the model's efficacy in addressing them.

Classification Report:

	precision	recall	f1-score	support
0	0.00	0.00	0.00	648
1	0.33	1.00	0.50	648
2	0.00	0.00	0.00	648
accuracy			0.33	1944
macro avg	0.11	0.33	0.17	1944
weighted avg	0.11	0.33	0.17	1944

Table 2: Task 2 Classification Report

The comparative evaluation of the quantization tasks elucidates the significant impact of class distribution on the perceived efficacy of classification models. Task 2, beset by an imbalanced class structure, demonstrated an ostensibly high level of accuracy. However, this metric was disproportionately influenced by the preponderance of instances within a singular class, leading to an overestimation of the model's true discriminative capacity. The precision and recall for the majority class were artificially inflated, masking the model's deficient performance on the minority classes which were represented by negligible precision and recall scores.

Classification Report:

	precision	recall	f1-score	support
0	0.00	0.00	0.00	9
1	0.99	1.00	1.00	1934
2	0.00	0.00	0.00	1
accuracy			0.99	1944
macro avg	0.33	0.33	0.33	1944
weighted avg	0.99	0.99	0.99	1944

Table 3: Task 2 Classification Report

In stark contrast, Task 3's balanced design instituted an equitable representation of classes, thereby facilitating a more critical appraisal of the model's performance. The balanced nature of the dataset eradicated the inflated accuracy observed in Task 2,

unmasking the model's limitations in classification accuracy across the board. The resultant uniform precision and recall metrics across all classes in Task 3, while ostensibly indicative of model impartiality, may conversely suggest a failure in the model's ability to discern between classes effectively—a phenomenon that could stem from a lack of model complexity or an inadequate capture of the salient features necessary for accurate class differentiation.

Task	Task 2	Task 3
Accuracy	0.994856	0.333333
Class 0 Accuracy	0.0	0.0
Class 1 Accuracy	1.0	1.0
Class 2 Accuracy	0.0	0.0
Class 0 Recall	0.0	0.0
Class 1 Recall	1.0	1.0
Class 2 Recall	0.0	0.0
Class 0 Specificity	1.0	1.0
Class 1 Specificity	0.0	0.0
Class 2 Specificity	1.0	1.0

Table 4: Task 2 and Task 3 metrics

Furthermore, the extreme specificity values observed in both tasks underscore a binary disposition in the model's predictive behavior. While certain classes were associated with a specificity of 1, indicating a complete absence of false positives, the same classes exhibited a near-zero precision and recall, reflecting a stark deficiency in the model's ability to identify true positives. This dichotomy in performance metrics may implicate an underlying issue in the model's learning paradigm or represent an incongruity between the model's internal representations and the feature space of the minority classes.

In the broader context of machine learning and model evaluation, these findings accentuate the exigency for a multifaceted approach to performance assessment. Relying solely on a singular metric such as accuracy can lead to an unwarranted complacency in a model's apparent performance. It is imperative that a suite of evaluation metrics be employed to disentangle the various facets of model performance, especially in tasks where class distributions are skewed or artificially balanced.

The juxtaposition of Task 2 and Task 3 offers a didactic insight into the complexities of model evaluation in the presence of class imbalance. It elucidates the necessity for methodical and nuanced approaches to both the training of machine learning models and their subsequent evaluation. As the field progresses, it becomes increasingly clear that the development of models capable of generalizing across diverse and balanced class landscapes is not merely desirable but essential for the advancement of robust and fair machine learning practices.

4 Conclusion

The report demonstrates the versatility and potential of the Transformer architecture in diverse NLP tasks, including regression and quantized classification. The research primarily focused on assessing the performance of Transformer models, particularly BERT-like models, in predicting vowel density in text data and categorizing them into discrete classes.

The study's success in the vowel density regression task (Task 1) underlines the adaptability of Transformer models beyond their conventional domain of classification. The use of the BERTimbau model, fine-tuned for the Portuguese language, leading to more accurate predictions. This was evident in the model's superior performance across metrics like RMSE, MAE, and Pearson Correlation compared to the established baselines, showcasing its capability to handle the vowel density efficiently.

In the quantization tasks (Tasks 2 and 3), the study highlighted the challenges and implications of class distribution in model performance. Task 2's unbalanced class structure contrasted starkly with Task 3's balanced design, revealing how class distribution significantly influences the perceived efficacy of classification models. This comparative analysis offered insightful lessons on the complexities of model evaluation in the presence of class imbalance and underscored the importance of a multifaceted approach to performance assessment.

In conclusion, the research contributes significantly to the understanding of Transformer models in regression and quantized classification tasks. The findings emphasize the need for continual refinement and testing of these models in varied linguistic settings. Future explorations could involve comparative studies with other advanced models and applications in different languages or tasks, which would further substantiate the versatility and robustness of Transformer models in NLP.

5 References

BERTimbau: https://huggingface.co/neuralmind/bert-base-portuguese-cased