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: This study explores the application of the Transformers architecture, specifically BERT-like models, in tasks beyond its usual domain of classification - focusing on numerical regression and quantized classification. The primary objective is to assess the performance of these models in predicting vowel density in text data and categorizing them into discrete classes. We utilize a corpus from B2W reviews, focusing on the 'review text' column, and adapt the BERTimbau model for our tasks.

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

1 Introdução

The Transformers architecture has gained prominence in various Natural Language Processing (NLP) tasks, predominantly in classification. However, its effectiveness in regression tasks remains less explored. This study aims to bridge this gap by applying Transformers to predict vowel density in text, a numerical regression task, and further extend its application to quantized classification.

The motivation stems from the typical training of Transformer models on classification tasks, which might limit their efficiency in regression scenarios. This research aims to explore and potentially improve the performance of these models in regression tasks.

Recent developments in the field of machine learning have seen the versatile Transformers architecture being adapted for regression tasks, a domain traditionally dominated by classification models. A groundbreaking approach in this direction is the Regression Transformer (RT). Transformers, predominantly used in Natural Language Processing (NLP), are now being adapted for regression tasks, such as predicting continuous properties in scientific domains. This method, that was introduced by Born and Manica (2023), abstracts regression as a conditional sequence modeling problem. This approach not only matches but occasionally surpasses traditional regression models in predicting properties of small molecules and proteins. It also demonstrates its utility in chemical reactions, showcasing the Transformer's capability in handling both predictive

and generative tasks efficiently. The RT's adaptability in seamlessly bridging sequence regression and conditional sequence generation marks a significant advancement in the field, indicating a promising direction for multitask language models in scientific research and material design.

On other hand, The advancement in Natural Language Processing (NLP) through Transformer models like BERT has brought about the need for efficient computational methods, especially in tasks requiring quantized classification. Recent studies have shown significant progress in this area. The development of Q8BERT, a quantized 8Bit BERT model, presents an innovative approach to maintain model accuracy while significantly reducing computational requirements. This model achieves a 4x compression rate compared to the standard BERT model with minimal accuracy loss (Zafrir et al., 2019).

Moreover, the implementation of hardware acceleration for fully quantized BERT models has shown promise in enhancing the efficiency of NLP tasks. By quantizing all parameters of the BERT model, researchers have been able to optimize the model for low-latency inference, crucial for real-time applications. This approach not only reduces the memory footprint but also accelerates inference speed, especially on hardware supporting 8bit Integer calculations. Such advancements are pivotal for deploying large-scale NLP models in production environments where resources are limited (Zafrir et al., 2019).

These developments in quantized classification and model optimization highlight the potential of Transformer models in more resource-constrained settings. They pave the way for broader applications of NLP technologies, from edge devices to large data centers, maintaining high efficiency and accuracy.

The BERTimbau model, a Portuguese language adaptation of the renowned BERT (Bidirectional Encoder Representations from Transformers) model, was used in our study. Developed specifically for the Portuguese language, BERTimbau leverages the powerful architecture of the original BERT model, which is known for its deep bidirectional training and ability to understand the nuances of language context. The model's proficiency stems from its training on a vast corpus of Portuguese text, enabling it to capture the intricacies and idiosyncrasies of the language. This characteristic is particularly advantageous for our study, as the BERTimbau model's nuanced understanding of Portuguese text allows for more accurate analysis and processing of the B2W review corpus, which is predominantly in Portuguese. The model's bidirectional nature ensures that each word is effectively contextualized based on its surrounding text, leading to more robust and meaningful language representations. This aspect is crucial in accurately predicting vowel density and

categorizing sentences in our quantization tasks, providing a deeper understanding of the linguistic patterns inherent in the reviews.

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.

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.

The BertForVowelDensityRegression model leverages the neural architecture known as BERTimbau, specifically, the neuralmind/bert-base-portuguese-cased. This model is tailored for the task of vowel density regression. It commences with the integration of a BERT model, pre-trained on Portuguese text, serving as the foundational feature extractor. Subsequently, a singular linear regression layer is appended atop this architecture.

Incorporating the same BERT model, neuralmind/bert-base-portuguese-cased, the BertForQuantizedClassification model is devised for quantized classification tasks. For the purposes of regularization, a dropout layer with a dropout rate of 0.1 is introduced subsequent to the BERT model's output. The final classification is executed via a linear classifier, with the number of classes specified as a hyperparameter.

Again, employing the BERT model, neuralmind/bert-base-portuguese-cased, the BertForBalancedClassification model is constructed for classification tasks characterized by balanced classes. A linear classifier is affixed to the BERT-based feature extractor. Additionally, a softmax layer is employed to compute class probabilities, facilitating the accommodation of balanced class distributions.

2.3.1 Hyperparameters

The model_name hyperparameter designates the specific pre-trained BERT model utilized as the foundational backbone. Across all models, the neuralmind/bert-base-portuguese-cased variant is employed, pre-trained on Portuguese language text.

The num_classes hyperparameter determines the number of distinct classes for classification tasks. It is pivotal in configuring the output layer of the model for accurate classification.

The dropout hyperparameter, set at a rate of 0.1, is incorporated as a means of regularization. It aids in mitigating overfitting by randomly dropping a fraction of units during training.

The hyperparameters, including the choice of pre-trained model, number of classes, and dropout rate, play a pivotal role in shaping the models' performance and adaptability to specific tasks.

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 1: 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.

While the Regression_BERT model shows strong performance, care must be taken to ensure it is not overfitting the training data. This could be examined by evaluating the model on a separate validation set and checking for consistency in performance.

The choice of baselines might not be appropriate for this task. Exploring other baseline models or metrics could provide a better comparative analysis.

The presence of outliers or anomalous data points could disproportionately affect the baseline models, leading to negative R2 values. 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	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 2: Task 2 Classification Report

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 3: Task 3 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.

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: Comparing Task 2 and Task 3 metrics

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.

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, "Exploring Transformers Architecture for Vowel Density Regression and Quantization Tasks," effectively 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, enabled nuanced language understanding, 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 continuous nature of 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. This study, therefore, not only reinforces the existing knowledge base but also opens avenues for more sophisticated language processing tools in diverse linguistic environments.

5 Referências

Born, J., & Manica, M. (2023). Regression Transformer enables concurrent sequence regression and generation for molecular language modelling. Nature Machine Intelligence, 5, 432-444. Disponível em: https://www.nature.com/articles/s42256-023-00639-z

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