Automated Writing Proficiency Classification: A Framework for Qualifying Skill Levels in Text Analysis

Daniel Skahill [1] **Rohit Menon** [1] **Mana Vahid** [1]

Abstract

Writing proficiency classification is essential for assessing English Language Learners (ELLs), yet existing methods often fail to capture nuanced writing characteristics. We propose a cross-domain application of state-of-the-art text classification approaches utilizing RoBERTa embeddings with CNN, LSTM, and GRU layers to classify writing proficiency levels accurately. Leveraging datasets like EFCamDat and OneStopEnglish, our model addresses prompt bias through data augmentation and enhances interpretability using both Local Interpretable Model-agnostic Explanations (LIME) and Large Language Models (LLMs). Our classification approach achieves high F1 and accuracy scores, demonstrating the potential to provide detailed, actionable feedback for ELL writers. This work advances the limited literature on writing proficiency classification and underscores the effectiveness of deep learning models in educational contexts.

1. Introduction

Writing proficiency is not well defined, especially in the United States where current methods, such as using the Flesch Kincaid score, (see Appendix A for definition and more methods) associate children and English language learners' (ELLs) writing with grade levels (i.e. stating they write at a "4th-grade level" or "12th-grade level") (Arnost et al., 2021). While grade level may be correlated with children's writing levels, this classification system does not work for ELLs as it does not qualify the characteristics of writing. Receiving text classification alongside justification is an important part of the learning process, and it could prove to be a valuable feedback loop to encourage improvement amongst developing writers. More formal grading systems, like the Common European Framework of Reference for Languages (CEFR), classify writing proficiency on a beginner, intermediate, advanced system with each classification having two sublevels (A1, A2, B1, B2, C1, C2) (Council of Europe). This system is more descriptive but presents a massive rubric with no indication of weights for the various criteria. Learners receiving classification of their writing would need to parse the lengthy rubric and guess what criteria they were evaluated from.

Existing literature for writing proficiency classification is relatively sparse. We postulate that the difficulty of obtaining a large amount of quality, manually annotated data may explain scarcity in recent studies. Another reason may be that most writing-proficiency related tasks are concerned with academic grading rather than proficiency scoring. The nuanced difference between these tasks is that grading is concerned with content and quality with respect to a specific topic, whereas proficiency is concerned with quality in general (not with respect to a specific topic).

Our primary objective is to develop and implement a robust method for classifying texts into meaningful writing proficiency levels, leveraging the large EFCamDat dataset (Geertzen et al., 2013) and the smaller OneStopEnglish dataset (Vajjala and Lučić, 2021). This approach aims to ensure our model focuses on writing quality rather than content-specific features, thereby creating a more generalizable classification system.

The secondary objective is to design a system for providing personalized, constructive feedback that will analyze multiple linguistic aspects, including grammar, vocabulary usage, sentence structure, and coherence. The feedback mechanism should go beyond simple error detection, offering context-specific suggestions and explanations that align with the progression of skills outlined in the CEFR rubric. This approach can provide a consistent framework for both learners and educators to understand and track writing proficiency improvement.

[1]School of Information, University of California, Berkeley, United States. Correspondence to: Daniel Skahill <dannyskahill@berkeley.edu>, Rohit Menon <rvmenon@berkeley.edu>, Mana Vahid <manavahid@berkeley.edu>.

2. Background

The automated assessment of writing proficiency remains an under-explored area within natural language processing (NLP), particularly in comparison to other well-established areas of research. While substantial work has been conducted in the domain of automated essay scoring for academic purposes, there exists a notable gap in the comprehensive evaluation of writing proficiency in broader contexts.

However, there is still a strong foundation to build upon. Arnold et al. 2018 employed a multifaceted approach, incorporating metrics such as lexical diversity, syntactic complexity, and readability indices, in conjunction with a Gradient Boosted Random Forest model to predict proficiency levels. Similarly, Kerz et al. 2021 utilized an ensemble of 57 linguistic features, encompassing measures of syntactic complexity and lexical richness, coupled with Recurrent Neural Networks (RNNs) for text classification. Both studies highlight the effectiveness of feature engineering and traditional machine learning paradigms in this domain.

Across domains, in 2023, Semary et al. developed a novel RoBERTa-CNN-LSTM hybrid model that demonstrated exceptional accuracy in sentiment classification, reaching 93% on a Twitter airline sentiment dataset and 96% on IMDB reviews dataset. This model effectively combines RoBERTa's deep contextual analysis with the sequential processing strengths of CNNs and LSTMs, capturing detailed textual nuances necessary for accurate classification. Semary et al. built upon research from Ullah et al. 2022 who utilized RoBERTa with a Gated Recurrent Unit (GRU) layer to receive 94% accuracy on the IMDB dataset. The capabilities of these models suggest potential for use across domains, particularly in educational technology for creating tools that can accurately assess writing proficiency.

Wang et al. 2022 Utilized BERT for the task of automated essay scoring and found success against 12 different LSTM and deep learning approaches. We propose using a simplified BERT classification architecture to compare against RoBERTa and baseline models.

Understanding and interpreting the decisions made by complex models remains a challenge. To address this, Ribeiro et al. 2016 introduces Local Interpretable Model-agnostic Explanations (LIME). LIME essentially uses a simpler, interpretable model to explain predictions made by the original complex model. However, LIME's reliance on local surrogates can limit its interpretability and scalability. We intend to experiment with LIME for explainability, but we also hope to improve explainability capabilities by leveraging Large Language Models (LLMs), which provide context-aware, comprehensive explanations that better capture the differences in writing proficiency classifications.

3. Methodology

3.1. Data

As alluded to, there is not a lot of existing research that has produced large, manually annotated writing proficiency datasets. Therefore, the two datasets we utilize (EFCAMDAT and OneStopEnglish) each have some advantages and limitations.

3.1.1. EFCAMDAT

The EF-Cambridge Open Language Database (EFCamDat), developed at the University of Cambridge, contains ground truth labels for ELL text excerpts from Education First's (EF) online school, Englishtown. EFCamDat uses the Common European Framework of Reference for writing proficiency scores, classifying text as one of the following: A1, A2, B1, B2, C1, C2– with A1 being the most basic, and C2 being the most advanced proficiency. Proficiency levels are determined by graded assessments throughout the Englishtown course. Once a learner is designated a proficiency level, they complete the associated coursework and move along to the next proficiency level. It is noted that if a learner were to complete all proficiency levels, from A1 to C2, they would write a total of 128 scripts and essays (Geertzen et al., 2013). Because specific prompts are associated with each level, there is an increased risk of models not classifying based on proficiency, but rather on content. The risk is mitigated by the fact that levels A1-C1 each have 24 prompts with C2 having 8; meaning, it will be very difficult for the model to identify common themes across the 24 prompts other than the writing structure and quality. However, we employ additional augmentation strategies such as clipping the first and last 5 words (to combat email/letter prompts), random insertion, deletion, sentence shuffling, and synonym replacement to reduce any additional prompt bias in the classification.

EFCamDat has two datasets: one with raw, unedited text, and one with texts that have inline error correction codes. The second dataset contains all 6 classes, whereas the first dataset does not contain any "C2" level data. Because we do not want to bias our model with inline error codes, we focus on utilizing the first dataset but use the second for additional experimentation.

3.1.2. ONESTOPENGLISH

OneStopEnglish was created for Automatic Readability Assessments (ARA) and is composed of subtexts from articles or textbooks intended for different audiences (Vajjala and Lučić, 2021). It is broken down into beginner, intermediate, and advanced groupings (with no subgroupings) for ELLs. While the texts are not written by ELLs, reading and writing levels are highly synonymous in terms of our end goal–

Table 1. EFCamDat Class Counts

Classes	EFCamDat 5-Class	EFCamDat 6-Class
A1	100000	100000
A2	100000	100000
B1	100506	100000
B2	40238	61329
C1	10006	14698
C2	N/A	1940

being able to classify the proficiency level of a piece of text (regardless of who wrote it). As this dataset does not have specific prompts associated with different proficiency levels (but only has 3 classes), we utilize it as a sanity check for comparison against the much larger, EFCamDat.

Table 2. OneStopEnglish Class Counts

Classes	OneStopEnglish
Beginner	2651
Intermediate	2595
Advanced	2651

Table 3. Tokenized Dataset Statistics

Statistics	EFCamDat	OneStopEnglish
Average Length	115	74
Median Length	107	68
Max Length	500	416
Min Length	3	3
Total Texts	377967	7397

3.2. Classification

To classify writing proficiency levels, we explore different transformer and hybrid architectures. We construct a metric-based model to act as a threshold baseline to compete against.

To evaluate our models, we use F1 and accuracy. Accuracy provides an overall measure of the proportion of correctly classified instances, giving a straightforward indication of the model's effectiveness. F1-score balances precision and recall, making it particularly valuable in scenarios where class distributions are imbalanced. Together, these metrics offer a comprehensive assessment of each model's ability to accurately and reliably classify writing proficiency levels.

3.2.1. Baseline

We extract twelve linguistic features (see Appendix A for a list of features) and measures of complexity to predict writing proficiency levels with a logistic regression model. This approach serves as a straightforward, interpretable baseline against which to compare more complex, transformer-based models.

3.2.2. BERT

Bidirectional Encoder Representations from Transformers (BERT) employs a transformer architecture with multiple layers of bidirectional self-attention. This enables the model to capture context from both directions in a sentence. Our implementation involves simply inputting BERT tokenized text into the pre-trained model, capturing the pooled output, and feeding into this into a dense layer, and finally a classification layer. We follow a simplified approach to BERT text classification that Wang et al. 2022 utilized for Automated Essay Scoring.

3.2.3. RoBERTa

RoBERTa-CNN-LSTM (Semary et al., 2023) combines the strengths of RoBERTa (a robustly optimized BERT approach), Convolutional Neural Networks (CNN), and Long Short-Term Memory (LSTM) networks for text classification. We feed tokenized text into the pre-trained RoBERTa model, pass it through a one-dimensional convolutional layer with 256 filters for extracting local features, feed the sequence into an LSTM layer with 256 units to capture long-term dependencies, and classify the text.

RoBERTa-CNN-GRU (Ullah et al., 2022) integrates the robust text representations of RoBERTa with the pattern recognition capabilities of Convolutional Neural Networks (CNN) and the sequence modeling strengths of Gated Recurrent Units (GRU). Tokenized texts are processed by the pre-trained RoBERTa model, passed through a one-dimensional convolutional layer with 256 filters, fed into a GRU layer with 256 units (which can capture more advanced contextual nuances), and classified.

3.2.4. Multi-Step

Due to significant class imbalances present in the EFCamDat dataset, we propose two advanced model architectures in hopes of enhancing classification capabilities. The first architecture stratifies EFCamDat data into ABC groupings, using the RoBERTA+CNN+GRU model to predict these three classes. Subsequently, three sub-models are used to further discriminate between A1/A2, B1/B2, and C1/C2 levels. This is referred to as the ”3-step model.” The second architecture utilizes the RoBERTA+CNN+LSTM model to perform initial classification, identifies the most frequently confused classes, and replaces and reclassifies these classes. A sub-model discriminator is then used to refine the distinctions between the confused classes. This is called the ”2-step model.” These strategies aim to address class imbalance and improve the robustness of writing proficiency classification. (See Appendix B for architecture)

3.3. Explainability

To address our secondary focus of providing explainable feedback for writing classification, two approaches are utilized: Local Interpretable Model-agnostic Explanations (LIME) and Large Language Model (LLM) engagement.

3.3.1. LIME

LIME is a technique that explains the predictions of any machine learning model in an interpretable manner. It creates a local surrogate model wrapper (typically a simpler linear model) around a prediction instance. By perturbing the instance (e.g., by removing words) and observing the changes in predictions, LIME identifies the words that cause the most significant change in prediction when perturbed, deeming them the most influential.

3.3.2. LLM

The recent explosion in Large Language Models and their ability to provide succinct answers to specific prompts provides the unique opportunity to explore prompt engineering in conjunction with a classification system. We obtain the CEFR rubric for language proficiency from the Council of Europe, filter out only writing-specific categorizations, provide Cohere's Command-R LLM with the rubric and a specific prompt (see Appendix C), and ask why a specific text was given a certain classification.

4. Results and Discussion

4.1. Classification

A glance at Table 5, reveals a close race between trainable BERT and all RoBERTa models. The baseline model performs relatively well on the EFCAMDAT datasets but shows weaker performance on the OneStopEnglish. Both multi-step models show strong results but are slightly outperformed by the simpler configurations. The untrainable BERT model underperforms significantly compared to its trainable counterpart, indicating the importance of fine-tuning.

Table 4. Misclassificatioins

Model	**Num**
Both	427
BERT Only	421
RoBERTa Only	706

When trained on the EFCamDat datasets (the 5 class raw text and 6 class error encoded text), all models perform surprisingly well (with the exception of untrainable BERT). While there may be a small degree of prompt bias encoded in classifications for EFCamDat, experimentation on the much smaller, non-prompt-based OneStopEnglish dataset indicates that the majority of classification power comes from writing quality/proficiency identification. Even the baseline models indicate that there is a degree of formulaic proficiency across levels. The confusion matrices for each model (see Appendix D) reveal that the majority of classification difficulty comes from identifying advanced proficiency classes, which could either be due to significantly less data for classes B2 and beyond, or the fact that as writing proficiency levels increase, writing quality becomes more subjective to the reader.

To better understand the differences between RoBERTa (CNN+LSTM) and BERT model classifications, we analyze the full picture (Table 4) of what both models have trouble classifying.

4.1.1. Misclassified by Both Models

Both BERT and RoBERTa misclassify certain texts due to their inherent limitations in handling nuanced and informal language. For example, the text *"Hi! I'm having a birthday party on Sunday. I invite you can go on. In the party, We can meet many friends. In the morning, We can watch TV. In the afternoon, We have a lunch in my house. I can make much delicious food, I think you must like it. Hope you go on. Johnson"* (A1) is misclassified by both models. Clearly, the text does not show basic proficiency, however, both BERT and RoBERTa predict A2. The simple vocabulary and grammatical errors lead both models to incorrectly assess the proficiency level. This suggests that both models may overly rely on certain lexical cues and fail to account for when language does not necessarily follow typical patterns.

4.1.2. BERT Correct, RoBERTa Incorrect

In some instances, BERT's ability to capture straightforward language patterns through its bidirectional self-attention mechanism allows it to outperform RoBERTa. For example, BERT accurately classifies the text *"Hi, Ed. Let me thing... I have 10 dollars too. Maybe we can buy special pens, I remember saw some one in the mall, those pens have a special form, its for a better writing. Well, tell me if you are agre. Sosa."* with a label of A1, aligning with the actual label. RoBERTa, however, predicts a higher proficiency level of B1. This discrepancy indicates that RoBERTa's advanced training regimen might lead it to expect more complexity, causing it to misinterpret the simple and conversational tone as more sophisticated than it actually is. Further analysis illuminates that most of RoBERTa misclassifications are over-predictions.

4.1.3. RoBERTa Correct, BERT Incorrect

Conversely, RoBERTa's optimizations and extensive training regime often give it an edge in understanding complex

Table 5. Classification Metrics for Different Models

Model	EFCamDat (5 Classes) F1	EFCamDat (5 Classes) Accuracy	OneStopEnglish F1	OneStopEnglish Accuracy	EFCamDat (6 Classes) F1	EFCamDat (6 Classes) Accuracy
Baseline	0.66	0.75	0.53	0.53	0.62	0.87
BERT (untrainable)	0.52	0.69	0.38	0.47	0.40	0.60
BERT (trainable)	**0.95**	**0.97**	0.74	0.75	**0.92**	**0.96**
roBERTa + CNN + LSTM	0.94	**0.97**	**0.83**	**0.84**	0.90	**0.96**
roBERTa + GRU	**0.95**	**0.97**	0.80	0.80	0.83	0.94
roBERTa 2 step	0.92	0.96	N/A	N/A	0.80	0.93
roBERTa 3 step	0.90	0.95	N/A	N/A	0.82	0.91

language patterns and contexts, when BERT falls short. For instance, the text *"At first I met my husband in elementary school. He was in class two and I in class one. He was very tall and he had a terrible suitcase instead of a satchel. A few years later I was sixteen years old, I went to the bowling with my friends. We went there with motorbike. One of my friends brought his best friend. He brought me home at 23 o'clock with his motorbike. It was winter and it was very cold. But we stood at the door and talked all night. When it was five o'clock in the Morning, we realized that are invited to the same party today. He brought me home again and we talked all night. He was very shy, but now I took all my courage and kissed him. This was the beginning of our relationship."* is correctly classified by RoBERTa with a label of B1, while BERT predicts a lower level of A2. RoBERTa's ability to accurately parse the narrative's complexity and emotional nuance underscores its strength in handling more sophisticated texts.

Table 6 computes features for misclassified texts from RoBERTa and BERT and takes the difference (RoBERTA minus (-) BERT). While there is no definitive trend that is generalizable across all classes, it is interesting to note that for labels A1, B2, B3 RoBERTa has a harder time classifying higher word count, syllable count, complex word count, vocab size, noun chunks, and index scores than BERT, but in classes A2 and C1, the inverse seems to be true. The fact these metrics don't necessarily tell a clear story is precisely why we utilized transformer architectures to classify text–prior research focusing on feature extraction was not able encapsulate the whole context.

The comparative analysis of BERT and RoBERTa on writing proficiency classification reveals intriguing insights into their respective strengths and weaknesses. BERT's ability to capture contextual information through its bidirectional self-attention mechanism allows it to excel in straightforward, simple texts with clear contexts. However, BERT tends to misclassify texts with nuanced or informal language, such as personal anecdotes and conversational tones, where it struggles to accurately interpret colloquial expressions and incomplete sentences. On the other hand, RoBERTa's optimizations (including the removal of the next sentence prediction task and longer training with larger batches) seem to give it an edge in understanding complex language patterns. RoBERTa accurately classified texts with sophisticated vocabulary and complex sentence structures, such as a formal complaint letter and a promotional text, showcasing its ability to handle varied linguistic patterns. However, RoBERTa occasionally misinterprets simpler, informal texts, leading to overestimation of proficiency levels. Furthermore, we hypothesize that the BERT architecture could be more influenced by prompt-based training compared to RoBERTa, which utilized additional CNN and LSTM layers and showed higher performance on our non-prompt-based dataset, OneStopEnglishCorpus. BERT might rely more heavily on context and prompt influence, whereas RoBERTa's architecture allows for greater flexibility and handling of varied linguistic patterns. Despite these differences, both models display complementary strengths, suggesting that a combined or ensemble approach might leverage the strengths of both for improved classification accuracy.

4.2. Explainability

4.2.1. LIME

Figure 1. LIME output with RoBERTa Model

As described, LIME perturbs the text and continuously searches to see which words have the greatest affect on misclassifications. The output in Figure 1 shows that the

Table 6. Average RoBERTa - BERT Misclassification Statistics

Class	Word Count	Syllable Count	Character Count	Complex Word Count	Vocab Size	Lexical Diversity	Noun Chunks	Flesch Kincaid	Dale Chall	Gunning Fog	Coleman Liau	Automated Readability
A1	11.708	11.292	50.292	3.833	9.333	0.024	2.875	1.692	1.963	2.011	1.456	2.375
A2	2.042	-4.042	-10.708	7.042	1.458	0.003	1.708	-2.683	-1.403	-3.251	-2.150	-3.438
B1	25.667	39.500	140.125	18.083	15.625	-0.006	4.375	2.913	0.725	3.013	1.265	3.096
B2	9.500	20.167	79.708	10.750	3.250	-0.028	4.542	1.229	-0.191	0.884	0.650	1.675
C1	-5.875	-9.500	-41.167	-2.583	-1.208	0.008	-1.833	-1.542	-0.360	-1.586	-0.563	-2.275

RoBERTa (CNN+LSTM) model correctly predicts the label as 3, with the LIME visualization highlighting "project" indicating the text relates to a complex topic such as project management. Words like "Order" and "achieved" suggest a structured and complex narrative, characteristic of higher-grade writing. Consequently, the model accurately assigns a writing proficiency level of 3, aligning with the true label (see Appendix E for more LIME examples).

While LIME is effective in providing minimal local explanations for model predictions, it has several drawbacks. It can be computationally expensive as it requires generating many perturbations of the input data to approximate the model's behavior. Additionally, the explanations provided by LIME are inherently local and might not capture the global behavior of the model, leading to potentially inconsistent interpretations across different instances. LIME's interpretability also relies on how well the simpler model can approximate the behavior of the complex model, namely, LIME does not seem to be completely contextually aware.

4.2.2. LLM

The limitations of LIME inspired the to transition to using large language models (LLMs) for explainability. LLMs can provide more comprehensive and coherent explanations by leveraging their extensive training on diverse datasets, offering both local and global insights into model behavior. This shift enhances the depth and consistency of model interpretability, making it more robust and reliable.

```
Student Text:
        Instructions for Frisbee bowling Please. We need to mark an area of 8 meters by 3 meters.  Lets use ten p
lastic bottles as 'bowling pins' with a little of water to make them heavier. Line up the bottles in line of 4, the
n 3, then 2 and then 1, like ten-pine bowling. Each player will get a frisbee and do two shots on each turn. To sco
re the player needs to knock down each pin. The player with the most points is the winner.

Proficiency Category: B1

'The text is classified as B1 because it demonstrates the ability to produce a straightforward, connected text on a
familiar topic, in this case, a set of instructions for a game. The description is clear and easy to follow, with a
logical structure that helps the reader understand the rules and objective of the game. \n\nHowever, there are seve
ral reasons why it is not classified at a higher level, such as B2 or C1: \n\n- Limited complexity: The text does n
ot demonstrate the ability to synthesize or evaluate information from multiple sources, which is a characteristic o
f higher levels. It focuses on providing clear instructions rather than analyzing or discussing the game in a more
complex manner. \n- Lack of argumentation: Higher levels, especially B2 and above, often involve developing argumen
ts and expressing opinions. This text does not include any personal views or attempts to persuade the reader, which
is typical of B1-level writing. \n- Restricted vocabulary and grammar: While the text is well-written and effective
for its purpose, it does not showcase a wide range of vocabulary or complex grammatical structures. Higher levels w
ould typically involve a more varied and sophisticated use of language, including idiomatic expressions and advance
d syntax. \n- Limited scope:'
```

Figure 2. LLM explaining text classification

In figure 2, the LLM gives a clear explanation for why the text was categorized as B1, but also why it was not higher or lower than the given class. The large language model is particularly useful in pointing to specific details from the provided rubric. There is enough feedback that a learner could easily implement into their edits or future work (see Appendix F for more examples of LLM explainability).

While both LIME and LLM techniques have tradeoffs, the LLM engagement for text classification explainability seems to be the most robust and informative.

5. Conclusion

Our research demonstrated the efficacy of both BERT and RoBERTa-based classifiers in assessing writing proficiency, highlighting the potential of transformer models augmented with CNN, LSTM, and GRU layers. Despite the success observed, our experimentation underscores the necessity for a comprehensive non-prompt-based language proficiency dataset encompassing all six CEFR levels to refine these models further. Given the dataset limitations, it is challenging to definitively identify the superior model, however, all models offer a solid foundation to build upon. The higher performance of RoBERTa models on the non-prompt-based dataset leads us to believe that this architecture may be more suitable for downstream tasks related to writing proficiency classification. Our framework of 1) classifying text and 2) explaining the classification via prompt engineering from LLMs delivered promising results, and should be implemented into ELL programs for more personalized and higher quality feedback. While LIME attempts to provide explanations for the classifications made by our models, it falls short in several critical areas. Specifically, LIME struggles to clearly identify and articulate the underlying criteria that drive the classification decisions. This lack of clarity and precision in explanation makes it difficult for educators and learners to understand the specific aspects of writing that are being evaluated, thereby limiting its practical usefulness. Given these shortcomings, we believe that LIME is not suitable for this application and does not warrant further exploration in the context of writing proficiency classification. Instead, we recommend investigating alternative interpretability methods (like engaging LLMs) that can offer more transparent and actionable insights into the model's decision-making process.

References

Taylor B Arnold, Nicolas Ballier, Thomas Gaillat, and Paul Lissón. Predicting cefrl levels in learner english on the basis of metrics and full texts. *ArXiv*, abs/1806.11099, 2018. URL https://statsmaths.github.io/

pdf/2018-learner-corpora.pdf.

William Arnost, Ellen Lull, Joseph Schueder, and Joseph Engler. Reading level identification using natural language processing techniques. *SMU Data Science Review*, 5(3), 2021. URL https://scholar.smu.edu/datasciencereview/vol5/iss3/7.

CoE Council of Europe. The cefr levels - common european framework of reference for languages (cefr) - www.coe.int. https://www.coe.int/en/web/common-european-framework-reference-languages/level-descriptions. Common European Framework of Reference for Languages (CEFR).

Jeroen Geertzen, Dora Alexopoulou, and Anna Korhonen. *Automatic linguistic annotation of large scale L2 databases: The EF-Cambridge Open Language Database (EFCAMDAT)*. 01 2013.

Elma Kerz, Daniel Wiechmann, Yu Qiao, Emma Tseng, and Marcus Ströbel. Automated classification of written proficiency levels on the cefr-scale through complexity contours and rnns. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 199–209. Association for Computational Linguistics, 2021. URL https://aclanthology.org/2021.bea-1.21.pdf.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322. URL https://doi.org/10.1145/2939672.2939778.

Noura A Semary, Wesam Ahmed, Khaled Amin, Paweł Pławiak, and Mohamed Hammad. Improving sentiment classification using a roberta-based hybrid model. *Frontiers in Human Neuroscience*, 17, 2023. URL https://doi.org/10.3389/fnhum.2023.1292010.

Farhat Ullah, Xin Chen, Syed Bilal Hussain Shah, Saoucene Mahfoudh, Muhammad Abul Hassan, and Nagham Saeed. A novel approach for emotion detection and sentiment analysis for low resource urdu language based on cnn-lstm. *Electronics*, 11(24), 2022.

Sowmya Vajjala and Ivana Lučić. Onestopenglish corpus for text simplification. In *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications*, pages 124–134. Association for Computational Linguistics, 2021. URL https://aclanthology.org/W18-0535.pdf.

Yongjie Wang, Chuan Wang, Ruobing Li, and Hui-Ching Lin. On the use of bert for automated essay scoring: Joint learning of multi-scale essay representation. *ArXiv*, abs/2205.03835, 2022. URL https://api.semanticscholar.org/CorpusID:248572476.

A. Text Features (Baseline Model)

- Word Count- Number of words in a text
- Syllable Count- Number of syllables in a text
- Character Count- Number of characters in a text
- Complex Word Count- Number of complex words (a word which has more than 2 syllables)
- Vocab Size- Number of unique words used
- Lexical Diversity- Ratio of unique words to text length
- Noun Chunks- Continuous noun phrases
- Flesch Kincaid Score- A score determining how difficult a piece of text is to read.

$$\text{FK} = 206.835 - 1.015\left(\frac{\text{Total Words}}{\text{Total Sentences}}\right) - 84.6\left(\frac{\text{Total Syllables}}{\text{Total Words}}\right) \tag{1}$$

Where the score can be evaluated on this table:

Score	School level (US)	Notes
100.00–90.00	5th grade	Very easy to read. Easily understood by an average 11-year-old student.
90.0–80.0	6th grade	Easy to read. Conversational English for consumers.
80.0–70.0	7th grade	Fairly easy to read.
70.0–60.0	8th & 9th grade	Plain English. Easily understood by 13- to 15-year-old students.
60.0–50.0	10th to 12th grade	Fairly difficult to read.
50.0–30.0	College	Difficult to read.
30.0–10.0	College graduate	Very difficult to read. Best understood by university graduates.
10.0–0.0	Professional	Extremely difficult to read. Best understood by university graduates.

- Dale Chall Score- A readability metric that addresses the number of ”difficult words”

$$\text{DC} = 0.1579\left(\frac{\text{difficult words}}{\text{words}} \times 100\right) + 0.0496\left(\frac{\text{words}}{\text{sentences}}\right) \tag{2}$$

Where the score can be evaluated on this table:

Score	Notes
4.9 or lower	easily understood by an average 4th-grade student or lower
5.0–5.9	easily understood by an average 5th- or 6th-grade student
6.0–6.9	easily understood by an average 7th- or 8th-grade student
7.0–7.9	easily understood by an average 9th- or 10th-grade student
8.0–8.9	easily understood by an average 11th- or 12th-grade student
9.0–9.9	easily understood by an average college student

- Gunning Fog Index- A readability test which describes what grade level someone should be able to read a specific text

$$\text{GF} = 0.4\left[\left(\frac{\text{words}}{\text{sentences}}\right) + 100\left(\frac{\text{complex words}}{\text{words}}\right)\right] \tag{3}$$

Where the score can be evaluated on this table:

Fog Index	Reading level by grade
17	College graduate
16	College senior
15	College junior
14	College sophomore
13	College freshman
12	High school senior
11	High school junior
10	High school sophomore
9	High school freshman
8	Eighth grade
7	Seventh grade
6	Sixth grade

- Coleman Liau Index- A readability index that estimates grade level through average metrics

$$\text{CL} = 0.0588 * L - 0.296 * S - 15.8 \tag{4}$$

Where L is the average number of letters per 100 words and S is the average number of sentences per 100 words.

- Automated Readability Index- Readability score that represents words, characters, and sentences as grade level

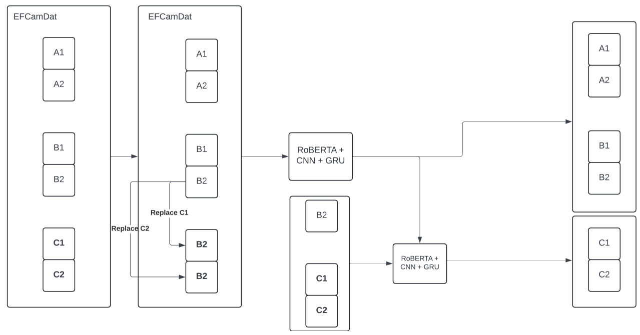
$$\text{ARI} = 4.71\left(\frac{\text{characters}}{\text{words}}\right) + 0.5\left(\frac{\text{words}}{\text{sentences}}\right) - 21.43 \tag{5}$$

Where the score can be evaluated on this table:

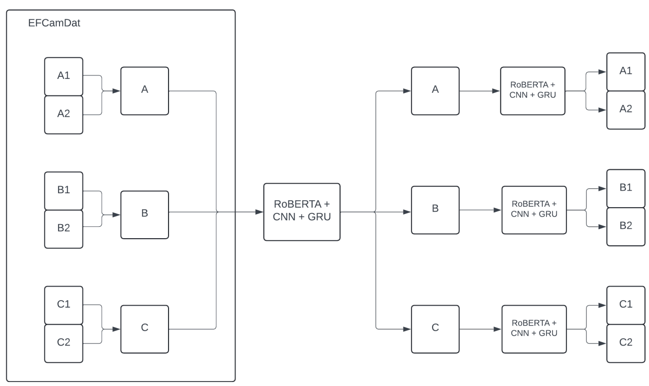
Score	Age	Grade Level
1	5-6	Kindergarten
2	6-7	First Grade
3	7-8	Second Grade
4	8-9	Third Grade
5	9-10	Fourth Grade
6	10-11	Fifth Grade
7	11-12	Sixth Grade
8	12-13	Seventh Grade
9	13-14	Eighth Grade
10	14-15	Ninth Grade
11	15-16	Tenth Grade
12	16-17	Eleventh Grade
13	17-18	Twelfth Grade
14	18-22	College student

B. Architectures

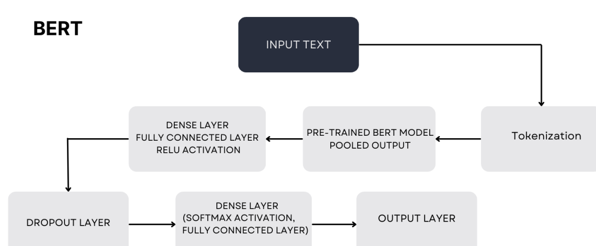
B.1. 2 Step Architecture

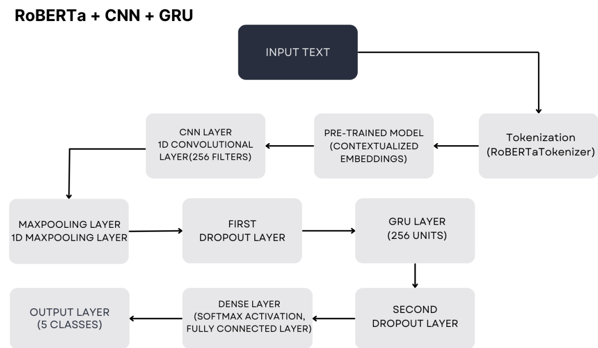


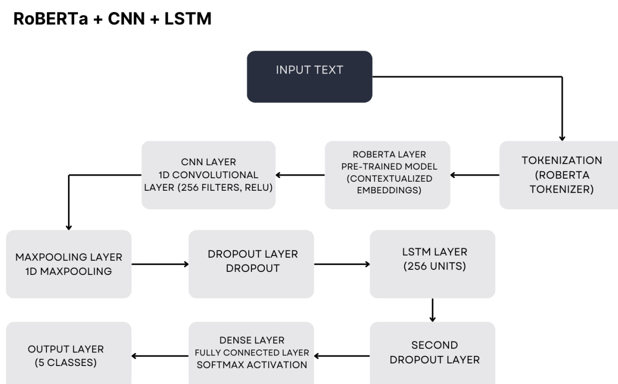
B.2. 3 Step Architecture

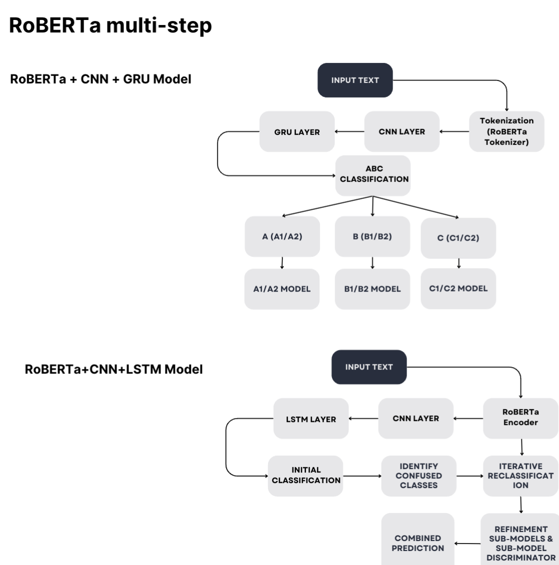


B.3. All Architectures Diagrams









C. LLM Prompt and CEFR Rubric

C.1. LLM Prompt

””””You are an agent designed to provide feedback on writing samples according to the provided rubric. You will first receive the rubric, then you will receive the writing sample and what classification it falls under. You will use this information to explain why the writing sample falls under that classification based on the rubric. The following rubric is given in the format of Classification level: Description, Classification level: Description, ...

Here is the rubric:

rubric ”””

C.2. Rubric

’A1’: ’Can give information about matters of personal relevance (e.g. likes and dislikes, family, pets) using simple words/signs and basic expressions.; Can produce simple isolated phrases and sentences.; Can produce simple phrases and sentences about themselves and imaginary people, where they live and what they do.; Can describe in very simple language what a room looks like.; Can use simple words/signs and phrases to describe certain everyday objects (e.g. the colour of a car, whether it is big or small).; No descriptors available; Can ask for or pass on personal details.; Can compose messages and online postings as a series of very short sentences about hobbies and likes/dislikes, using simple words and formulaic expressions, with reference to a dictionary.; Can compose a short, simple postcard.; Can compose a short, very simple message (e.g. a text message) to friends to give them a piece of information or to ask them a question.; Can fill in numbers and dates, own name, nationality, address, age, date of birth or arrival in the country, etc., e.g. on a hotel registration form.; Can leave a simple message giving information regarding for instance where they have gone, or what time they will be back (e.g. “Shopping: back at 5 p.m.”).’,

’A2’: ’Can produce a series of simple phrases and sentences linked with simple connectors like “and”, “but” and “because”.; Can produce a series of simple phrases and sentences about their family, living conditions, educational background, or present or most recent job.; Can create short, simple imaginary biographies and simple poems about people.; Can create diary entries that describe activities (e.g. daily routine, outings, sports, hobbies), people and places, using basic, concrete vocabulary and simple phrases and sentences with simple connectives like “and”, “but” and “because”.; Can compose an introduction to a story or continue a story, provided they can consult a dictionary and references (e.g. tables of verb tenses in a course book).; Can produce simple texts on familiar subjects of interest, linking sentences with connectors like “and”, “because” or “then”.; Can give their impressions and opinions about topics of personal interest (e.g. lifestyles and culture, stories), using basic everyday vocabulary and expressions.; Can compose short, simple formulaic notes relating to matters in areas of immediate need.; Can convey personal information of a routine nature, for example in a short e-mail or letter introducing themselves.; Can compose very simple personal letters expressing thanks and apology.; Can compose short, simple notes, e-mails and text messages (e.g. to send or reply to an invitation, to confirm or change an arrangement).; Can compose a short text in a greetings card (e.g. for someone’s birthday or to wish them a Happy New Year).; Can formulate short, simple notes and messages relating to matters in areas of immediate need.; Can fill in personal and other details on most everyday forms (e.g. to open a bank account, or to send a letter by recorded delivery).’,

’B1’: ’Can produce straightforward connected texts on a range of familiar subjects within their field of interest, by linking a series of shorter discrete elements into a linear sequence.; Can give straightforward, detailed descriptions on a range of familiar subjects within their field of interest.; Can give accounts of experiences, describing feelings and reactions in simple, connected text.; Can give a description of an event, a recent trip – real or imagined.; Can narrate a story.; B1; B1; B1; Can produce very brief reports in a standard conventionalised format, which pass on routine factual information and state reasons for actions.; Can present a topic in a short report or poster, using photographs and short blocks of text.; Can compose personal letters and notes asking for or conveying simple information of immediate relevance, getting across the point they feel to be important.; Can compose personal letters describing experiences, feelings and events in some detail.; Can compose basic e-mails/letters of a factual nature (e.g. to request information or to ask for and give confirmation).; Can compose a basic letter of application with limited supporting details.; B1; Can formulate notes conveying simple information of immediate relevance to friends, service people, teachers and others who feature in their everyday life, getting across comprehensibly the points they feel are important.; Can take messages over the phone containing several points, provided the caller dictates these clearly and sympathetically.; B1’,

’B2’: ’Can produce clear, detailed texts on a variety of subjects related to their field of interest, synthesising and evaluating information and arguments from a number of sources.; Can give clear, detailed descriptions on a variety of subjects related to their field of interest.; Can give a review of a film, book or play.; B2; B2; Can produce an essay or report which develops an argument, giving reasons in support of or against a particular point of view and explaining the advantages and disadvantages of various options.; Can synthesise information and arguments from a number of sources.; B2; B2; B2; Can express news and views effectively in writing, and relate to those of others.; B2; Can compose letters conveying degrees of emotion and highlighting the personal significance of events and experiences and commenting on the correspondent's news and views.; Can use formality and conventions appropriate to the context when writing personal and professional letters and e-mails.; Can compose formal e-mails/letters of invitation, thanks or apology using appropriate registers and conventions.; Can compose non-routine professional letters, using appropriate structure and conventions, provided these are restricted to matters of fact.; Can obtain, by letter or e-mail, information required for a particular purpose, collate it and forward it by e-mail to other people.; B2; B2; B2; B2; Can take or leave complex personal or professional messages, provided they can ask for clarification or elaboration if necessary.; B2; B2',

'C1': 'Can produce clear, well-structured texts of complex subjects, underlining the relevant salient issues, expanding and supporting points of view at some length with subsidiary points, reasons and relevant examples, and rounding off with an appropriate conclusion.; Can employ the structure and conventions of a variety of genres, varying the tone, style and register according to addressee, text type and theme.; Can produce clear, detailed, well-structured and developed descriptions and imaginative texts in an assured, personal, natural style appropriate to the reader in mind.; Can incorporate idiom and humour, though use of the latter is not always appropriate.; Can give a detailed critical review of cultural events (e.g. plays, films, concerts) or literary works.; C1; Can produce clear, well-structured expositions of complex subjects, underlining the relevant salient issues.; Can expand and support points of view at some length with subsidiary points, reasons and relevant examples.; Can produce a suitable introduction and conclusion to a longer report, article or dissertation on a complex academic or professional topic provided the topic is within their field of interest and there are opportunities for redrafting and revision.; C1; C1; C1; Can express themselves with clarity and precision, relating to the addressee flexibly and effectively.; Can express themselves with clarity and precision in personal correspondence, using language flexibly and effectively, including emotional, allusive and joking usage.; Can, with good expression and accuracy, compose formal correspondence such as letters of clarification, application, recommendation, reference, complaint, sympathy and condolence.; C1; C1; C1; C1; No descriptors available; see B2',

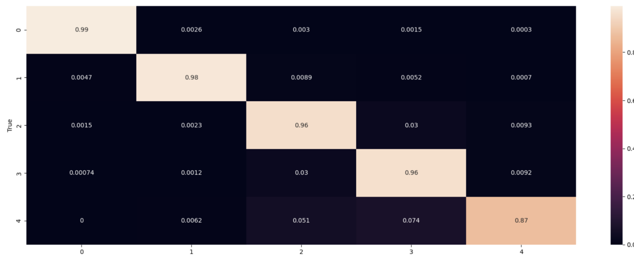
'C2': 'Can produce clear, smoothly flowing, complex texts in an appropriate and effective style and a logical structure which helps the reader identify significant points.; Can relate clear, smoothly flowing and engaging stories and descriptions of experience in a style appropriate to the genre adopted.; Can exploit idiom and humour appropriately to enhance the impact of the text.; Can produce clear, smoothly flowing, complex reports, articles or essays which present a case, or give critical appreciation of proposals or literary works.; Can provide an appropriate and effective logical structure which helps the reader identify significant points.; Can set out multiple perspectives on complex academic or professional topics, clearly distinguishing their own ideas and opinions from those in the sources.; Can express themselves in an appropriate tone and style in virtually any type of formal and informal interaction.; Can compose virtually any type of correspondence necessary in the course of their professional life in an appropriate tone and style.; No descriptors available; see B2'

D. Model Confusion Matrices

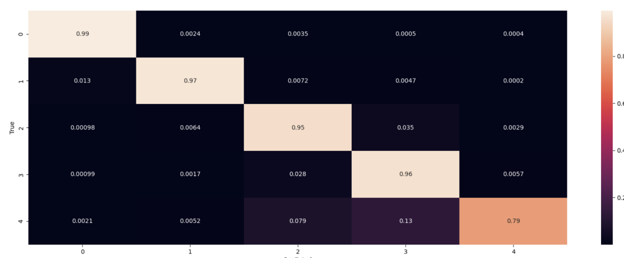
Confusion Matrices are for EFCamDat 5 Class

D.1. BERT Untrainable

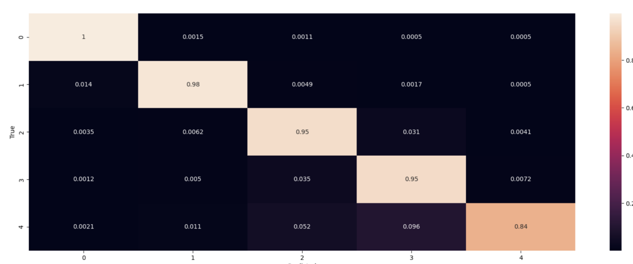
D.2. BERT Trainable



D.3. RoBERTa+CNN+LSTM

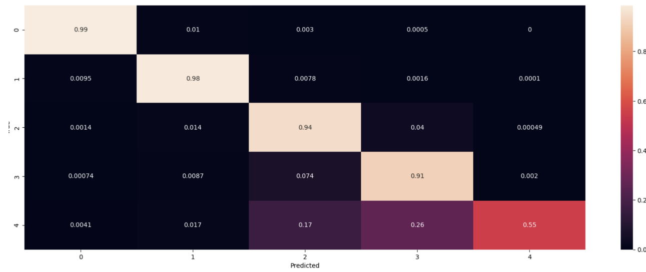


D.4. RoBERTa+CNN+GRU

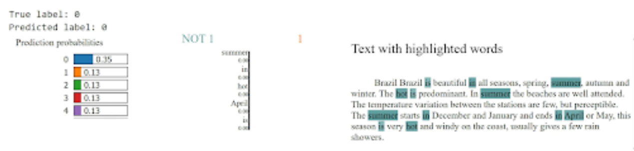


D.5. RoBERTa 2 Step

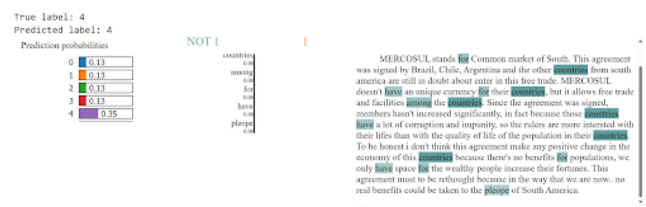
D.6. RoBERTa 3 Step



E. LIME Explainability Examples



In this example, the RoBERTa model correctly predicts the label as 0, with the LIME visualization highlighting words like ”summer,” ”hot,” ”is,” “in,” and ”April” as significant to the model’s prediction. The repeated use of ”summer” indicates its importance in identifying seasonal descriptions typical of a particular grade level. The words ”hot” and ”April” further support the context about weather and seasons, providing clues about time-related content. These words inform the model that the text is a simple description of seasons, influencing its prediction for the grade level.



The model identifies the term ”Countries” as indicative of geopolitical or economic complexity, suggesting a higher grade-level discussion. The repeated use of functional words like ”Have” and ”For” points to sentence complexity and overall readability, reflecting a higher writing proficiency. Additionally, the preposition ”Among” implies relationships and comparisons, further contributing to the text’s complexity and readability.

F. LLM Explainability Examples

```
Student Text:
         Hi, can you buy me:  A bottle of red wine, a loaf of bread, a bag of rice, a packet of chips, one box of
ice cream, some tomatoes and some pork.  Thanks.  When you go back, I'll invite you for dinner in my house!

Proficiency Category: A1

'This piece of text is classified as A1 because it demonstrates the ability to use simple words and phrases to conv
ey a message about matters of personal relevance (in this case, requesting someone to buy some groceries and inviti
ng them for dinner). Here are the specific reasons why it falls under the A1 classification and not higher:\n\n- **
Limited vocabulary and sentence structure:** The text uses simple words and phrases like "a bottle of," "a loaf o
f," and "a bag of," indicating a basic vocabulary level. The sentences are short and mainly consist of subject-verb
-object structures, which is typical of A1-level language production.\n- **Lack of complex connectors:** The text d
oes not demonstrate the use of complex connectors like "because," "although," or "despite," which are characteristi
c of higher-level language production. The only connector used is "and," which is a basic connector.\n- **No abstra
ct or complex ideas:** The content of the text is concrete and practical, focusing on specific items to be purchase
d. There are no abstract concepts, complex arguments, or detailed descriptions, which would be expected in higher-l
evel writing (e.g., B1, B2, or C1).\n- **Limited'
```

```
Student Text:
        Instructions for Frisbee bowling Please. We need to mark an area of 8 meters by 3 meters.  Lets use ten p
lastic bottles as 'bowling pins' with a little of water to make them heavier. Line up the bottles in line of 4, the
n 3, then 2 and then 1, like ten-pine bowling. Each player will get a frisbee and do two shots on each turn. To sco
re the player needs to knock down each pin. The player with the most points is the winner.

Proficiency Category: B1

'The text is classified as B1 because it demonstrates the ability to produce a straightforward, connected text on a
familiar topic, in this case, a set of instructions for a game. The description is clear and easy to follow, with a
logical structure that helps the reader understand the rules and objective of the game. \n\nHowever, there are seve
ral reasons why it is not classified at a higher level, such as B2 or C1: \n\n- Limited complexity: The text does n
ot demonstrate the ability to synthesize or evaluate information from multiple sources, which is a characteristic o
f higher levels. It focuses on providing clear instructions rather than analyzing or discussing the game in a more
complex manner. \n- Lack of argumentation: Higher levels, especially B2 and above, often involve developing argumen
ts and expressing opinions. This text does not include any personal views or attempts to persuade the reader, which
is typical of B1-level writing. \n- Restricted vocabulary and grammar: While the text is well-written and effective
for its purpose, it does not showcase a wide range of vocabulary or complex grammatical structures. Higher levels w
ould typically involve a more varied and sophisticated use of language, including idiomatic expressions and advance
d syntax. \n- Limited scope:'
```

```
Student Text:
        I'd like to start by describing my so called physiolocical needs, which I've achieved generally. Secondly
on the ladder appears the safety aspect. Well, I am actually quite comfortable. Everything's fine with my family an
d every family member feels fine. So there is nothing to criticize. I'd like to get on with the "love thing ". I am
happily married for about 30 years. That's all I'd like to tell you. The intension is on happily here. There is not
hing else to say. :) Esteem: I am a hard working guy and trying to give the best in job every day. I am expecting j
ust a little  acknowledgment from time to time. Perhaps a wee bit more as usally . :) Right, last but not least I a
m not on the top of the ladder. There are a few things which will have to be improved. I really should spend more t
ime with my wife. Work should not play a main role in our life anymore. I'd like to perfect my language skills with
in the next 3 years as well as  perhaps to start learning a second foreign language. I'd like to travel to New York
next year as a short term goal.

Proficiency Category: C1

'The text is classified as C1 because it demonstrates the ability to produce a clear and well-structured text with
complex sentences and an advanced vocabulary. The writer effectively conveys their thoughts and feelings on a perso
nal topic, using appropriate tone and style for the audience. \n\nHowever, there are a few reasons why it is not cl
assified higher, at C2:\n- While the text is clear and well-structured, it does not demonstrate the same level of c
omplexity and sophistication in sentence structure and vocabulary that would be expected at the C2 level. It does n
ot, for example, include much idiomatic or humorous expression, which is something that a C2-level text might incor
porate. \n- The text also does not demonstrate the same level of flexibility in tone and style that a C2-level text
might. While the writer does adapt their tone and style appropriately for the audience and topic, there is room for
more variation and nuance in their expression. \n- Finally, while the text is well-organized and logical, it does n
ot demonstrate the same level of sophistication in structure and argumentation that a C2-level text might. It does
not, for example, present multiple perspectives on a topic or include extensive justification of the writer'
```