Automatic Essay Multi-dimensional Scoring with Fine-tuning and Multiple Regression

Kun Sun*[1] and Rong Wang †[2]

[1]Department of Linguistics, University of Tübingen, Germany
[2]Institute of Natural Language Processing, Stuttgart University, Stuttgart, Germany

Abstract

Automated essay scoring (AES) involves predicting a score that reflects the writing quality of an essay. Most existing AES systems produce only a single overall score. However, users and L2 learners expect scores across different dimensions (e.g., vocabulary, grammar, coherence) for English essays in real-world applications. To address this need, we have developed two models that automatically score English essays across multiple dimensions by employing fine-tuning and other strategies on two large datasets. The results demonstrate that our systems achieve impressive performance in evaluation using three criteria: precision, F1 score, and Quadratic Weighted Kappa. Furthermore, our system outperforms existing methods in overall scoring.

Keywords: multiple scoring, classifier, mutiple regression, fine-tuning, LLMs

*email: kun.sun@uni-tuebingen.de
†email: rongw.de@gmail.com

1 Introduction

Automated Essay Scoring (AES) could automatically evaluate the proficiency of written essays. When AES works with high effectiveness, this technology not only saves educators substantial time that would otherwise be spent on manual essay grading but also provides students with immediate and free feedback. Moreover, AES systems offer more consistent and impartial assessments compared to human evaluators. More importantly, AES could be commercially applied in computer-aided language learning market.

Over the past half-century, a diverse array of methodologies has been proposed to tackle the challenges of AES. These methodologies range from learning from rule-based features to machine learning methods [10]; [16]. The recent development is to leverage neural approaches, including pre-trained language models [6], [24]. After the revolution of transformer-based language models and large language models (LLMs), the AES work based on LLMs has achieved STOA performance. This process is quite similar to other tasks in NLP. The primary objective of most AES research is to predict an overall holistic score that aligns closely with human judgment. Additionally, other studies have focused on providing detailed feedback by estimating quality scores across multiple traits of an essay.

We summarize the past work on AES. First, there are several approaches to AES: regression model [20] where the goal is to predict the score of an essay; [4]; [6], classifier model, where the goal is to classify an essay as belonging to one of a small number of classes (e.g., low, medium, or high, as in the TOEFL11 corpus) [14]. [21], [9] and ranking model [25], [2], [5], where the goal is to rank two or more essays based on their quality. Second, the vast majority of existing AES systems were developed for holistic scoring. Dimension-specific scoring did not start until 2004. So far, several dimensions of quality have been examined, such as, organization [15], argument persuasiveness [8], coherence [19]. However, these systems have not been developed based on LLMs. Third, the datasets were mostly relying on two: ASAP (Automated Student Assessment Prize) and TOEFL11, and both merely include the information holistic score.

Recent advancements in AES have been driven by the use of transformer-based language models, which achieve state-of-the-art performance through combined regression and ranking optimization techniques [24], [23], [7]. Furthermore, with proper prompting strategies, general-purpose language models like ChatGPT and LLaMA can also facilitate AES for small-scale essays [12], [13], [11]. However, when it comes to processing a large volume of essays, specialized APIs and appropriate prompts are required for efficient handling. Our focus is on these specific AES that enables spontaneous and comprehensive processing of numerous essays (e.g., [22], [23]).

Overall, AES stands as a pivotal innovation in educational technology, promising to enhance the efficiency and fairness of essay assessments. As NLP continues to evolve, AES systems are expected to become even more sophisticated, further

bridging the gap between automated and human scoring.

Despite the impressive results achieved by models designed for AES, several challenges and limitations persist. The following specifies these limitations.

Model training on `BERT`: A number of models for AES have been trained using BERT, a general-purpose language model. However, BERT might not be the optimal choice for directly enhancing AES effectiveness. Scoring essays could be fundamentally takne as a text classification task to some degree, and improvements can be made from an engineering perspective by fine-tuning existing models specialized for text classification.

Limitations of training datasets: The training datasets for AES often rely heavily on the ASAP essays, which are written by students in grades 7 and 10 in USA (i.e., junior secondary school students). These essays may not provide the necessary complexity and depth for robust AES training. Currently, there are superior datasets available that include essays from students in grades 10 to 12. These datasets offer a richer variety of language use and structure, making them more suitable for training advanced AES systems.

Holistic scoring limitations: Traditional AES systems typically provide a single holistic score for each essay. While this approach simplifies the scoring process, it fails to meet the nuanced expectations of real-world users. Ideally, an advanced AES system should deliver multiple scores across different dimensions such as vocabulary, grammar, coherence, and overall quality. This multi-dimensional evaluation would offer a more comprehensive and objective assessment of essays, providing more valuable feedback to second-language learners and researchers.

In summary, while current AES models have achieved significant progress, there is room for improvement in several areas. By addressing these challenges, we can develop more effective and versatile AES systems that provide richer, more detailed feedback and are better suited to the diverse needs of educational contexts.

To overcome these limitations, we propose new strategies and utilize an enhanced dataset to train a more robust AES system capable of grading essays across multiple dimensions. This new system is termed as "automatic essay multi-dimensional scoring" (**AEMS**). This is the primary objective of this paper. The following sections detail the methods employed to train AEMS and present our experimental results.

2 Methods

2.1 Technique route

As discussed in the introduction section, we adopted a novel approach to design a new multi-dimensional scoring system. We first selected several effective classifiers and fine-tuned them. Next, we aimed to implement multi-dimensional scoring, a process akin to multiple regression, for which we employed a multiple regression model during fine-tuning. Additionally, some essays include extra information,

such as requirements, topics, and types. We applied contrastive learning to enable the models to better understand and incorporate this information, enhancing the accuracy of the scoring. The following details these techniques.

First, for multi-class classification using BERT or RoBERT or other BERT-based classifiers, we obtain the representation of the `[CLS]` token from the BERT model. This representation is then passed through a dense layer with a `softmax` activation to predict class probabilities.

Let BERT(x) be the output representation of the [CLS] token for input x; W be the weight matrix of the dense layer; b be the bias vector; $\hat{y}$ be the predicted probability vector for K classes. The predicted logits z are given by:

$$z = W \cdot \text{BERT}(x) + b$$

We then apply the `softmax` function to obtain the class probabilities:

$$\hat{y}_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \quad \text{for } i = 1, 2, \ldots, K$$

Second, to incorporate multiple regression, we add a separate regression head to the `[CLS]` token representation. Let β be the regression weight vector; γ be the regression bias; y be the continuous target variable. The regression model is expressed as:

$$y = \beta \cdot \text{BERT}(x) + \gamma$$

The combined BERT-based model has two heads: one for classification and one for regression. The overall loss function combines cross-entropy loss for classification and mean squared error (MSE) loss for regression.

Let $\mathcal{L}_{\text{CE}}$ be the cross-entropy loss; $\mathcal{L}_{\text{MSE}}$ be the mean squared error loss; λ be the weight balancing the two losses. The combined loss function is:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda \mathcal{L}_{\text{MSE}}$$

By combining these losses, the BERT-based model can be trained to perform multi-class classification and regression simultaneously. The contrastive learning allows to better learn the information on essay prompts. The formal description on this is seen in **Appendix A**.

Our primary approach involves fine-tuning existing text classification models and applying multiple regression methods to improve the effectiveness for multi-dimensional scoring, and enhancing their capabilities through contrastive learning to better understand essay prompts. To develop a more practical AES for second language (L2) learners' essays, we employed a new dataset, the English Language Learner Insight, Proficiency, and Skills Evaluation Corpus (ELLIPSE) [3], which contains 9,000 essays written by students in grades 8 to 12 in USA. Additionally, to ensure applicability in real-world scenarios, we trained our models on the official IELTS (International English Language Testing System) exam dataset The following sections specify our strategies and training datasets in detail.

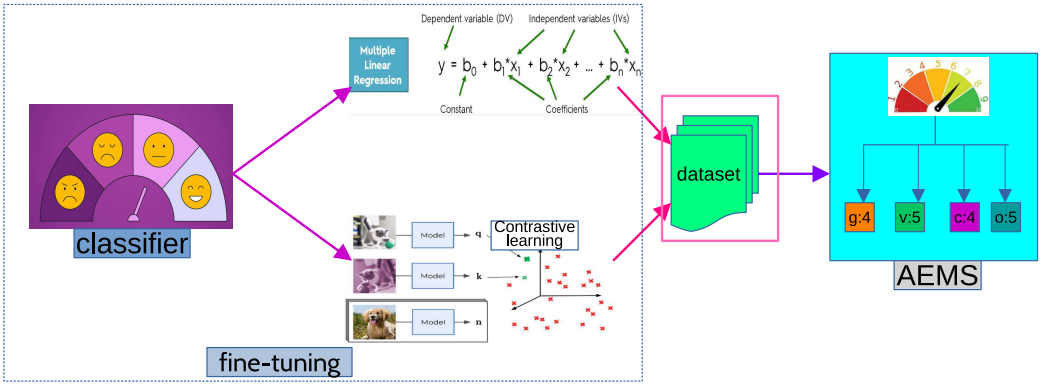

Figure 1: The roadmap of developing AEMU in the present study

2.2 Fine-tuning existing Models

Given that scoring is a classification task, we can fine-tune existing high-performance models to tailor them for the AES purpose. To validate our strategies, we selected two existing models for fine-tuning:

The RoBERTa-base model is good at making classifications, and we can take advantage of this classifier to fine-tune. The RoBERTa works well in making fine-tuning. `DistilBERT` is a smaller, more efficient language representation model pre-trained using knowledge distillation. This method reduces the BERT model size by 40%, retains 97% of its language understanding capabilities, and is 60% faster. The model employs a triple loss function combining language modeling, distillation, and cosine-distance losses. Through proof-of-concept and on-device studies, DistilBERT has demonstrated its efficacy as a cost-effective solution for high-quality NLP tasks in constrained environments.

In order to make these models suitable for AES, we fine-tune them to scoring essay quality across various dimensions such as syntax, vocabulary, and coherence. We used multiple regression and contrastive learning aim to enhance their ability to assess and score essays more comprehensively. The tech details were described above. The tech roadmap is seen in Fig. 1. The following sections will detail the datasets used, and our fine-tuning process.

2.3 Training datasets

To support our research, we utilize two comprehensive datasets. The following details the two datasets for training and testing.

The ELLIPSE Corpus is a freely available resource containing approximately 9,000 writing samples from English Language Learners (ELL) [3]. These samples are scored for overall holistic language proficiency and analytic proficiency in various dimensions, including cohesion, syntax, vocabulary, phraseology, grammar, and conventions. Additionally, the corpus provides individual and demographic

information about the writers, such as economic status, gender, grade level (8-12), and race/ethnicity. Developed to support research in corpus and NLP approaches, the ELLIPSE Corpus offers detailed language proficiency scores, aiding in the assessment of both overall and specific aspects of language proficiency. For our research, the training dataset consists of 8,100 essays, while the test dataset includes 900 essays.

The second dataset comprises essays from the International English Language Testing System (IELTS). This dataset includes the prompt, the essay itself, and scores across six dimensions: task achievement, coherence and cohesion, vocabulary, grammar, and an overall score. This dataset is particularly valuable as it reflects real-world scoring criteria used in high-stakes language assessments. For our purposes, the training dataset includes 14,500 essays, and the test dataset contains 2,000 essays.

By leveraging these datasets, we aim to fine-tune our models to provide detailed, multi-dimensional scoring of essays. The following sections will detail our methodologies for fine-tuning, the specific strategies we employed, and the results of our experiments.

2.4 Evaluation criteria

Following these strategies, we employed either of the selected models and utilized comprehensive training datasets to fine-tune them into the AEMU systems. Upon completing the model training, we validated the model's effectiveness using the corresponding test datasets. The evaluation standards included the following metrics: `precision`, `F1 score`, and `Quadratic Weighted Kappa (QWK)` for each dimension in each test dataset. This means that each essay in the test dataset was evaluated across various dimensions—such as vocabulary and grammar—using all three criteria. For example, in the first test dataset, the dimension "vocabulary" was assessed based on precision, F1 score, and QWK. Similarly, the "grammar" dimension was also evaluated using these same three criteria. This comprehensive evaluation approach ensures that the models' performance is thoroughly assessed from multiple angles, providing a detailed understanding of their strengths and weaknesses in scoring different aspects of essay quality.

Overall, to facilitate an explicit comparison, we employed the same fine-tuning strategies on two different existing models— two training datasets (ELLIPSE and IELTS). This resulted in four distinct models. We also made comparison with the previous work.

3 Results

3.1 Study 1

Following the similar strategies to fine-tune and retrain the two existing models on the same two datasets, we obtained two AEMU models. The first one is termed `RoAEMS` (RoBERT-based Automatic Essay Multi-dimensional Scoring). Table 1 provides the results on the the first model on the test dataset of ELLIPSE. The three criteria were taken to evaluate the model performance. The overall performance is beyond 0.8.

Table 1: The performance of two models in the test dataset of ELLIPSE using the three criteria

Dimension	**Precision**		**F1 Score**		**QWK**	
	Ro-AEMS	**Distil-AEMS**	**Ro-AEMS**	**Distil-AEMS**	**Ro-AEMS**	**Distil-AEMS**
Cohesion	0.85	0.84	0.86	0.85	0.81	0.82
Syntax	0.89	0.91	0.92	0.93	0.83	0.85
Vocabulary	0.90	0.90	0.89	0.90	0.84	0.85
Phraseology	0.83	0.85	0.84	0.85	0.80	0.81
Grammar	0.87	0.89	0.90	0.91	0.83	0.84
Conventions	0.91	0.82	0.92	0.92	0.84	0.85

3.2 Study 2

Table 2 provides the results on the the second model on the test dataset of IELTS. Note that there are five different dimensions in scoring. The same three criteria were taken to evaluate the model performance. The overall performance is also beyond 0.8. Compared with Table 1, the model performance has slightly improved, probably because of the larger size of the training dataset and reduced number of scoring tasks (only five.). The information on hyperparameters and other setups during training is seen in **Appendix B**.

Table 2: The performance of two models in the test dataset of IELTS using the three criteria

Dimension	**Precision**		**F1 Score**		**QWK**	
	Ro-AEMS	**Distil-AEMS**	**Ro-AEMS**	**Distil-AEMS**	**Ro-AEMS**	**Distil-AEMS**
Task Achievement	0.90	0.91	0.90	0.9	0.86	0.87
Coherence and Cohesion	0.89	0.88	0.91	0.90	0.86	0.85
Vocabulary	0.93	0.92	0.94	0.93	0.88	0.87
Grammar	0.91	0.89	0.89	0.87	0.84	0.85
Overall	0.89	0.88	0.91	0.89	0.85	0.86

4 Discussion

Table 3 lists the main achievements of various models on holistic scoring. Compared to these models, our model demonstrated better performance in holistic scoring based on the criterion of `QWK`. Additionally, our model can provide scores for multiple dimensions, with performance in these dimensions closely matching the overall scoring.

Table 3: QWK performance of in the past AES models and closed-sourced models for overall scoring

Task	**QWK Average (overall scoring)**
Xie et al. (2022)	0.82
Jiang et al. (2023)	0.70
ChatGPT (Mansour et al., 2024)	0.31
Llama (Mansour et al., 2024)	0.30
GPT-4V (Lee et al., 2023)	0.43

Furthermore, the two models exhibited similar performance across two types of test datasets, despite the differences in their dimensions. This cross-validation indicates that our models maintain stable performance across diverse datasets. It also demonstrates that our strategies are highly effective and robust in enhancing the AEMS system. The consistent results across varied datasets highlight the generalizability of our approach, ensuring reliable scoring under different conditions and validating the robustness of our model improvements.

Data Availability

One AEMS model for the ELLIPSE corpus is available at: `https://huggingface.co/Kevintu/Engessay_grading_ML`. The usage of the model is detailed in this huggingface repo.

References

[1] Jose Camacho-Collados, Kiamehr Rezaee, Talayeh Riahi, Asahi Ushio, Daniel Loureiro, Dimosthenis Antypas, Joanne Boisson, Luis Espinosa-Anke, Fangyu Liu, Eugenio Martínez-Cámara, et al. Tweetnlp: Cutting-edge natural language processing for social media. *arXiv preprint arXiv:2206.14774*, 2022.

[2] Hongbo Chen and Ben He. Automated essay scoring by maximizing human-machine agreement. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1741–1752, 2013.

[3] Scott Crossley, Yu Tian, Perpetual Baffour, Alex Franklin, Youngmeen Kim, Wesley Morris, Meg Benner, Aigner Picou, and Ulrich Boser. The english language learner insight, proficiency and skills evaluation (ellipse) corpus. *International Journal of Learner Corpus Research*, 9(2):248–269, 2023.

[4] Ronan Cummins, Meng Zhang, and Ted Briscoe. Constrained multi-task learning for automated essay scoring. Association for Computational Linguistics, 2016.

[5] Fei Dong and Yue Zhang. Automatic features for essay scoring–an empirical study. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1072–1077, 2016.

[6] Fei Dong, Yue Zhang, and Jie Yang. Attention-based recurrent convolutional neural network for automatic essay scoring. In *Proceedings of the 21st conference on computational natural language learning (CoNLL 2017)*, pages 153–162, 2017.

[7] Zhiwei Jiang, Tianyi Gao, Yafeng Yin, Meng Liu, Hua Yu, Zifeng Cheng, and Qing Gu. Improving domain generalization for prompt-aware essay scoring via disentangled representation learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12456–12470, 2023.

[8] Zixuan Ke, Winston Carlile, Nishant Gurrapadi, and Vincent Ng. Learning to give feedback: Modeling attributes affecting argument persuasiveness in student essays. In *IJCAI*, pages 4130–4136, 2018.

[9] Zixuan Ke, Hrishikesh Inamdar, Hui Lin, and Vincent Ng. Give me more feedback ii: Annotating thesis strength and related attributes in student essays. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 3994–4004, 2019.

[10] Zixuan Ke and Vincent Ng. Automated essay scoring: A survey of the state of the art. In *IJCAI*, volume 19, pages 6300–6308, 2019.

[11] Ehsan Latif and Xiaoming Zhai. Fine-tuning chatgpt for automatic scoring. *Computers and Education: Artificial Intelligence*, page 100210, 2024.

[12] Gyeong-Geon Lee, Ehsan Latif, Xuansheng Wu, Ninghao Liu, and Xiaoming Zhai. Applying large language models and chain-of-thought for automatic scoring. *Computers and Education: Artificial Intelligence*, page 100213, 2024.

[13] Watheq Mansour, Salam Albatarni, Sohaila Eltanbouly, and Tamer Elsayed. Can large language models automatically score proficiency of written essays? *arXiv preprint arXiv:2403.06149*, 2024.

[14] Huy Nguyen and Diane Litman. Argument mining for improving the automated scoring of persuasive essays. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[15] Isaac Persing, Alan Davis, and Vincent Ng. Modeling organization in student essays. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 229–239, 2010.

[16] Dadi Ramesh and Suresh Kumar Sanampudi. An automated essay scoring systems: a systematic literature review. *Artificial Intelligence Review*, 55(3):2495–2527, 2022.

[17] Margarita Rodríguez-Ibánez, Antonio Casánez-Ventura, Félix Castejón-Mateos, and Pedro-Manuel Cuenca-Jiménez. A review on sentiment analysis from social media platforms. *Expert Systems with Applications*, page 119862, 2023.

[18] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[19] Swapna Somasundaran, Jill Burstein, and Martin Chodorow. Lexical chaining for measuring discourse coherence quality in test-taker essays. In *Proceedings of COLING 2014, the 25th International conference on computational linguistics: Technical papers*, pages 950–961, 2014.

[20] Kaveh Taghipour and Hwee Tou Ng. A neural approach to automated essay scoring. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1882–1891, 2016.

[21] Sowmya Vajjala. Automated assessment of non-native learner essays: Investigating the role of linguistic features. *International Journal of Artificial Intelligence in Education*, 28:79–105, 2018.

[22] Yongjie Wang, Chuan Wang, Ruobing Li, and Hui Lin. On the use of bert for automated essay scoring: Joint learning of multi-scale essay representation. *arXiv preprint arXiv:2205.03835*, 2022.

[23] Jiayi Xie, Kaiwei Cai, Li Kong, Junsheng Zhou, and Weiguang Qu. Automated essay scoring via pairwise contrastive regression. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2724–2733, 2022.

[24] Ruosong Yang, Jiannong Cao, Zhiyuan Wen, Youzheng Wu, and Xiaodong He. Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking. In

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 1560–1569, 2020.

[25] Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. A new dataset and method for automatically grading esol texts. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 180–189, 2011.

[26] Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Jialin Pan, and Lidong Bing. Sentiment analysis in the era of large language models: A reality check. *arXiv preprint arXiv:2305.15005*, 2023.

Appendix A: Incorporating contrastive learning for additional information

Contrastive learning aims to make representations of similar inputs closer in the embedding space, while pushing representations of dissimilar inputs apart.

Let:

- $\text{BERT}_{\text{cls}}(x)$ be the output representation of the [CLS] token for input x.
- $\text{BERT}_{\text{additional}}(r)$ be the representation of additional information r (e.g., essay requirements, topic, type).
- $\mathbf{z}_i$ be the combined representation of the essay and its additional information.

The combined representation $\mathbf{z}_i$ is obtained by concatenating the essay representation and the additional information representation:

$$\mathbf{z}_i = \text{BERT}_{\text{cls}}(x_i) \oplus \text{BERT}_{\text{additional}}(r_i)$$

where $\oplus$ denotes concatenation.

For contrastive learning, we define the NT-Xent (Normalized Temperature-scaled Cross Entropy) loss. Let $(\mathbf{z}_i, \mathbf{z}_j)$ be a positive pair (e.g., essays with the same topic), and let τ be a temperature parameter. The NT-Xent loss for a positive pair is given by:

$$\mathcal{L}_{\text{contrastive}}(\mathbf{z}_i, \mathbf{z}_j) = -\log \frac{\exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_k)/\tau)}$$

where $\text{sim}(\mathbf{z}_i, \mathbf{z}_j)$ denotes the cosine similarity between $\mathbf{z}_i$ and $\mathbf{z}_j$, and N is the batch size.

The overall loss function combines the classification loss, regression loss, and contrastive loss:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda_1 \mathcal{L}_{\text{MSE}} + \lambda_2 \mathcal{L}_{\text{contrastive}}$$

where:

- $\mathcal{L}_{\text{CE}}$ is the cross-entropy loss for classification.
- $\mathcal{L}_{\text{MSE}}$ is the mean squared error loss for regression.
- $\mathcal{L}_{\text{contrastive}}$ is the contrastive loss.
- λ_1 and λ_2 are hyperparameters that balance the contributions of the different loss components.

By optimizing this combined loss function, the model can better understand and utilize additional information such as essay requirements, topics, and types, leading to improved automatic essay scoring.

Appendix B: Hyperparaters and setups in training

Tables 4 and 5 provides the information on some of the hyperparameters in training based on the two existing models. The full hyperparameters are given on request.

Table 4: Training configuration parameters in RoBERT

Parameter	**Value**
num_train_epochs	28
per_device_eval_batch_size	16
warmup_steps	500
logging_steps	10
evaluation_strategy	epoch
learning_rate	2e-5
contrastive_learning_batch_size	128

Table 5: Training configuration parameters in DistilBERT

Parameter	**Value**
num_train_epochs	22
per_device_eval_batch_size	64
warmup_steps	500
logging_steps	10
evaluation_strategy	epoch
learning_rate	2e-5
contrastive_learning_batch_size	130