Automatic Assessment of Student Answers using Large Language Models: Decoding Didactic Concepts

Daniel Schönle
IDACUS Insitute
Furtwangen University
Furtwangen, Germany
email:schonledanielhfu@gmail.com

Christoph Reich
IDACUS Insitute
Furtwangen University
Furtwangen, Germany
email:christoph.reich@hs-furtwangen.de

Djaffar Ould Abdeslam
Institut IRIMAS
Université de Haute Alsace
Mulhouse, France
email:djafar.ould-abdeslam@uha.fr

Daniela Fiedler
Department of Science Education
University of Copenhagen
Copenhagen, Denmark
email:dfiedler@ind.ku.dk

Ute Harms
Department of Biology Education
IPN - Leibniz Institute for
Science and Mathematics Education
Kiel, Germany
email:harms@leibniz-ipn.de

Johannes Poser
Department of Biology Education
IPN - Leibniz Institute for
Science and Mathematics Education
Kiel, Germany
email:poser@leibniz-ipn.de

***Abstract*—This study evaluates machine learning for automating the evaluation of textual responses in virtual learning environments, particularly by applying advanced linguistic enhancement techniques. Techniques such as Transformer-based data augmentation, Part-of-Speech enhanced feature selection, and LinPair tokenisation were employed. The evaluation focused on classification quality and training efficiency using a synthetically created question-and-answer dataset, characterised by its limited sample size, extensive class range, and the complexity of identifying didactical elements. The findings indicate that while the Support Vector Machine (SVM) consistently outperforms the distilled version of the large language model Bidirectional Encoder Representations from Transformers (DistilBERT) in quality metrics, the integration of linguistic elements improved DistilBERT's performance significantly—achieving a 7.62% increase in F1-Score and a 17.02% rise in Hamming-Score. Despite these gains, DistilBERT recorded lower efficiency scores compared to SVM. This suggests that while SVM excels with synthetic data, Large Language Models demonstrate substantial potential in processing complex linguistic data when provided with linguistic information. These insights confirm the viability of both approaches as effective tools for automated assessment in educational settings.**

Keywords-machine learning; efficiency; linguistic; text classification; assessment.

I. Introduction

The advent of Virtual Learning Environments (VLE) marks a profound shift in educational paradigms, driven by the fusion of digital technologies and Machine Learning (ML) algorithms. This shift addresses the growing demand for educational experiences that are accessible, adaptable, and personalised to meet the needs of a diverse global learner population [1][2]. Sophisticated ML techniques enable VLEs to analyse learner data and deliver personalised content along with adaptive learning paths, significantly improving engagement and outcomes. The instrumental role of ML in fostering this adaptivity is paramount, as it dynamically refines content and pedagogical approaches based on learner interactions, optimising the educational pathway [3].

The typical interaction between teacher and student during the learning process is illustrated in Figure 1. When answering textual diagnostic questions, students provide open-text responses. The automation of diagnostic responses can be effectively integrated by analysing these open-text responses based on both the content of the student's response and the underlying didactic principles embedded within it. This integration facilitates a more nuanced understanding of student understanding and learning needs. This study evaluates the use of ML, especially Large Language Models (LLM), to automate the evaluation of text in VLEs. This research highlights the usefulness of advanced configurations in real-world educational settings by comparing established state-of-the-art methods with innovative techniques, such as LLM-based data augmentation, LLM-based text classification, Part-of-Speech enrichment (POS-Enrichment), *LinPair* Tokenization [4], and *UnImportant-Part-of-Speech* (UIP) feature selection [5]. Challenges related to the training dataset include small sample sizes, often reflecting data scarcity, the use of artificially created curated datasets, and the complexity of accurately identifying nuanced labels. Method setups are evaluated based on quality by F1-Score [6] and Hamming Loss [7]. A particular focus was on the integration of *LinPair*-Tokenization, *UIP* feature selection, and a quality-focused evaluation of efficiency via the *COmpact Efficiency* (CO) score [8]. This approach is novel in this domain.

The *Teacher questions and student answers for the SCRBio in the context of evolution* (QASCRBio) dataset [9], integral to the FiSK-Research-Project within the domain of didactic science, forms the foundation of this research. It includes questions, student responses, and corresponding assessment results as specific didactical diagnostic aspects. These elements are used for text classification to identify didactic attributes within student answers, which can be used for automated formative feedback or as assistive information for educators. The study presents a reliable setup for the automated assessment of

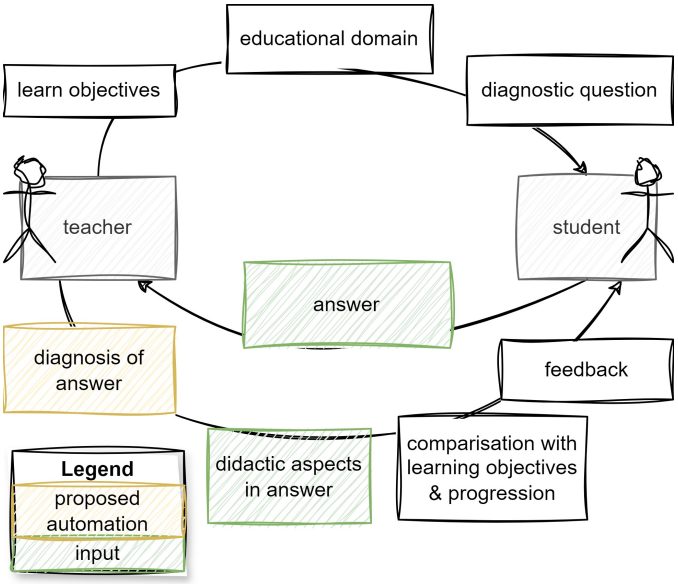

Figure 1. Context of Automatisation.

students' responses that can be used and is of value in real-life educational scenarios.

A. Didactic Background

In education, assessment tasks are crucial for improving the learning and teaching process. Constructing questions, underpinned by sound educational theory and practice, is a fundamental mechanism for diagnosing student understanding, revealing misconceptions and encouraging more profound engagement with the subject matter. Figure 2 provides an overview of the assessment and feedback process. Diagnostic questions are designed to reveal the basis of students' responses, highlighting correct and incorrect thinking patterns. They are particularly adept at identifying misconceptions by using distractors that target known misconceptions, providing insights into students' conceptual understanding [10]. Basic characteristics and theoretical frameworks of practical assessment questions have been established, drawing on the seminal contributions of scholars such as Popham (1995) [11], Brookhart (2017) [10], Black and Dylan (1998) [12].

B. Structure

This paper is organised into six sections. (i) Introduction; (ii) Related work, reviewing relevant e-learning and assessment literature; (iii) Automated didactic assessment, providing definitions, research questions, a description of the methodology and limitations of the work; (iv) Experiment, elaborating on the approach including the data set, implementation and evaluation; (v) Discussion, offering insights and implications of the findings; and (vi) Conclusion, reflecting on the broader implications and suggesting avenues for further research.

II. Related Work

This section provides an overview of the application of ML in VLEs. First, a summary of existing research on ML applications in VLEs is presented, positioning this research within the broader landscape of technological interventions. The focus then shifts to examining approaches that emphasise educational assessment automation, highlighting the progressive integration of ML to streamline and improve assessment processes. Finally, the discussion extends to the study of simulated learning environments.

A. Machine Learning in Virtual Learning Environments

The integration of ML into VLEs has been increasingly recognised for its potential to tailor education to individual learning needs, a concept referred to as precision education. Luan and Tsai systematically reviewed 40 empirical studies, revealing a focus on predicting student performance and dropout rates within online or blended learning settings, particularly among students in Science, Technology, Engineering and Mathematics (STEM) fields [1]. Dogan et al. conducted a systematic review on the use of Artificial Intelligence (AI) in online learning and distance education, noting a significant increase in research, with substantial contributions from China, India, and the United States. Their analysis identified three dominant clusters of research themes, underscoring the versatility of AI in enhancing online teaching, learning processes, and personalisation [13].

B. Automated educational assessment

In their comprehensive survey, Das et al. examine the burgeoning field of automatic question generation and answer assessment, pivotal for enhancing learning through internet-based platforms [14]. The study aggregates and critiques a decade's worth of research, elucidating the state-of-the-art techniques that automate the creation and evaluation of questions across textual, pictorial learning resources. The survey underscores the growing integration of such methodologies in intelligent education systems, reflecting on their potential to transform self-paced learning by identifying learning gaps effectively. This synthesis of past and current methodologies provides a critical baseline for future explorations in automated educational assessments.

The systematic review by González-Calatayud et al. [15] delves into the use of artificial intelligence in student assessments, analysing data from over 450 papers to discern the impact and implications of AI on educational practices. The review reveals a marked focus on formative assessment and grading, albeit with a noted deficiency in pedagogical integration within the AI applications reviewed. Highlighting the need for educational models that synergise with technological advancements, this work calls for enhanced teacher training and research that bridges the gap between AI capabilities and pedagogical needs, ensuring that AI supports rather than supplants the educational process.

INCEpTION, a novel annotation platform detailed by Klie et al., integrates ML to support and enhance the annotation

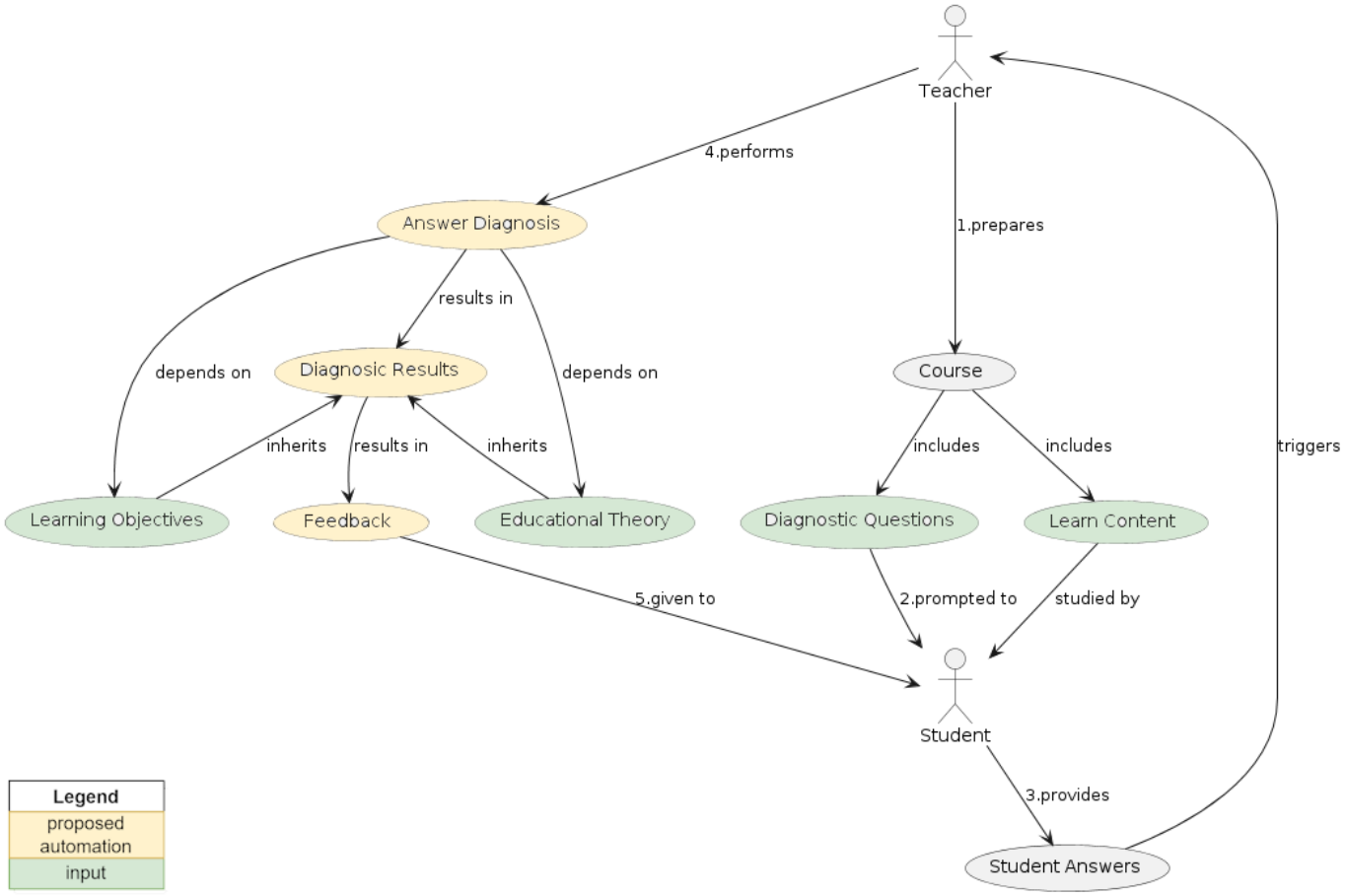

Figure 2. Use Case of Assessment and Feedback.

process. Tailored for semantic tasks like concept linking and semantic frame annotation, INCEpTION addresses the complex demands of creating high-quality annotated corpora by incorporating active learning and entity linking. This platform is designed to be adaptable across various fields, demonstrating its utility in collecting and managing domain-specific knowledge through an interactive, machine-assisted environment. This innovation represents a significant step forward in the semantic annotation domain, offering robust support for researchers and annotators alike [16].

Hartmann et al. compare ten text classification methods to understand their efficacy in analysing social media content for marketing applications. Their empirical study identifies Naive Bayes (NB) [17] and Random Forest [18] as superior in aligning with human intuition over traditional methods like Support Vector Machine (SVM) [19] and lexicon-based approaches. By demonstrating the relative performance of these methods across varied datasets, this research provides valuable insights into the optimisation of text classification in marketing, suggesting a pivot towards more dynamic and statistically robust methods [20]. Das et al. presented a survey of automatic question generation and assessment strategies from textual and pictorial learning resources, emphasising the importance of assessment systems in identifying learning gaps [14]. Furthermore, González-Calatayud, Prendes-Espinosa, and Roig-Vila analysed the application of AI in student assessment, revealing a prevalent focus on formative evaluation and the necessity for pedagogical grounding in AI applications [15].

C. Simulated Learning Environments

In the field of simulated learning environments, incorporating digital technologies offers novel ways to improve pre-service biology teachers' pedagogical skills, particularly in diagnostic competence. Fiedler et al. explored this through a classroom simulation integrated with a chatbot designed to enhance teachers' ability to accurately assess students' understanding of evolutionary processes - a crucial skill given the scarcity of practical teaching opportunities in university settings [21]. Their research showed that while participants were able to diagnose clear, naive, or scientific explanations, they struggled with mixed model explanations and identifying specific misconceptions, highlighting the need for targeted feedback to refine diagnostic strategies.

Adelana et al. [22] explored pre-service biology teachers' attitudes and intentions towards using AI-based intelligent tutoring systems for teaching genetics, a subject known for its teaching challenges. Through the Theory of Planned Behaviour lens, their study highlighted the influence of perceived usefulness and subjective norms on these teachers' behavioural intentions while noting the non-significant impact of perceived behavioural control on such intentions. Furthermore, the research highlighted gendered nuances in subjective norms, particularly among female pre-service teachers, while revealing consistent attitudes across other dimensions. This research not only highlights the importance of social norms and attitudes in adopting AI technologies in education but also points to broader implications for integrating AI in the promotion of effective teaching strategies in science. The classroom simulation *Simulated Classroom Biology* (SCRBio) [23] demonstrates the validation of action-oriented pedagogical content knowledge of pre-service biology teachers, focusing specifically on evolution education.

Rogers et al. [24] explore the educational potential of Virtual Reality (VR) technologies in STEM education through the operation of a virtual CNC milling machine. Their study evaluates the usability and pedagogical effectiveness of immersive VR environments, providing evidence from usability studies that highlight the benefits of such technologies in enhancing hands-on learning without physical constraints. The findings suggest that VR can significantly enhance the educational experience by providing intuitive and engaging ways to learn complex machine operations, marking a significant step forward in the integration of immersive technologies in education.

III. Automatic assessment

Automatic assessment in e-learning aims to accurately assess learner responses using computational methods, thereby increasing the scalability and efficiency of educational systems. This is achieved by integrating ML algorithms that automate the assessment process, thereby reducing the burden on instructors and providing timely feedback to learners.

Information Sources:

1) Learner input: Primary data includes textual responses, quiz results and interactive logs that capture learner interactions within the e-learning environment.
2) Instructional materials: Secondary sources include the instructional content against which learner responses are assessed, including guidelines for correct answers and grading rubrics.
3) Historical data: Archived assessments and their results contribute to the training of ML models, enabling them to learn from past instructional scenarios.

Techniques:

1) Text classification: Used to classify open-ended responses into predefined response categories or to identify thematic consistencies within learner submissions.
2) Natural Language Processing (NLP): Uses linguistic analysis to understand and assess the quality of text responses, focusing on grammar, relevance and content accuracy.
3) ML algorithms: Applies techniques such as supervised learning to recognise patterns in responses and unsupervised learning to discover underlying patterns in unstructured data.
4) Feedback generation: Algorithms generate automated feedback based on assessment results tailored to individual learner needs and performance, supporting personalised learning pathways.

A. Research Questions

This study investigates automatic assessment in eLearning frameworks where free text input needs to be evaluated. The expected training data consists of textual responses accompanied by diagnosed labels of corresponding assessment results.

RQ1 Which machine learning methods offer the best performance and efficiency for automated assessment? This question focuses on selecting pre-processing, tokenisation, and classification approaches that enable automated assessment of responses to learning content.

RQ2 What are the indicators of quality and efficiency in automated assessment? This question aims to select benchmarks for measuring the quality and efficiency of automated assessment processes.

RQ3 What factors significantly influence the performance of automated assessment? This question examines the factors that are crucial in determining the performance of automated assessment tools. These factors include algorithmic, data attributes, and contextual variables.

This research aims to examine the challenges and opportunities related to the automatic assessment of e-learning data. It is assumed that progress in ML could significantly contribute to the evolution of digital education.

B. Methodology

This research proposes and evaluates new technologies for automated assessment through empirical validation. The methodology involves preparing the dataset, applying ML algorithms and critically analysing the results to validate the effectiveness of these technologies. A publicly available dataset that represents the real-world conditions in which the technology will be used is selected. The design and conduct of the experiment is documented and justified. Systematic evaluation allows for a detailed comparison with traditional methods, highlighting potential accuracy, efficiency and scalability improvements in educational assessment.

C. Limitations

This study focuses on scenarios that require the evaluation of free text input. The validation of the methods used is empirical and depends on the parameters of the experiment. To mitigate this dependency, a real-world dataset is selected, accompanied by a variety of methods for pre-processing, tokenisation and classification. The behaviour of the appliance may differ in response to alternative use cases or datasets, depending on the specific requirements, the text and the quality of the labels. Furthermore, performance and efficiency results

TABLE I
QASCRBIO DATASET EXCERPT

QG[1] German	QL[1] Labels	Q[2] English	Q_W Word-List Word	Q_L Lemmatized Lem	Q_{LP}[3] Token+POS-Tag LemPair	Q_{WP}[3] Token+POS-Tag WordPair	Q_{WP-Us}[4] UIP Feature Selection Us Us
Die Natur bewirkte die Veränderung beim See-pferdchen	F1	Nature brought about the change in the seahorse	'Nature', 'brought', 'about', 'the', 'change', 'in', 'the', 'seahorse'	'nature', 'bring', 'about', 'the', 'change', 'in', 'the', 'seahorse'	'nature_NN-nsubj', 'bring_VBD', 'about_RP', 'the_DT', 'change_NN-dobj', 'in_IN', 'the_DT', 'seahorse_NN-pobj'	'Nature_NN-nsubj', 'brought_VBD', 'about_RP', 'the_DT', 'change_NN-dobj', 'in_IN', 'the_DT', 'seahorse_NN-pobj'	'Nature_NN-nsubj', 'brought_VBD', 'about_RP', 'change_NN-dobj', 'in_IN', 'seahorse_NN-pobj'

Sample Excerpt of QASCRBio dataset along with the results of the pre-processing variants.
[1] QASCRBio dataset [9], [2] Deepl-Translator [25], [3] UIP feature selection [5], [4] LinPairTokenization [4]

TABLE II
QASCRBIO DATASET LABELS

Principle			Threshold			Misconception		
P1_Variability	P2_Inheritance	P3_Selection	T1_Chance	T2_Probability	T3_Time	F1_Anthropomorphic	F2_Teleological	F3_Usage

QASCRBio dataset [9]: Multilabel-Dataset: single label or multiple labels per sample

may be influenced by the host setup. This study does not extend to the subsequent application of assessment results, such as feedback determination or generation, nor does it explore integration with learner models for predicting student profiles.

IV. EXPERIMENT

The use of automated assessment in educational settings requires the execution of several software engineering steps. First, a text classification system is designed to facilitate the assessment process. Next, a relevant data set is carefully curated. The implementation phase involves setting up an automated system to accurately process and classify student responses. The culmination of this process is the systematic evaluation, where the effectiveness and accuracy of the automated assessment are rigorously tested to ensure its reliability and pedagogical utility. This methodological approach ensures a robust framework for integrating automated assessment tools into educational contexts, improving student assessment's efficiency and accuracy.

A. Classification Procedure

The design of the classification procedure is targeted at the presentation and evaluation of automatic evaluation using innovative techniques (Figure 3). This includes an extensive pre-processing phase that integrates state-of-the-art methods to optimise the input data for subsequent classification. The key pre-processing steps are (i) selective feature selection, which focuses on removing text segments that are not considered essential for the classification objectives, and (ii) information enrichment strategies, which enhance the dataset by incorporating Part-Of-Speech (POS) tags to provide syntactic context. The classification process uses sophisticated tokenisation techniques designed to minimise data loss. This is followed by the application of selected classification algorithms that categorise the text according to the pre-trained labels. The

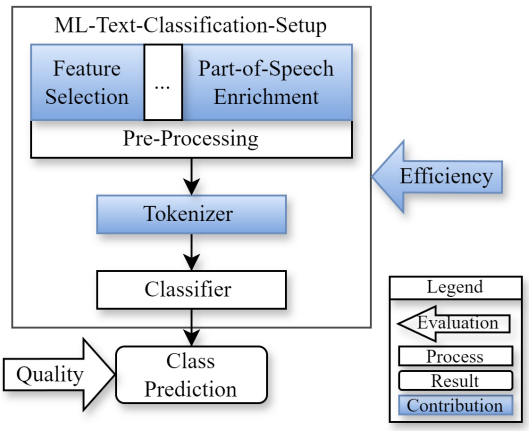

Figure 3. Text Classification Setup Overview.

overall approach ensures a robust framework for tackling complex text classification challenges in diverse applications.

B. Dataset

This study investigated the automation of the didactic diagnosis process for German university student responses using the QASCRBio dataset. A diagnostic question was used to assess the students' learning outcomes (Table I). The results are not suitable for grading but for providing formative feedback. The diagnostic aspects are divided into nine labels, reflecting the analytical challenges of sentiment analysis. These aspects are grouped into three categories (Table II). The main engineering objective was to classify the texts into one or more classes, overcoming the challenges posed by a dataset limited to 540 samples, a multi-classification problem and the complicated detection complexity of the labels. The dataset

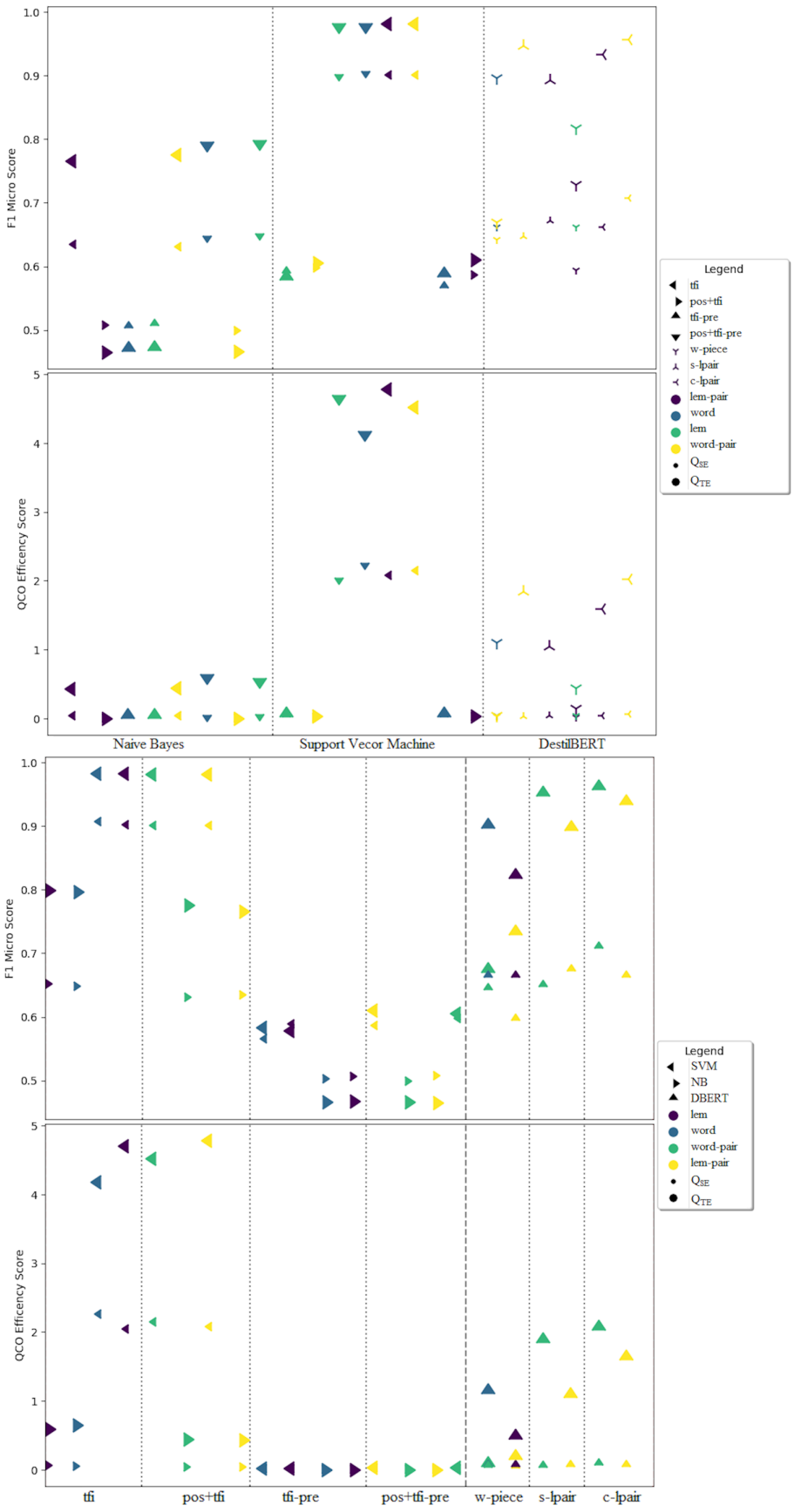

Figure 4. Quality and Efficiency of Classifiers and Tokenizers.

TABLE III
UNIMPORTANT PART-OF-SPEECH LISTS

Origin	UIP-List	POS-Tags	Description
UIP for NB	U_N	UH;NNPS	Interjection;Proper Nouns, Plural
UIP for SVM	U_S	WP;JJS;RP;CC;WRB;EX; MD;PRP;WDT;IN;TO; POS;JJ;VBP;JJR;NNS	Wh-pronoun; Superl. Adjective; Particle; Coord. Conjunction; Wh-adverb; Exist. There; Modal; Pers. Pronoun; Wh-determiner; Prep. Conjunction; To; Poss. Ending; Adjective; Verb, Sing. Present; Comp. Adjective; Plural Noun
Linguistic Set 1	U_X	DT	Determiner
Linguistic Set 2	U_Y	DT;IN;CC	Determiner; Preposition or Subordinating Conjunction; Coordinating Conjunction

used for research was created by specialists in the didactic field who synthetically generated and annotated the texts.

TABLE IV
TOKENIZERS

Tokenizer	Abbrev.	Elements	Method
WordPiece[26]	w-piece	tokens	Split into subtokens
SmartLinPair[4]	s-slpair	pairs of token and POS-tag	Split into subtokens, use POS-tag on OOV
CompleteLinPair[4]	c-lpair	pairs of token and POS-tag	Split into subpairs

TABLE V
DEFINITION OF DIMENSIONS

Dimension	Measurements & Scores	Weight
Quality	(F1MacroScore + F1MicroScore + Hamming) / 3	6
Work	FLOPS [count] / DatasetSize [kB]	1
Space	AverageRSS [MB] * DurationTime [s]	1
Duration	DurationTime [s]	1

TABLE VI
HOST-SETUP

No.	Type	CPU-Model	Clock	Threads	RAM
1	Virtualised	AMD EPYC 7742	2,2 GHz	16	32 GB

OS: Linux Ubuntu 22, Language: Python3.10,
Libraries: Scikit-learn [27], DistilBERT [28], torch [29], pandas [30].

C. Implementation

Data augmentation was performed to overcome the size limitations of the dataset (RQ3) using Transformer-based text translation methods: DeeplTranslator [25], a combination of DeeplTranslator and DeeplWrite, and Google Translator [31]. Two datasets were created: QASCRBio-SingleEnglish (QASCRBioSE) with the Google Translator, consisting of 432 training and 108 test samples; and QASCRBio-TripleEnglish (QASCRBioTE) with all three translations, consisting of 1296 training and 324 test samples after removing duplicates.

The pre-processing included lemmatising, POS-tagging, and feature selection, resulting in seven text variants: original text, lemmatised text - and both texts with added POS information (Table I). For feature selection, UIP [5] was used to select tokens based on their importance in English and specific classifiers. In addition to the two available UIP lists for NB and SVM, two standard sets (x, y) were used (Table III). The tokenisation methods selected were Term Frequency–Inverse Document Frequency (TF-IDF) [32], WordPiece [26], and the CompleteLinPair and SmartLinPair tokenisation techniques from the LinPair framework [4] (Table IV). The classifiers were chosen to compare the fast models SVM and NB with the more sophisticated Large Language Model, DistilBert (DBERT) [28].

D. Evaluation

In response to RQ2, F1-Micro, F1-Macro, and the Hamming score were used to evaluate the quality of class prediction. The Quality-Focused Compact Efficiency Metric (QCO) from the Compact Efficiency Metrics Framework [8] was used to assess computational efficiency. QCO provides a score to compare the training effectiveness of the model and the operational efficiency as defined by the metric configuration. Training effectiveness was captured based on the measurements of the efficiency dimensions in Table V. QCO was calculated as defined by Equation (1). The impact of the UIP feature selection was evaluated by comparing the quality of the classification results and by measuring the data size savings. The computation of the implementation was performed on a system whose specifications are documented in Table VI, thus ensuring the reproducibility and reliability of the results. This comprehensive setup included two datasets, up to 3 feature selection methods, five tokenisation methods, and three classifiers mentioned above.

V. DISCUSSION

The analysis of the results, shown in Figure 4, provides valuable insights into candidate outcomes regarding research questions RQ1 and RQ3. Figure 4 consists of four plots, the top two displaying the results for each classification method and the bottom two illustrating the results associated with each tokeniser method. Regarding classification accuracy, SVM and DBERT exceeded the threshold of a 0.9 F1 Micro Score. The top-performing SVM configurations used either TF-IDF (tf) or TF-IDF with pre-processing (tfpre) and showed similar performance levels across various pre-processing methods. Notably, SVM did not show significant improvement when using POS-optimised TF-IDF tokenisers or incorporating POS tags, indicating a degree of robustness to pre-processing variations.

$$QCO(M) = \frac{((\frac{F1MAC+F1MIC+HUM}{3})^6)}{(\log_{47,5B} FLOPS/DS[kB] + \log_{3,5K} RSS[MB] * log_{3,7T} D[s] + \log_{3,7T} D[s]} * 10 \quad (1)$$

where HUM = Humming Sc., $FLOPS$ = Float. Point OP, DS = Dataset-S., RSS = Resident Set S., D = Duration

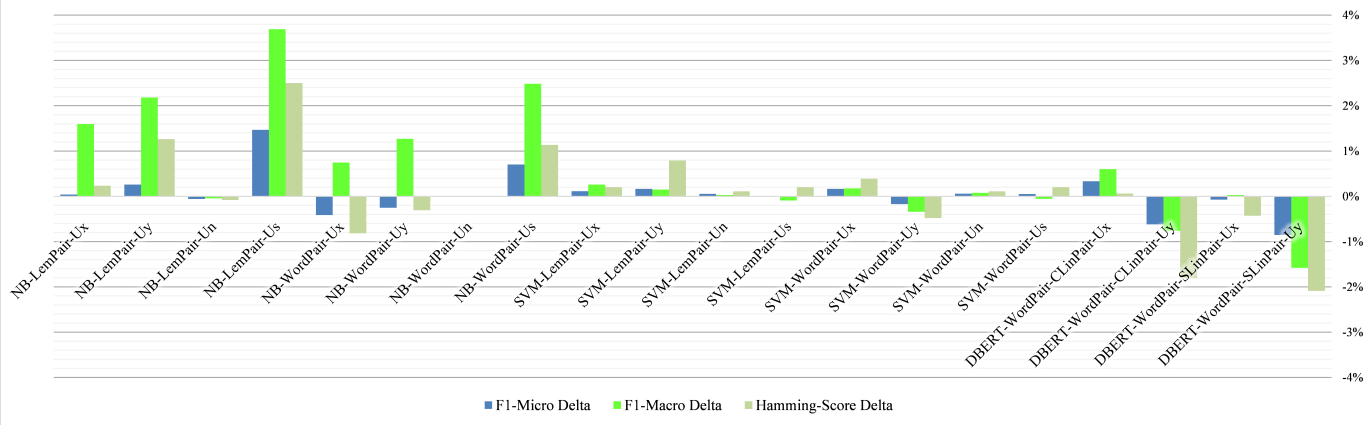

Figure 5. UIP-Effects: Relative Quality Gain against baseline without UIP feature selection.

DBERT significantly improved by data augmentation, particularly with outcomes related to the $WordPair$ pre-processing method. This indicates that performance has been enhanced in most cases by including POS information, but performance has decreased in a few cases. The LinPair-Tokenizers SmartLinPair and CompleteLinPair consistently outperformed the WordPiece Tokenizer when comparing tokenisation strategies across the board. This disparity is accentuated when taking into account the size of the dataset and the pre-processing technique. Regarding efficiency, SVM proved to be the superior option when using the QASCRBioTE dataset due to its fast processing times, low memory requirements, and high classification accuracy. DBERT exhibited its most effective results with $WordPair$ pre-processing, significantly enhancing its efficiency by applying LinPair tokenisers.

The results show significant differences between the setups regarding quality (F1 Score) and efficiency (QCO-Efficieny) as seen in Figure 4. Surprisingly, the SVM outperformed the standard DBERT model in terms of quality and efficiency. DBERT setups with innovative improvements in pre-processing, feature selection, and tokenisation achieve similar quality to SVM. POS enrichment, POS-based filtering, and LinPair tokenisation lead to quality improvements with production-ready results. SVM and DBERT benefited from Transformer-based data augmentation (Deepl-Translation), while the quality of NB deteriorated with augmentation. POS enhancement improved the results of DBERT but degraded those of SVM and NB.

LinPair is currently only available for subset-based techniques, such as DBERT. It significantly improves quality by up to 12% (Figure 6) while increasing efficiency (Figure 4). Therefore, the quality improvement compensates for the additional computation required for LinPair. This highlights the potential of customised tokenisation strategies in enhancing

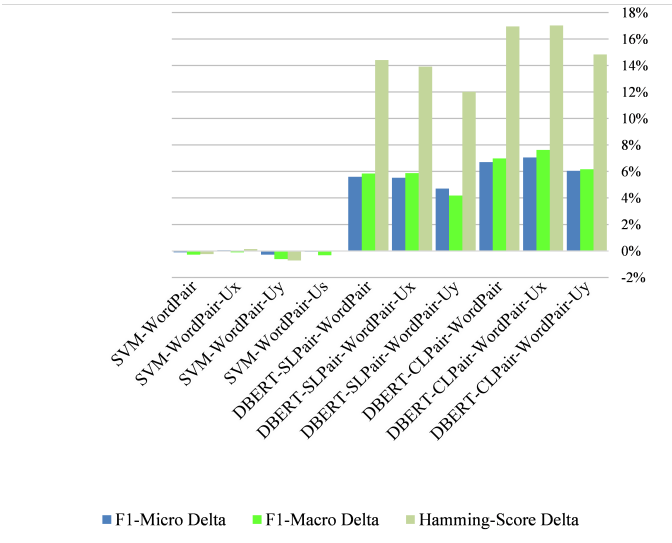

Figure 6. Overall Effects: Relative Quality Gain against baseline without POS usage.

the effectiveness of ML models in automatic student response evaluation.

The UIP feature selection's evaluation demonstrated mixed outcomes, as depicted in Figure 5. Notably, NB exhibited gains, enhancing the F1-Macro Score by up to 3.7%. Conversely, SVM showed no discernible benefits from UIP feature selection, while DBERT registered only marginal improvements when employing the minimal UIP-Set Ux with Complete LinPair. Given that DBERT is trained on complete sentences to capitalise on contextual relationships, the elimination of parts of sentences through feature selection might adversely impact its performance.

	Word	Lem	LemPair	WordPair	WordPairUn	WordPairUs	WordPairUx	WordPairUy	LemPairUn	LemPairUs	LemPairUx	LemPairUy
Word	0	-8	61	58	58	29	41	23	61	32	44	26
Lem	8	0	74	71	71	40	53	33	74	43	56	36
LemPair	-38	-43	0	-2	-2	-20	-12	-23	0	-18	-11	-22
WordPair	-37	-42	2	0	0	-18	-11	-22	2	-16	-9	-20
WordPairUn	-37	-42	2	0	0	-18	-11	-22	2	-16	-9	-20
WordPairUs	-23	-28	25	22	22	0	9	-4	25	2	11	-2
WordPairUx	-29	-34	14	12	12	-8	0	-13	14	-6	2	-11
WordPairUy	-19	-25	30	28	28	5	14	0	30	7	17	2
LemPairUn	-38	-43	0	-2	-2	-20	-12	-23	0	-18	-11	-22
LemPairUs	-24	-30	22	20	20	-2	7	-7	22	0	9	-5
LemPairUx	-30	-36	12	10	10	-10	-2	-14	12	-8	0	-12
LemPairUy	-21	-27	28	25	25	2	12	-2	28	5	14	0

Figure 7. UIP Effects on Dataset Size.

To explore potential side effects correlating with these outcomes (RQ2), variations in dataset size induced by preprocessing techniques, such as POS enrichment and feature selection, were analysed. The heatmap in Figure 7 illustrates the relative changes in dataset size, tracing the trajectory from initial word counts to lemmatised forms and subsequent modifications through POS-Enrichment and UIP feature selection. The transformations between comparable and successive processing stages are particularly telling, which align with the aggregate findings presented in Figures 4 and 6.

DBERT's response to different text processing methods was marked. The WordPair configuration yielded the most substantial quality increase by 58%, attributed to the inclusion of POS tags. All feature selection techniques generally resulted in reduced data sizes. Noteworthy are two specific combinations: WordSVM, which decreased the size of WordPair by 18% while maintaining classification performance—evidenced by SVM-WordPair-Us in Figure 5; and WordPairUx, which lessened the size by 11% and had only a slight impact on DBERT performance when using tokens by SmartLinPair (DBERT-WordPair-SLPair-Ux). NB presented intriguing results with the LemSVM configuration, which diminished the size of LemPair by 12% and led to an improvement in F1-Macro Score by 3.7% (NB-LemPair-Us in Figure 6) for the pared-down dataset.

VI. Conclusion and Future Work

The implementation of transformer models trained on large datasets typically improves ML performance in a variety of tasks. This research has shown that tuning the model with specialised data generally improves text classification performance. However, in the context of the QASCRBio dataset, all optimisation strategies failed to improve the performance of DBERT beyond that of SVM. This study attempted to maximise the utility of the data by exploiting latent linguistic information, such as Part-of-Speech, which inevitably increased the computational requirements and improved the classification quality of DBERT.

Despite these efforts, LLM setups were consistently outperformed by statistical methods, such as SVM, in terms of quality and efficiency. The synthetic nature of the QASCR-Bio dataset, with carefully crafted texts and perfectly balanced classes, may inherently favour statistical approaches. In contrast, LLMs are adept at processing texts of varying orthographic quality and skewed information levels, such as tweets, possibly due to their training in different text types.

This research has opened up new avenues for exploration. While the selected LLM has proven its efficacy in analysing well-structured student responses, its application to naturally written responses presents a promising area for future research. The QASCRBio dataset, which primarily evaluates the biological and didactic elements within students' responses, may not fully capture the complexity of real student submissions. These submissions often contain spelling and grammatical errors, as well as extraneous information, such as emotional expressions, which could provide valuable insights for a comprehensive assessment.

This research highlights the need for advanced methods capable of interpreting such complexities within student responses. For example, the use of Part-of-Speech information has demonstrated potential benefits for automated assessment, suggesting that deeper linguistic analysis could provide significant benefits. Further studies should explore the refinement of LLM capabilities to handle better the nuanced and diverse nature of authentic student responses, thereby increasing the effectiveness and applicability of machine learning in educational assessment.

Acknowledgments

This research was funded by the Federal Ministry of Education and Research of Germany in the framework of FiSK (Project-Number 16DHB4005).

References

[1] H. Luan and C.-C. Tsai, "A review of using machine learning approaches for precision education," *Educational Technology & Society*, vol. 24, no. 1, pp. 250–266, 2021.

[2] W. Villegas-Ch, M. Román-Cañizares, and X. Palacios-Pacheco, "Improvement of an online education model with the integration of machine learning and data analysis in an LMS," *Applied Sciences*, vol. 10, no. 15, p. 5371, 2020.

[3] H. A. El-Sabagh, "Adaptive e-learning environment based on learning styles and its impact on development students' engagement," *International Journal of Educational Technology in Higher Education*, vol. 18, no. 1, p. 53, 2021.

[4] D. Schönle, C. Reich, and D. Ould-Abdeslam, "Linguistic-Aware WordPiece Tokenization: Semantic Enrichment and OOV Mitigation," in *6th International Conference on Natural Language Processing (ICNLP 2024)*, 2024, p. tba. Forthcoming.

[5] D. Schönle, C. Reich, and D. O. Abdeslam, "Linguistic driven feature selection for text classification as stop word replacement," *Journal of Advances in Information Technology*, vol. 14, no. 4, pp. 796–802, 2023.

[6] J. Makhoul, F. Kubala, R. Schwartz, R. Weischedel, *et al.*, "Performance measures for information extraction," in *Proceedings of DARPA broadcast news workshop*, Herndon, VA, 1999, pp. 249–252.

[7] R. W. Hamming, "Error detecting and error correcting codes," *The Bell system technical journal*, vol. 29, no. 2, pp. 147–160, 1950.

[8] D. Schönle, C. Reich, and D. Ould-Abdeslam, "Streamlining AI: Techniques for Efficient Machine Learning Model Selection," *The International Journal on Advances in Intelligent Systems*, vol. 17, no. 12, p. tba. 2024, forthcoming.
[9] D. Fiedler, J. Poser, and U. Harms, *Teacher questions and student answers for the SCRBio in the context of evolution*, 2024.
[10] S. M. Brookhart, *How to give effective feedback to your students*. Ascd, 2017.
[11] W. J. Popham, *Classroom assessment*. Allyn and Bacon Boston, 1995.
[12] P. Black and D. Wiliam, "Assessment and classroom learning," *Assessment in Education: principles, policy & practice*, vol. 5, no. 1, pp. 7–74, 1998.
[13] M. E. Dogan, T. Goru Dogan, and A. Bozkurt, "The use of artificial intelligence (AI) in online learning and distance education processes: A systematic review of empirical studies," *Applied Sciences*, vol. 13, no. 5, p. 3056, 2023.
[14] B. Das, M. Majumder, S. Phadikar, and A. A. Sekh, "Automatic question generation and answer assessment: a survey," *Research and Practice in Technology Enhanced Learning*, vol. 16, no. 1, p. 5, 2021.
[15] V. González-Calatayud, P. Prendes-Espinosa, and R. Roig-Vila, "Artificial intelligence for student assessment: A systematic review," *Applied Sciences*, vol. 11, no. 12, p. 5467, 2021.
[16] J.-C. Klie, M. Bugert, B. Boullosa, R. E. de Castilho, and I. Gurevych, "The inception platform: Machine-assisted and knowledge-oriented interactive annotation," in *Proceedings of the 27th international conference on computational linguistics: system demonstrations*, 2018, pp. 5–9.
[17] I. Kononenko, "Comparison of inductive and naive bayesian learning approaches to automatic knowledge acquisition," *Current trends in knowledge acquisition*, vol. 8, p. 190, 1990.
[18] T. K. Ho, "Random decision forests," in *Proceedings of 3rd international conference on document analysis and recognition*, IEEE, vol. 1, 1995, pp. 278–282.
[19] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.
[20] J. Hartmann, J. Huppertz, C. Schamp, and M. Heitmann, "Comparing automated text classification methods," *International Journal of Research in Marketing*, vol. 36, no. 1, pp. 20–38, 2019.
[21] D. Fiedler, D. Schönle, C. Reich, and U. Harms, "When practical situations are rare: Improving pre-service biology teachers' diagnostic competency in a classroom simulation with chatbot," *Herausforderung Zukunft*, p. 294, 2023.
[22] O. P. Adelana, M. A. Ayanwale, and I. T. Sanusi, "Exploring pre-service biology teachers' intention to teach genetics using an AI intelligent tutoring-based system," *Cogent Education*, vol. 11, no. 1, p. 2 310 976, 2024.
[23] J. Fischer, N. Machts, T. Bruckermann, J. Möller, and U. Harms, "The Simulated Classroom Biology—A simulated classroom environment for capturing the action-oriented professional knowledge of pre-service teachers about evolution," *Journal of Computer Assisted Learning*, vol. 38, no. 6, pp. 1765–1778, 2022.
[24] C. Rogers, H. El-Mounaryi, T. Wasfy, and J. Satterwhite, "Assessment of STEM e-learning in an immersive virtual reality (VR) environment," *Computers in Education Journal*, vol. 8, p. 15 724, Oct. 2017.
[25] DeepL SE, *How does DeepL work?* 2024. [Online]. Available: www.deepl.com/en/blog/how-does-deepl-work (visited on 05/29/2024).
[26] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
[27] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
[28] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.
[29] R. Collobert, K. Kavukcuoglu, and C. Farabet, "Torch7: A Matlab-like Environment for Machine Learning," in *BigLearn, NIPS Workshop*, 2011.
[30] W. McKinney, "Data Structures for Statistical Computing in Python," in *Proceedings of the 9th Python in Science Conference*, Stéfan van der Walt and Jarrod Millman, Eds., 2010, pp. 56–61.
[31] La Vivien, *Google Translate Architecture illustrated*, 2022. [Online]. Available: https://www.lavivienpost.com/google-translate-and-transformer-model/ (visited on 05/29/2024).
[32] K. Sparck Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of documentation*, vol. 28, no. 1, pp. 11–21, 1972.