A data-driven approach for the identification of features for automated feedback on academic essays

Mohsin Abbas, Peter van Rosmalen, and Marco Kalz

***Abstract*—For predicting and improving the quality of essays, text analytic metrics (surface, syntactic, morphological and semantic features) can be used to provide formative feedback to the students in higher education. In this study, the goal was to identify a sufficient number of features that exhibit a fair proxy of the scores given by the human raters via a data-driven approach. Using an existing corpus and a text analysis tool for the Dutch language, a large number of features were extracted. Artificial neural networks, Levenberg Marquardt algorithm and backward elimination were used to reduce the number of features automatically. Irrelevant features were eliminated based on the inter-rater agreement between predicted and human scores calculated using Cohen's Kappa (κ). The number of features in this study was reduced from 457 to 28 and grouped into different categories. The results reported in this paper are an improvement over a similar previous study. Firstly, the inter-rater reliability between the predicted scores and human raters was increased by tweaking the corpus for overfitting for average scores. The resulting maximum value of κ showed substantial agreement compared to moderate inter-rater reliability in the prior study. Secondly, instead of using a dedicated training and test set, the training and testing phases in the new experiments were performed using k-fold cross validation on the corpus of texts. The approach presented in this research paper is the first step towards our ultimate goal of providing meaningful formative feedback to the students for enhancing their writing skills and capabilities.**

***Index Terms*—Natural Language Processing, Artificial Neural Networks, Levenberg Marquardt, Backward Elimination, Dimensionality reduction, Feature selection, Feature reduction, k-fold Cross Validation**

I. Introduction

In academic environments, text-based assignments are an important form of assessment to judge learners' progress. On these assignments, individual feedback to learners is required for improvement of their performance and enhancement in learning. Such feedback, known as formative feedback, makes the learning environment more conducive for learners [1]. Formative assessment also supports self-regulated learning to enrich the learning experience [2].

Mohsin Abbas is with Chair of Open Education, Faculty of Management Science, Department of Strategic Management, Open University of the Netherlands, Heerlen, The Netherlands. He is also working as an Assistant Professor at Faculty of Information Technology, University of Central Punjab, Lahore, Pakistan. E-mail: mohsin.abbas@gmail.com, mohsin.a@ucp.edu.pk

Dr. P. van Rosmalen is an Associate Professor at Department of Educational Development and Research, School of Health Professions Education, Faculty of Health, Medicine and Life Sciences, Maastricht University, Maastricht, The Netherlands. E-mail: p.vanrosmalen@maastrichtuniversity.nl

Prof. Dr. Marco Kalz is Full Professor of Technology-Enhanced Learning at the Heidelberg University of Education, Im Neuenheimer Feld 561, 69120 Heidelberg, Germany. E-mail: kalz@ph-heidelberg.de

Manuscript received July 24, 2021; revised August 03, 2022; revised June 28, 2023; revised August 30, 2023. accepted September 20, 2023.

Instead of relying solely on teacher-provided feedback, learners can participate in activities to promote self-regulation [3]. Technology can assist here by providing feedback as input to these activities. By using technology, students do not need to wait for teachers to provide feedback, rather, they can add further details, start thinking in different directions or completely change the course of study based on the feedback given to them [4]. Both for on-campus and on-line students, teacher-provided feedback is a time-consuming task and it takes a considerable amount of a teacher's time to provide feedback on textual assignments [5]. Providing feedback to the learners immediately is not possible because reading and analyzing written material requires a lot of effort. For this purpose, several technological solutions have been researched and developed to support the process of providing feedback to learner's texts and textual assignments. As an alternative to teacher-provided feedback, another widely studied topic is peer feedback for improving quality of one's writing [5]–[10]. Further, due to the coronavirus pandemic, educational institutions have been forced to abandon on-campus education and are trying to continue the learning process through online education [11]–[13]. This has had an impact on both students and teachers [13]. For this reason, a system that provides automatic feedback to the students is essential to combat such situations in the future.

Concerning technology based approaches, PEG (Project Essay Grade) built in the 1960s by Ellis Page, is regarded as the very first system to use computers for assessment tasks [14]. The system automatically graded essays while the scores given by PEG were comparable with the scores given by human judges. The correlation scores between the human graders and PEG varied between 65% to 71%. The reduction in workload of the teachers is one of the motivations of our work, similar to PEG. MI-Write, the current version of PEG [15] provides automated essay scoring along with immediate feedback on texts through recommendations on how to improve the scores. IntelliMetric [16] another early Automatic Essay Scoring (AES) system used artificial intelligence to score essays. IntelliMetric calculated more than 300 discourse, semantic and syntactic features to give a final score based on coherence, organization, elaboration, sentence structure and overall mechanics of the essay [17]. Another approach that focused more on content was introduced with Latent Semantic Analysis [18], a technique that can be used to find the similarity among texts. A plethora of applications have been developed that apply latent semantic analysis for learning [19]–[21]. Similar work was done in Pearson's Intelligent Essay Assessor (IEA) [17]. Final scores in IEA were based on the correlation of unknown essays to a pool of already existing

Please cite as: Abbas, M., Van Rosmalen, P. & Kalz, M. (2023). A data-driven approach for the identification of features for automated feedback on academic essays. *IEEE Transactions on Learning Technologies*. https://doi.org/10.1109/TLT.2023.3320877

and scored essays. In English language, Educational Testing Services (ETS) has developed E-rater [22] to automatically score GMAT essays. In order to provide scores, E-rater uses a huge corpus of graded responses to train its system. The first version of E-rater used approximately 50 features with an agreement of 0.87 to 0.94 between the system and expert readers' scores on GMAT essay prompts [23]. In the newer version of E-rater (version 2.0), 12 more features were added with a kappa (κ) value of 0.58 [22]. Despite the existence of these systems, there is still a need to further elaborate these systems, in particular, on the kind and quality of the feedback they give to help students to improve their writing. In addition, to develop approaches to make them less dependent on huge corpora and to extend their reach to languages other than English.

For the development of these systems, one of the critical questions is, which textual features are most important for automated feedback and how these features can be identified. The textual features (surface, syntactic, morphological and semantic features) that contribute the most in predicting the quality of students' texts can be extracted using machine learning techniques to provide formative feedback to the students. Once a limited number of features has been calculated, a final selection can be made in cooperation with teachers and students to determine the most useful features to provide meaningful actionable feedback.

Several approaches for feature selection exist. In a study [24], an automatic linguistic and textual feature extraction tool Coh-Metrix [25] was used to select the features required to predict the essay quality; this selection was based on the highest values of Pearson correlation of features compared to scores given by human raters. Writing-Pal [26], an Intelligent Tutoring System, also uses features selected from Coh-Metrix using statistical procedures [27]. Features were selected in another study [28] using Principal Component Analysis and the effectiveness of chosen features was analyzed for providing formative feedback to the writers. Feature selection techniques in text mining using deep learning have been reviewed in [29].

Several existing text analysis tools can calculate a huge number of textual features against input texts. ReaderBench [30] is an open source framework that makes use of natural language processing techniques to provide text-analysis tools. The framework is multilingual [31] – text analysis tools are available in Dutch, French, Romanian and English. Readerbench provides more than 200 textual complexity indices related to linguistic features of the text including surface, syntactic, morphological, semantic, and discourse features. Using ReaderBench, research to choose features that contribute the most towards the scores given by human raters has already been conducted for the French language [32]. That research uses a different approach, namely Discriminant Function Analysis. T-scan [33], [34] is a Dutch language analysis tool that calculates more than 400 text features which can be used for lexical and syntactic analysis. T-Scan derives its features from tools such as Alpino parser [35] and Frog [36]. To date, no research is available that has identified in a data-driven manner the metrics for automated essay feedback for the Dutch language. Thus, the goal of our research is to find the text analytic metrics (surface, syntactic, morphological and semantic features) that can be used to predict the quality of student's essays in the Dutch language. Students may feel overwhelmed if all possible indicators are shown to them and in addition, only a few indicators could have sufficient predictive power to provide automated feedback on essays written by students. Therefore, the idea is to identify a small number of features that are required to provide meaningful feedback. Machine learning algorithms such as Neural Networks can be used to create models using a corpus of scored texts and subsequently backward elimination [37] can be used to choose those features which are the most meaningful ones.

The study on the types, learning outcomes, and implications of automated writing evaluation (AWE) feedback [38] found that such feedback can, to an extent, improve students' writing, albeit not as effectively as human feedback. This aligns seamlessly with our study's premise, which envisions our study as promoting self-regulated learning rather than a substitute for human guidance. Moreover, the research states that students generally found AWE feedback useful and motivating. The research on multi-dimensional analysis of writing flexibility [39] specifically focuses on the English language using Writing-Pal. Our study extends this utility to the Dutch language, highlighting its potential application across various languages. Further, the authors have indicated that lower-level feedback (i.e., spelling and grammar mistakes related feedback) has little to no impact on the properties of students' essays. Our study assumes that the essays have no spelling or grammatical mistakes in them. A review study on Automated Corrective Feedback (ACF) [39] emphasizes that ACF tools should only be used to assist, not replace, instructors and that learners need to understand the functions and limitations of these tools. Our research aligns with these recommendations. Research on the effects of automated feedback on students' scientific argumentation has shown that most students make revisions based on the feedback, leading to higher final scores [40], [41]. Each revision is associated with an average increase in scores, indicating the effectiveness of automated feedback in improving learning outcomes.

In one of our previous studies [42], we reported a set of text metrics that may be used to provide formative feedback. The analysis was done by calculating more than 457 features against a scored corpus of Dutch students' essays extracted using T-Scan. However, there were certain limitations reported in this experiment. The biggest constraint was the issue of overfitting of the learned models leading to limitations in reliability. The original corpus of scored texts contained 436 scores over representing texts with an average score. The machine learning model depicted the phenomena of overfitting [43] for prediction of scores using the models learned through this corpus. The models worked well for those texts which had an average score. However, the accuracy was reduced dramatically if those scores were predicted which were either well above or below average. The current study extends the approach of the earlier study by avoiding overfitting and thus improving the validity of our results. Secondly, instead of using all features we eliminated those features at the start of the experiment for which either the variance was negligible

or the values were not being calculated for texts. Lastly, to expand the training set, instead of using a dedicated training and a test set, the training and testing phases in the new experiments were performed using k-fold cross validation [44] on the corpus of texts.

Following the approach discussed above, the main research questions of this study were:

1) Which text analytic metrics (surface, syntactic, morphological and semantic features) can be derived to predict the quality of student essays in the Dutch language?
2) Is there any external evidence available that the identified dimensions contribute to essay quality and/or and can be used for formative feedback?

In addition, for question 1 following the improvements suggested, we investigate the difference in results with the results of our previous study [42].

This paper is divided into four parts: the methods, approach and the techniques used in the experiment performed are discussed in the next section. After that, the results of the current research are presented. Finally, we discuss the significance of our findings and discuss limitations of the research and conclude implications for future research that can be conducted using our approach.

An Overview of LLMs in Educational Evaluation

Recent years have seen the rise of Large Language Models (LLMs) as a pervasive technology in different fields [45]–[48]. LLMs like GPT-4 have exhibited significant capabilities in various tasks [49]–[51], including natural language understanding and generation, often reducing the need for manual feature engineering. Yet, the role of feature engineering remains critical, particularly in the domain of educational technology [52], [53].

The advantages of feature engineering lie in its ability to fine-tune models for specialized tasks. For example, when evaluating academic essays, feature engineering enables the inclusion of important metrics such as surface, syntactic, morphological, and semantic features [42]. LLMs often overlook these specifics unless specially customized. Thus, feature-engineered models offer a more tailored approach, whereas LLMs serve a general range of tasks.

Despite their wide range, LLMs can exhibit limitations in the educational sector. They often struggle with complex logic and reasoning tasks [54] . Moreover, they can produce biased outputs, a significant concern when accurate and impartial evaluation is needed. One major drawback is their lack of transparency in decision-making, making it difficult to understand the rationale behind the feedback provided [55], [56]. This could be particularly problematic for educational applications, where understanding the reasoning behind the feedback is crucial for student improvement. Their "black-box" nature often renders the decision-making process opaque, making it challenging to generate specific, actionable feedback for students.

The feature engineering approach starts by analyzing a comprehensive set of variables to identify which factors predict student performance and behavior most. The method often employs techniques like backward elimination, where a model is initially built with all available features, followed by iterative removal of the least significant features based on statistical tests. This iterative process allows the model to focus on a reduced set of highly informative features, enhancing both its interpretability and performance. Large Language Models, on the other hand, are not inherently designed to handle this sort of specialized feature selection [29], [57], often leading to a less targeted and potentially less effective analysis.

II. METHODS

The current study explores a data-driven approach to identify textual features and metrics for an essay feedback system for the Dutch language. We have implemented an Automatic Essay Scoring (AES) methodology to automatically predict the scores against the input texts. Using this methodology, the first research question that is addressed is to identify the text analytic metrics (surface, syntactic, morphological and semantic features) that can be used to predict the quality of student essays in the Dutch language. Prediction models can be created using Natural Language Processing approaches where the input of these models are the features extracted from the texts. The basis of our experimentation is the proposition that the scores obtained in the essays by students can be correlated with the features extracted from those texts.

A. Materials: Corpus preparation

The texts for our research have been taken from CLiPS Stylometry Investigation (CSI) corpus [58]. The CSI corpus was designed for stylometric research purposes such as detection of age, gender, authorship, personality, sentiment, deception, topic, and genre. The goal of the collection of this corpus was to provide a Dutch corpus that is freely available for research to overcome the problems related to the non-disclosure agreements and anonymization problems in the existing corpora. The CSI corpus is licensed under Creative-Commons Attribution Non-Commercial Share-Alike 3.0 (CC BY-NC-SA 3.0); we were therefore allowed to use it freely. The corpus contains a vast amount of meta-data available on the author. For each author, the corpus provides information on age, gender, region of origin, and personality scores on the Big Five Inventory (BFI) scale [59].

The version that we used in our research was assembled in February 2016. This corpus provides 436 essays scored by humans. The raters of these texts were professors teaching the Dutch proficiency course at the University of Antwerp. They were experts in the Dutch proficiency courses for native speakers. The authors of these texts were first-year, and second-year Linguistics & Literature students enrolled in the Dutch proficiency course at the University of Antwerp. These students were Dutch native speakers. There were 333 authors in total. The male authors were 57, while the remaining 276 authors were female. The corpus contains documents of two genres: essays/papers and reviews. In their first year, students had to write a shorter text, here called 'essay'; in their second year, they wrote a longer text called 'paper.' The topics of these assignments were not the same; therefore, the category/topic

was different for the students. The reviews were assignments for the students. Our research included documents related to essays and papers only.

The region-wise division of these authors is given below:

- Antwerp (Belgium): 228 (68.5%)
- The Netherlands: 32 (9.6%)
- Oost-Vlaanderen (Belgium): 30 (9%)
- Limburg (Belgium): 18 (5.4%)
- Vlaams-Brabant (Belgium): 13 (3.9%)
- West-Vlaanderen (Belgium): 6 (1.8%)
- Others: 6 (1.8%)

The human raters gave a single score to the essays ranging from 0 to 20. The maximum score obtained in this corpus is 18, whereas, the minimum score is 5. The original corpus contained a large number of texts with an average score. Using all the texts, the models didn't perform well and were overfitting [42]. The first step therefore was tweaking the corpus of scored texts during the training phase of the machine learning process. We reduced the size of our corpus to 406 by excluding texts which had an average score. The Fig.1 shows a comparison of texts in the original and the ones used in this study.

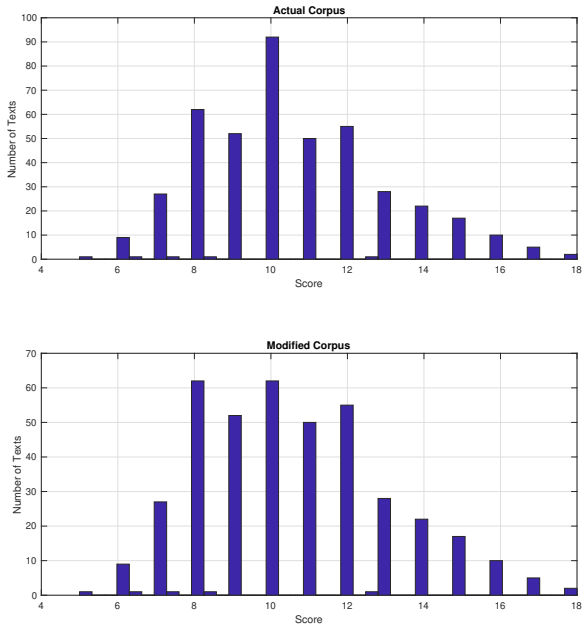

Fig. 1. Comparison between histograms of the actual corpus (top) the modified corpus (bottom). The histogram at the bottom has flatter peaks compared to histogram at the top.

The collectors of the CSI corpus have not disclosed the rating behaviour; however, we can confidently state that the rating of professors was not based merely on essay length. This is evident from the fact that the essay having 3292 words had a score of 10, while another essay having 327 words had a score of 12. Hence, the essay length did not influence the scoring of the raters. The maximum, minimum, and average length of the texts is given below:

- Maximum Length: 3292 words
- Minimum Length: 327 words
- Average Length: 1126 words

B. Tool: T-Scan for feature extraction

The tool that we used to extract features for our study was T-Scan. The complete form of T-Scan is **T**ext - **S**oftware for **C**omplexity **An**alysis. This feature extraction tool provides stylometric features; hence, using the CSI corpus (primarily for stylometric research) in conjunction with T-Scan was a suitable option. T-Scan intends to find features that influence the complexity of the Dutch texts using different tools for calculating its features; these are:

1) Frog [36] for tokenization, lemmatization, PoS tagging, and named entity recognition
2) Alpino [35] for dependency parsing;
3) SoNaR [60] for frequency lists;
4) SUBTLEX-NL [61] also for frequency lists;
5) Wopr [62] for measuring tri-gram probability, entropy, and perplexity;
6) Dutch Reference Document for semantically annotated word lists.

A brief description of the categorization of features calculated using T-Scan is shown in Table I. This classification of features is useful in finding out the effect of features present in each class on the overall score hence indicating the importance of individual classes.

TABLE I
CLASSIFICATION OF T-SCAN FEATURES SHOWING THE NUMBER OF FEATURES PRESENT IN DIFFERENT CATEGORIES.

No.	Class Name	No. of Features
1	General Characteristics	4
2	Word Difficulty	88
3	Sentence Complexity	73
4	Lexical Diversity and Referential Coherence	31
5	Relational Coherence and Situation Model Measures	36
6	Semantic Classes, Concreteness and Generality	128
7	Personal Elements	5
8	Other Lexical Information	76
9	Probability Measures	16
	Total	**457**

The features in the first category relate to **general characteristics** such as the number of paragraphs, sentences, and words in the text. **Word difficulty** features are based on the mechanics of the text in the essays, such as the word frequency and word preferences. Examples of two of such 88 features are the frequency of top 1000 words and the number of letters per sentence. The linguistic complexity of texts is measured in the 73 features the **sentence complexity**. **Lexical diversity and coherence** are essential elements of writing quality; there are 31 features related to this category in T-Scan. **Relational coherence** refers to the connectives of words such as: causal, comparative, constructive, enumerating, and temporal, and the **situation models** are words related to time, space, and

emotion. In this class, there are 36 features. **Semantic Classes, Concreteness, and Generality** features sum up to 128 and are related to nouns, adjectives, verbs, and adverbs. These features have been calculated based on the annotations given in the Dutch Reference Document [63]. The following two classes are related to Personal elements and Other Lexical Information. These classes have 5 and 75 features, respectively. The class **personal elements** include features referring to persons (for example, nouns referring to human beings, personal names, etc.) **Other lexical information** has features based on named entity recognition (calculated using Frog [36]), verb characteristics, modal verbs and auxiliary verbs of time, linking verbs, non-conjugated verbs, prepositions, and parts of speech (POS) related features. T-Scan calculates word probabilities in the category of **probability measures**; these 16 features are computed using Wopr [62].

C. Procedure: Feature selection

The features required as an input for our machine learning models are extracted using T-Scan. Eliminating those features at the start of the experiment for which either the variance was negligible or the values were not being calculated by T-Scan for a majority of texts reduced the number of features calculated by T-Scan from 457 to 382. Further reduction in features was performed by creating machine learning models for automatic prediction of the score of unknown texts written in Dutch language. The experiments in our research consist of the training and testing phases. The neural networks Levernberg Marquardt algorithm [64], [65] was used in both phases. The Levenberg–Marquardt (LM) Algorithm is used to solve nonlinear least squares problems. Instead of the LM algorithm, we could have chosen one of the two iterative algorithms: Gauss-Newton or Gradient-Descent. These curve-fitting methods minimize the sum of squared errors to find the optimal values of a function. Compared to the two, the LM algorithm is slower, but accurate. The LM algorithm is a combination of these two methods where a choice between either of the two methods is made based on the damping parameter (λ) [66]. Levenberg [64] and Marquardt [65] provided a solution to find the local minima of non-linear least squares problems [67].

The algorithm is explained considering the loss function (f) below as a sum of least squared errors (e) for m instances present in data-sets:

$$f = \sum_{i=1}^{m} e_i^2$$

We define a m x n Jacobian Matrix with n parameters of the Neural Network for $i = 1, ..., m$ and $j = 1, ..., n$:

$$\mathbf{J}_{i,j} = \frac{\partial e_i}{\partial \mathbf{w}_j}$$

A gradient vector of the loss function with e as vector of errors is computed as:

$$\bigtriangledown f = 2\mathbf{J}^T\mathbf{.}\ \mathbf{e}$$

Next, a Hessian Matrix can be approximated using the following expression:

$$\mathbf{H}f \approx 2\mathbf{J}^T\mathbf{.}\ \mathbf{J} + \lambda\mathbf{I}$$

Finally the expression for parameters improvement process for LM algorithm are defined for $i = 0, 1, ...$:

$$\mathbf{w}^{i+1} = \mathbf{w}^i - (\mathbf{J}^{iT}\mathbf{.}\ \mathbf{J}^i + \lambda^i\mathbf{I})^{-1}\mathbf{.}(2\mathbf{J}^{iT}\mathbf{.}\ \mathbf{e}^i)$$

After each iteration, the Levenberg–Marquardt Algorithm chooses either the gradient descent or Gauss-Newton and updates the solution based on the damping parameter (λ):

1) for large values of λ, Gradient-Descent method is used since parameters are far from their optimal value, and
2) for $\lambda = 0$ Gauss-Newton method is used as the parameters are close to their optimal value.

For the learning phase of these models, the input features are correlated with the scores given by human raters. Once the learned models have been created, automatic prediction of scores for unknown texts can be done in the testing phase. The performance of these models is measured by finding the inter-rater reliability between the predicted scores and human raters. Cohen's Kappa (κ) is one such statistic that is used in our research to measure the inter-rater reliability to filter out features which were relatively unimportant.

Backward elimination technique was used to reduce the number of features based on the inter-rater reliability between the predicted scores and the human scores. The corpus was divided into two parts for creating training and testing models. In this technique, texts for training were used and N input features were used to train the Artificial Neural Network models. The experiment was repeated N times, leaving one feature at a time. The feature left out was replaced (put back) for the next iteration, therefore, the algorithm used backward elimination technique with replacement. After each iteration, the value of Kappa was calculated. At the end of N iterations, N values of Kappa were calculated and that feature was eliminated for which the value of Kappa was maximum. The Kappa value being maximum was an indication that even without the feature eliminated, the inter-rater reliability was the best of all the Kappa values calculated, thus, the left-out features had little or no effect on predicting the final score. The features were then reduced by using Backward Elimination Technique which recursively downsized the number of features based on a stopping criterion. The algorithm of backward elimination is shown in the block diagram shown in Fig.2.

For increasing the validity of our results during the training and testing phases in the experiments using k-fold cross validation on the corpus of texts with a value of k = 7 with 58 texts in each bin. The Fig.3 shows the complete flow of our experiments with backward elimination and k-fold cross validation.

We regard the scores given by the human raters as evidence of essay quality; the greater the score obtained, the higher is the essay quality. The experiment did run until our stopping criteria (κ = 0.2) was met indicating a slight agreement between the human and our machine predicted scores serving as an evidence that the identified features contribute towards the essay quality. Consequently, we consider those features

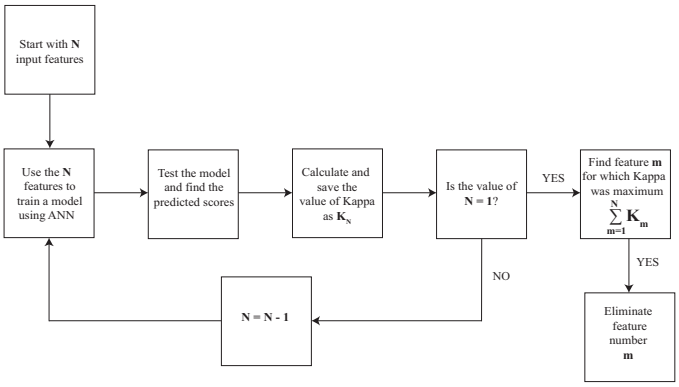

Fig. 2. Backward elimination Algorithm to identify a minimum set of text features.

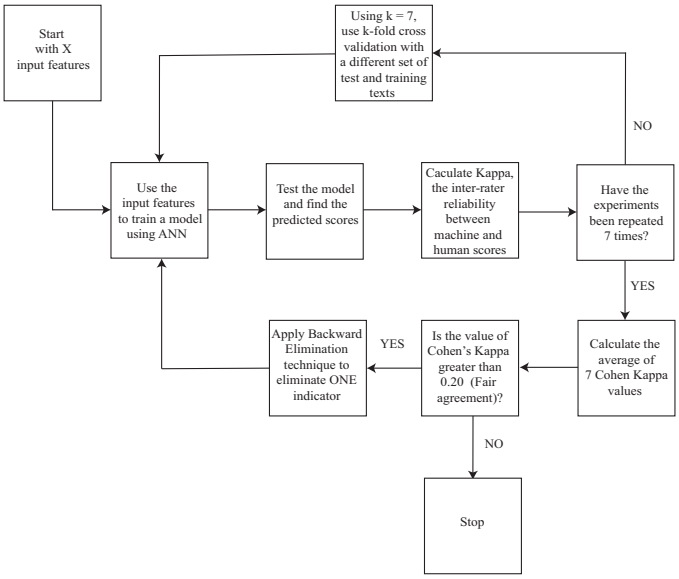

Fig. 3. Complete flow of the experiments with backward elimination and k-fold cross validation.

as the most important ones, which, when left behind at the end of the experiments, still show an agreement between the machine predicted and the given scores. We regard the number of features left at the end of the experiment as the most important ones since with only these few features, the inter-rater agreement did not fall below a range where there existed some agreement between the raters. Prior research had a goal to increase the value of Kappa to produce models for accurate prediction of scores for automatically grading of texts [15]–[17], [22], [23]. However, the goal of our research is to identify the minimum number of features for a particular inter-rater agreement and to investigate if these features can be used to provide automated formative feedback to students on their text. If the number of features is reduced, there should still be a correlation of the given scores with the ones predicted by our machine learning models.

For our second research question we focused on identifying external evidence to support that the features we identified contribute to essay quality and/or can be used for formative feedback.

III. Results

The number of features were reduced using the Levenberg Marquardt algorithm and backward elimination running on MATLAB R2017b on a computer having core i7 4.0 GHz processor and 32 GB RAM. The experiment ran for 19 days before the stopping criteria was reached. The maximum value of Cohen's Kappa was calculated as 0.7175 showing substantial agreement compared to 0.52 in the prior study which fell just above the middle range of moderate inter-rater reliability. Fig. 4 shows a comparison between the current results and the ones reported in the prior study [42].

At the end of the experiment, the number of features was reduced to 28 as compared to 23 in the earlier study. The Table II shows the features that were selected at the end of the experiment. These features belong to different categories/classes. A detailed explanation of these features is given in Appendix A. There are 17 features that are common in both sets (results of prior study and that of current research), these features are shown in the table given in Appendix B. 10 new features that were not present in the prior study have been listed in Appendix C. Six features that were present in the final list of prior study have now been eliminated.

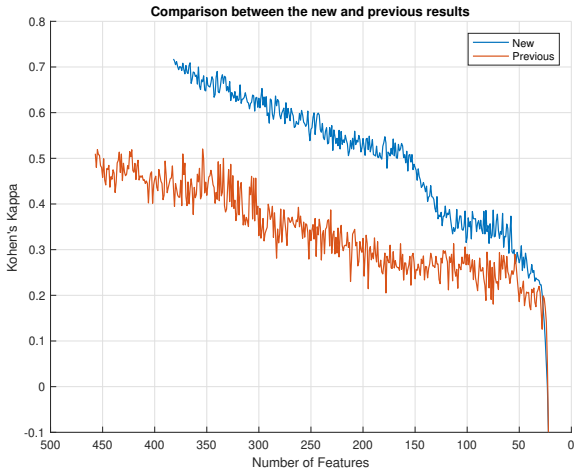

Fig. 4. The value of Cohen's Kappa (κ) at the end of each experiment by eliminating 1 feature at a time.

IV. Discussion

In the current study we have investigated a data-driven approach to reduce the number of features for the further usage of these features for automated feedback on Dutch student essays. Our first research question was, which text analytics metrics (surface, syntactic, morphological and semantic features) can be derived to predict the quality of student essays in the Dutch language?

Out of the 457 input features, 75 features with low variance were discarded before the start of the experiment reducing the total features to 382. The final results contain features for which there remained a fair agreement between machine predicted scores and human ratings at the end of our study. The number of features in this research was reduced to 28 by using

TABLE II
LIST OF SELECTED FEATURES

No.	Feature	Class[1]
1	*Wrd_freq_zn_log:* Logarithm of word frequency without names	WD
2	*Wrd_freq_log_sam_nw:* Logarithm of word frequency of the noun phrases	WD
3	*Freq2000:* The word that belongs to the most frequent 2000 words	WD
4	*Freq1000_inhwrd:* The proportion of content words associated with the most frequent 1000 words	WD
5	*Freq20000_nw:* Proportion of nouns associated with the most frequent 20000 words	WD
6	*Freq1000_corr:* Corrected proportion of words pertaining to the most frequent 1000 words	WD
7	*Lem_freq_log:* Lemma frequency, logarithm	WD
8	*Wrd_per_zin:* Words per sentences	SC
9	*Attr_bijv_nw_d:* Density of attributive adjectives	SC
10	*MTLD_wrd:* Measure of textual lexical diversity for words	LDRC
11	*Tijd_MTLD:* Measure of textual lexical diversity for time words	LDRC
12	*MTLD_inhwrd:* Measure of textual lexical diversity for content words	LDRC
13	*MTLD_lem:* Measure of textual lexical diversity for lemmas	LDRC
14	*TTR_wrd:* Type token ratio for words	LDRC
15	*Inhwrd_d:* Density of content words	LDRC
16	*Vnw_ref_d:* Density of referring pronouns	LDRC
17	*Tijd_d:* Density of time words	LDRC
18	*Causal_d:* Density of causal words	LDRC
19	*Conc_nw_ruim_d:* Density of broadly-concrete nouns	SCCG
20	*Gedekte_nw_p:* Proportion of nouns and names in the list	SCCG
21	*Alg_nw_d:* Density of general nouns	SCCG
22	*Ep_ev_bvnw_p:* Proportion of nouns that evaluate epistemically	SCCG
23	*Conc_ww_p:* Proportion of concrete verbs	SCCG
24	*Alg_bijw_d :* Density of general adverbs	SCCG
25	*Spec_bijw_p:* Proportion of specific adverbs to adverbs	SCCG
26	*Conc_nw_strikt_p:* Proportion of strictly concrete nouns	SCCG
27	*Concr_ov_nw_p:* Proportion of other specific nouns	SCCG
28	*Perplexiteit_bwd:* Perplexity, backwards	PM

[1] WD = Word Difficult; SC = Sentence Complexity; LDRC = Lexical Diversity and Referential Coherence; SCCG = Semantic Classes, Concreteness and Generality, PM = Probability Measures

a combination of machine learning, backward elimination techniques. We discuss the most important clusters of features and their relation to feedback (on writing) and potential to be used for automated feedback next.

1) Word Difficulty: Of the total features shown in Appendix A, there are seven features related to word difficulty. These features are based on the mechanics of the text in the essays such as the word frequency and word preferences. The understanding of the texts increases if the users are aware of the words used in those texts [68]. Word prevalence [69] and word frequency lists [70] are two ways to estimate the chances of a reader's knowledge about a word. Word frequency is strongly correlated with both perceived and actual text difficulty [71]. We therefore believe that all the frequency related features in the list of selected features may be important to provide feedback to the students. These features in this category are all related to the frequency of words in the texts, these are discussed in detail in Appendix A.

2) Sentence Complexity: Sentence Complexity features are important for measuring linguistic complexity [72], [73]. In our study, we have used "Sentence Complexity" as a broader term. While "Sentence Complexity" often refers broadly to the combination of sentence length and structural variety, "Syntactic Complexity" refers to the diversity of the grammatical structures within sentences. We have selected two features under this category to provide a comprehensive understanding of complexity in language use. The first feature, "Words per sentence," measures sentence length, often reflecting intricate ideas and sophisticated writing structures. The second feature,"Density of attributive adjectives," captures the interplay of syntactic and semantic complexity, indicating more intricate sentence structures and richer language use. These Sentence Complexity features are vital for measuring linguistic complexity and can be further explored to track the progression of complexity over time, such as in monitoring student growth and development. In our final list of features, there are two features from this category. If at a later stage it appears to be necessary (e.g., to measure complexity of expressions over time of a student) we might consider to trace back and look for other features in this category.

3) Lexical Diversity and Referential coherence: The variety of words used in the texts is called lexical diversity. The consistency of sentences in a text is called coherence. Lexical diversity and coherence are important elements of writing quality [74], [75]. The greater the vocabulary of words used, greater is the lexical diversity [76]. Lexical diversity is a measure of one's written and spoken language proficiency [77]–[80]. Traditionally, lexical diversity has been measured through Type Token Ratio (TTR) [81], however, other ways to measure it [82] have also been defined. In our experiment, 9 out of 28 features are related to the lexical diversity. Explanation of these features is given in Appendix A. We consider features in the class Lexical Diversity to be important for providing feedback to the learners.

4) Semantic Classes, Concreteness and Generality: Grouped into these classes are features related to nouns, adjectives, verbs and adverbs. These classes have been created based on the annotations given in the Dutch Reference Document [63]. Brief explanation of these features is given in Appendix A.

From our final list of eliminated features, the 9 features in this category seem too complex and require further investigation before it may be concluded if these can be used to give feedback to the students in order to improve the essay quality.

5) Probability Measures: Lastly, a feature calculates the logarithm of the back- ward perplexity. In Natural Language Processing, "perplexity" is a way to evaluate the language model [83] and has an inverse relation with the probability. A

lower value of perplexity refers to a higher value of probability. This feature is also quite technical and may not be useful to provide meaningful feedback to the students.

Of these features, we especially expect the categories "Word Difficulty" and "Lexical Diversity" as most useful for providing automated formative feedback to students. Providing feedback to students about the frequency of certain words (such as the proportion of content words associated with the most frequent 1000 words, nouns associated with the most frequent 20000 words or proportion of words pertaining to the most frequent 1000 words etc.) used in the texts and the diversity of the vocabulary (such as the measure of textual lexical diversity for words, time words, content words; the type token ratio for words etc.) may help them in improving the quality of their writing. The features present in the categories "Sentence Complexity" and "Semantic Classes" need to be explored further. The results obtained from these categories serve as a starting point for our future research where the teachers in writing will analyze if these features can be used to provide meaningful formative feedback. The only feature present in the category "Probability Measures" that calculates the logarithm of the backward perplexity is too technical and may not be helpful in providing meaningful feedback to the students.

We would like to highlight that each feature is not an isolated entity; instead, they interact and converge to form a coherent and meaningful text, which is fundamental to effective writing [84]. For instance, word difficulty, including the frequency of words [85], is correlated with word familiarity and influences readability [86]. Sentence complexity pertains to syntactic diversity [87] and sophistication [88]–[91], which are crucial for advanced writing. Lexical diversity [89], [92], [93] and coherence [94] are critical for ensuring a rich and varied text. The correct usage of semantic classes [72], [95]–[97] can reflect a writer's maturity and depth of understanding.

The second research question was to explore if there is any external evidence available that the identified dimensions contribute to essay quality and/or and can be used for formative feedback. Numerous studies have explored the use of computational indices and automated feedback in assessing and improving the quality of writing. These studies have investigated various linguistic and textual features, providing insights into their predictive capabilities and effectiveness [84], [85], [87], [89], [90], [92], [97].

The first version of e-rater used a huge corpus of graded responses to train its system. Their scoring engine computes the score of GRE essays with the help of approximately 50 features [98]. The final list contains lexical complexity features, average word length and use of sophisticated vocabulary to automatically predict the scores. For non-native English language speakers, e-rater distinguishes good essays from bad ones through the use of vocabulary by converting essays into vectors of word frequency [99]. Most of our features in the category "Word Difficulty" are word frequency features. In the newer version of e-rater (version 2.0) [22], 12 more features were added in the scoring engine. Out of these new features that were 3 features related to Lexical Diversity (type-token ratio, average word length and measure of vocabulary level) - our final list have similar features.

Several types of complexity features related to the surface, lexical, syntactic, and semantic properties of the texts were computed using the ReaderBench [31] framework. More studies utilizing the ReaderBench framework [100]–[102] have integrated a wide range of features for assessing writing quality, including surface indices, word complexity indices, and syntactic and cohesion indices. In terms of feedback analysis [103], automated content analysis of educational feedback suggested the use of linguistic features such as those developed in LIWC [104] and Coh-Metrix [25] as they better capture the content and quality of feedback. In another study, statistical techniques (discriminant analysis and stepwise regression) were used [57] to select Coh-Metrix features significant in predicting the quality of high and low scoring essays. The feature classes related to lexical diversity, word frequency and syntactic complexity were reported to be the most predictive ones in determining the essay proficiency. Features related to lexical diversity, word difficulty and sentence complexity were selected in another study [28] using Principal Component Analysis and the effectiveness of chosen features was analyzed for providing formative feedback to the writers. 211 features used in the study were extracted from 3 different tools: Coh-Metrix, Linguistic Inquiry and Word Count and the Writing Assessment Tool [27]. Using Coh-Metrix, a recent study [92] was employed to evaluate the linguistic and discourse features of texts with the aim of predicting the quality of assessments provided by judges. It was demonstrated that a combination of four key variables - word count, lexical diversity, hypernymy of verbs, and frequency of first person singular, significantly impacted the quality prediction model.

Investigations into adolescent academic writing [105] have highlighted the importance of specific lexical and discourse features in academic writing. Additionally, research on automated paraphrase quality assessment [106] has demonstrated the usefulness of assessing paraphrase quality using recurrent neural networks and language models, which can facilitate literacy skills and provide timely feedback to learners. Using different Machine Learning approaches, features that identify difficult texts have been discovered in [107]. A collection of 16 features have been suggested in this study from different categories. Final results of this study include features corresponding to frequency of words and characters, part-of-speech (nouns, adjectives, verbs, adverbs) and vocabulary.

These studies collectively demonstrate the potential of computational indices and automated feedback, suggesting that textual features not only show promise in predicting essay quality, but can also enhance the feedback process, thereby supporting learners in their writing development.

When comparing this study with its predecessor, our approach and findings might be of use also in other cases. In general, it tends to be challenging to collect dedicated, fit for purpose corpora for less-spread languages such as for instance Dutch. The approach followed did alleviate this challenge with two closely connected steps. The first step, a general improvement also applicable in large corpora, was tweaking the corpus for overfitting for average scores. Next, in particular of relevance for small corpora, the decrease in corpus size was

more than compensated for by expanding the training set by using k-fold cross validation on the corpus of texts. Obviously, in what cases and to what degree this approach is sufficient to use small corpora has to be part of further study.

This reduction in features was imperative because this huge number of features cannot be understood easily by the learners without prior knowledge and understanding of these features. The results in our study are restrained by the corpus used in the experiments - the corpus used in this work does not have texts that belong to the same subject or topic. There could be certain features that correspond to higher values for certain domains and types of writing (e.g. a news article versus an academic article) and lower values for others. Another problem in the corpus is regarding the scores. The human raters could give a single score to the essays ranging from 0 to 20, however, the maximum score obtained in this corpus is 18, whereas, the minimum score is 5. The corpus does not have essays that have scores less than 5 or greater than 18. Using a domain specific corpus with essays having scores spread across the scoring range may improve the results further. Lastly, the texts in the corpus used in our experiments have been written by people having different backgrounds, age groups and levels of education. The type of writing may have different features that distinguish the type of writer (such as their age, gender etc.). Conducting the experiment with texts written by people having the same age group, same level of education and similar background also needs to be investigated. In future, the same experiment can be repeated using machine learning algorithms other than neural networks to explore whether there is an improvement in results by using a different algorithm. Further, applying the algorithm on features extracted from texts using a different tool such as ReaderBench [31] may add to the existing set of our chosen features.

Pros and Cons of LLMs for Automated Feedback Generation

The paper recognizes the potential applicability of Large Language Models (LLMs) like ChatGPT and GPT-4 for automated feedback in education [108]. LLMs do bring certain advantages to the table. For instance, their extensive training on a wide array of topics equips them to provide quick feedback on a multitude of subjects without requiring training data [109]. This makes them particularly useful for scaling up operations where immediate responses are needed. They also excel at generating human-like, fluent text, which can make the feedback feel more personalized and engaging for students [110]. Furthermore, the ever-improving robustness and scalability of LLMs indicate a promising future [111] where these models could be adapted for more specialized tasks, potentially even matching the performance of feature-engineered models in certain scenarios. While these models can serve various functions, from casual conversation to content generation, they come with limitations that become apparent in an educational context [112].

First, the all-purpose nature of LLMs can be a disadvantage. They may lack the specialized, domain-specific training necessary for precise academic evaluation [113]. In contrast, our proposed model focuses on key features of academic writing, providing a more targeted evaluation. Models like ChatGPT, GPT-4, and similar ones have the potential to play a role in feedback generation [114], however, it's essential to recognize that they are tools requiring refinement and augmentation to effectively serve educational contexts [115]. These models, while proficient in generating language, might require additional layers of expertise, domain-specific knowledge, and insights from pedagogy to provide feedback tailored to individual learners' needs and growth.

Moreover, LLMs are sensitive to the types of prompts they receive. Adversarial or ambiguous prompts can lead to incorrect or biased outputs [116]. Ongoing efforts to mitigate these biases remain challenging due to the complexity of identifying and rectifying biased patterns [117]. Further, it's important to discuss the limitations intrinsic to models like ChatGPT. The model's reliance on supervised training can result in inconsistencies in responses [118], as the ideal answer depends on the model's knowledge rather than that of the human demonstrator. This challenge becomes apparent when the model encounters uncertain queries, where it might try to guess the user's intent rather than asking for clarification [119].

The method presented in this research utilizes a data-driven approach that includes artificial neural networks, the Levenberg Marquardt algorithm, and backward elimination for feature selection. The model, reduced from 457 to 28 features, achieves a substantial agreement in inter-rater reliability when compared to human evaluators. The proposed model's tailored nature, efficiency, and customization capabilities present it as a strong alternative to generalized LLMs in educational applications.

V. Conclusion

The current study has several methodological and practical implications. The approach presented in this research paper is the first step towards the development of automated feedback for essay writing for Dutch learners and higher education lecturers. The initial research [42] explored the data-driven reduction of features but had also some shortcomings with regard to overfitting and number of features. The current research has addressed this issue pointing to a potential approach which could be useful for other languages to identify important features for feedback on essays. In the current study, the dimensionality of the input features was reduced automatically via an existing corpus. In future studies, the usefulness of the identified features shall be confirmed with the help of human experts (teachers/experts and students). Based on their responses, the most relevant features for essay feedback will be chosen. An automated feedback service for Dutch essays will be built on the basis of the final list of features. Understandable feedback such as suggestions on how to improve will be shown to learners for improving the overall quality of responses. The final version of the automated essays feedback in Dutch will focus on providing feedback to the students using visualizations of their text. Such a combination of visualizations with written feedback is expected to help students in improving the quality of written responses. It will

also help in reducing the workload of teachers and tutors. Based on the feedback provided, the students will make changes in their essays. The implementations of the research will provide automatic analysis of Dutch essays using natural language processing and text mining techniques.

ACKNOWLEDGMENT

This paper is an expanded version of a paper titled "Identifying critical features for formative essay feedback with artificial neural networks and backward elimination," that was presented at the 14th European Conference on Technology Enhanced Learning, held in Delft, the Netherlands, in September 2019 and was published in the proceedings of EC-TEL 2019, Transforming Learning with Meaningful Technologies Lecture Notes in Computer Science, volume 11722. Springer, Cham. The authors would like to thank the Faculty of Information Technology, University of Central Punjab, Pakistan for providing technological resources to conduct this research.

REFERENCES

[1] S. Brown, "Assessment for Learn." *Learn. and Teaching in Higher Educ.*, no. 1, pp. 81–89, 2005.

[2] I. Clark, "Formative Assessment: Assessment Is for Self-regulated Learn." *Educational Psychol. Rev.*, vol. 24, no. 2, pp. 205–249, 2012.

[3] D. Boud and E. Molloy, "Rethinking models of feedback for Learn.: The challenge of design," *Assessment and Eval. in Higher Educ.*, vol. 38, no. 6, pp. 698–712, 2013.

[4] E. Ras, D. Whitelock, and M. Kalz, "The promise and potential of e-assessment for learn." *P. Reimann, S. Bull, M. Kickmeier-Rust, R. Vatrapu, & B. Wasson (Eds.), Measuring and visualizing Learn. in the Inf.-rich classroom*, pp. 21–40, 2015.

[5] J. Kasch, P. V. Rosmalen, M. Henderikx, and M. Kalz, "The factor struc. of the peer-feedback orientation scale (PFOS): toward a measure for assessing students' peer-feedback dispositions," *Assessment & Eval. in Higher Educ.*, pp. 1–14, 2021.

[6] T. M. Paulus, "The effect of peer and teacher feedback on student writing," *J. of 2nd Lang. Writing*, vol. 8, no. 3, pp. 265–289, 1999.

[7] M. M. Nelson and C. D. Schunn, "The nature of feedback: How different types of peer feedback affect writing Perform." *Instructional Sci.*, vol. 37, no. 4, pp. 375–401, 2009.

[8] H. Nguyen, W. Xiong, and D. Litman, "Iterative Design and Classroom Eval. of Automated Formative Feedback for Improving Peer Feedback Localization," *Int. J. of Artif. Intell. in Educ.*, vol. 27, no. 3, pp. 582–622, 2017.

[9] L. Ramachandran, E. F. Gehringer, and R. K. Yadav, "Automated Assessment of the Quality of Peer Reviews using Natural Lang. Process. Techn." *Int. J. of Artif. Intell. in Educ.*, vol. 27, no. 3, pp. 534–581, 2017.

[10] B. A. Simonsmeier, H. Peiffer, M. Flaig, and M. Schneider, "Peer Feedback Improves Students' Academic Self-Concept in Higher Educ." *Res. in Higher Educ.*, vol. 61, no. 6, pp. 706–724, 2020.

[11] L. Sun, Y. Tang, and W. Zuo, "Coronavirus pushes educ. online," *Nature Materials*, vol. 19, no. 6, pp. 687–687, 2020.

[12] E. M. Onyema, N. C. Eucheria, F. A. Obafemi, S. Sen, F. G. Atonye, A. Sharma, and A. O. Alsayed, "Impact of coronavirus pandemic on educ." *J. of Educ. and Pract.*, vol. 11, no. 13, pp. 108–121, 2020.

[13] P. Sahu, "Closure of universities due to coronavirus disease 2019 (covid-19): impact on educ. and mental health of students and academic staff," *Cureus*, vol. 12, no. 4, 2020.

[14] E. B. Page, "The imminence of... grading essays by Comput." *Phi Delta Kappa Int.*, vol. 47, no. 5, pp. 238–243, 1966.

[15] "MI-Write." [Online]. Available: https://measurementinc.com/miwrite

[16] L. M. Rudner, V. Garcia, and C. Welch, "An Eval. of the IntelliMetric essay scoring system," *The J. of Technol., Learn., and Assessment*, vol. 4, no. 4, pp. 1–22, 2006.

[17] M. Shermis and J. Burstein, *Automated Essay Scoring: A Cross-Disciplinary Perspective*, J. B. Mark D. Shermis, Ed., jan 2003.

[18] T. Landauer and P. Foltz, "An introduction to latent semantic Anal." *Discourse processes*, vol. 25, pp. 259–284, 1998.

[19] K. Zupanc and Z. Bosnić, "Automated essay eval. with semantic anal." *Knowl.-Based Syst.*, vol. 120, pp. 118–132, 2017.

[20] T. K. Landauer and J. Psotka, "Simulating text understanding for educational Appl. with latent semantic Anal.: Introduction to LSA," *Interactive Learn. Environments*, vol. 8, no. 2, pp. 73–86, 2000.

[21] L. Handayani, W. Alika, B. Negara *et al.*, "A latent semantic anal. method for autom. scoring system at essay test," in *J. of Phys.: Conf. Ser.*, vol. 1566, no. 1. IOP Publishing, 2020, p. 012119.

[22] Y. Attali and J. Burstein, "Automated Essay Scoring With E-Rater® V.2.0," *The J. of Technol., Learn. and Assessment*, vol. 4, no. 3, pp. 1–21, 2006.

[23] J. Burstein, K. Kukich, S. Wolff, C. Lu, and M. Chodorow, "Comput. Anal. of Essays," *Proc. of the NCME Symp. on Automated Scoring*, pp. 1–13, 1998.

[24] S. A. Crossley, R. Roscoe, and D. S. McNamara, "Predicting human scores of essay quality using Comput. indices of linguistic and textual features," in *Int. Conf. on Artif. Intell. in Educ. (AIED 2011)*, 2011, pp. 438–440.

[25] A. C. Graesser, D. S. McNamara, M. M. Louwerse, and Z. Cai, "Coh-Metrix: Anal. of text on cohesion and Lang." *Behav. Res. Methods, Instruments, and Comput.*, vol. 36, no. 2, pp. 193–202, 2004.

[26] R. D. Roscoe, L. K. Allen, J. L. Weston, S. A. Crossley, and D. S. McNamara, "The Writing Pal Intell. Tutoring System: Usability Testing and Develop." *Comput. and Composition*, vol. 34, pp. 39–59, 2014.

[27] D. S. McNamara, S. A. Crossley, and R. Roscoe, "Natural Lang. Process. in an Intell. writing strategy tutoring system," *Behav. Res. Methods*, vol. 45, no. 2, pp. 499–515, 2013.

[28] S. A. Crossley, K. Kyle, and D. S. Mcnamara, "To Aggregate or Not? Linguistic Features in Autom. Essay Scoring and Feedback Syst." *J. of Writing Assessment*, vol. 8, no. 1, pp. 1–16, 2015.

[29] H. Liang, X. Sun, Y. Sun, and Y. Gao, "Text feature extraction based on deep Learn.: a Rev." *Eurasip J. on Wireless Commun. and Networking*, vol. 2017, no. 1, pp. 1–12, 2017.

[30] M. Dascalu, W. Westera, S. Ruseti, S. Trausan-Matu, and H. Kurvers, "ReaderBench Learns Dutch: Building a Comprehensive Automated Essay Scoring System for Dutch Lang." in *AIED (Artif. Intell. in Educ.)*, vol. 10331. Cham: Springer, 2017, pp. 52–63.

[31] M. Dascalu, G. Gutu, S. Ruseti, I. C. Paraschiv, P. Dessus, D. S. Mcnamara, S. A. Crossley, and S. Trausan-matu, "ReaderBench: A Multi-lingual Framework for Analyzing Text Complexity," in *12th Eur. Conf. on Technol. Enhanced Learn. (EC-TEL 2017)*. Tallinn: Springer, 2017, pp. 495–499.

[32] M. Dascalu, P. Dessus, L. Thuez, and S. Trausan-matu, "How Well Do Student Nurses Write Case Studies? A Cohesion-Centered Textual Complexity Anal." in *12th Eur. Conf. on Technol. Enhanced Learn. (EC-TEL 2017)*. Tallinn: Springer, 2017, pp. 43–53.

[33] R. Kraf and H. Pander Maat, "Leesbaarheidsonderzoek: oude problemen, nieuwe kansen," *Tijdschrift voor taalbeheersing*, vol. 31, no. 2, pp. 97–123, 2014.

[34] H. P. Maat, R. Kraf, A. Van Den Bosch, N. Dekker, M. Van Gompel, S. Kleijn, T. Sanders, and K. Van Der Sloot, "T-Scan: A new tool for analyzing Dutch text," *Comput. Linguistics in the Netherlands J.*, vol. 4, pp. 53–74, 2014.

[35] G. Bouma, G. van Noord, R. Malouf, and G. V. Noord, "Alpino: Wide-coverage Comput. Anal. of Dutch." in *CLIN*, vol. 37, Jan 2000, pp. 45–59.

[36] A. Van Den Bosch, B. Busser, S. Canisius, and W. Daelemans, "An efficient memory-based morphosyntactic tagger and parser for Dutch," in *Proc. of the 17th Meeting of Comput. Linguistics in the Netherlands, CLIN17*, no. October, 2007, pp. 191–206.

[37] D. Koller and M. Sahami, "Toward optimal feature selection," *ICML'96 Proc. of the 13th Int. Conf. on Mach. Learn.*, pp. 284–292, 1996.

[38] Q.-K. Fu, D. Zou, H. Xie, and G. Cheng, "A review of awe feedback: types, learn. outcomes, and implications," *Computer Assisted Lang. Learn.*, pp. 1–43, 2022.

[39] L. K. Allen, A. D. Likens, and D. S. McNamara, "A multi-dimensional analysis of writing flexibility in an automated writing evaluation system," in *Proc. of the 8th International Conference on Learn. Analytics and Knowl.*, 2018, pp. 380–388.

[40] M. Zhu, H.-S. Lee, T. Wang, O. L. Liu, V. Belur, and A. Pallant, "Investigating the impact of automated feedback on students' scientific argumentation," *Int. J. of Science Education*, vol. 39, no. 12, pp. 1648–1668, 2017.

[41] M. Zhu, O. L. Liu, and H.-S. Lee, "The effect of automated feedback on revision behav. and learn. gains in formative assessment of scientific argument writing," *Computers & Education*, vol. 143, p. 103668, 2020.

[42] M. Abbas, P. Van Rosmalen, and M. Kalz, "Identifying Crit. Features for Formative Essay Feedback with Artif. Neural Networks and Backward Elimination," in *EC-TEL 2019: Transforming Learn. with Meaningful Technologies*, vol. 11722, no. September. Delft, The Netherlands: Springer, Cham, 2019, pp. 396–408.

[43] T. Dietterich, "Overfitting and Undercomputing in Mach. Learn." *ACM Comput. Surveys (CSUR)*, vol. 27, no. 3, pp. 326–327, 1995.

[44] M. Stone, "Cross-Validatory Choice and Assessment of Statistical Predictions," *J. of the Royal Statistical Soc.: Ser. B (Methodological)*, vol. 36, no. 2, pp. 111–133, 1974.

[45] Y. Wu, A. Q. Jiang, W. Li, M. Rabe, C. Staats, M. Jamnik, and C. Szegedy, "Autoformalization with large language models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 32 353–32 368, 2022.

[46] A. Yuan, A. Coenen, E. Reif, and D. Ippolito, "Wordcraft: story writing with large language models," in *27th s Conference on Intelligent User Interfaces*, 2022, pp. 841–852.

[47] C. W. Safranek, A. E. Sidamon-Eristoff, A. Gilson, and D. Chartash, "The role of large language models in medical education: Applications and implications," p. e50945, 2023.

[48] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, and D. S. W. Ting, "Large language models in medicine," *Nat. Med.*, pp. 1–11, 2023.

[49] Z. Fan, X. Gao, M. Mirchev, A. Roychoudhury, and S. H. Tan, "Automated repair of programs from large language models," in *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*. IEEE, 2023, pp. 1469–1481.

[50] A. Madani, B. Krause, E. R. Greene, S. Subramanian, B. P. Mohr, J. M. Holton, J. L. Olmos Jr, C. Xiong, Z. Z. Sun, R. Socher *et al.*, "Large language models generate functional protein sequences across diverse families," *Nat Biotechnol*, pp. 1–8, 2023.

[51] S. Sarsa, P. Denny, A. Hellas, and J. Leinonen, "Automatic generation of programming exercises and code explanations using large language models," in *Proceedings of the 2022 ACM Conference on International Computing Education Research-Volume 1*, 2022, pp. 27–43.

[52] M. Kuhn and K. Johnson, *Feature engineering and selection: A practical approach for predictive models*. Chapman and Hall/CRC, 2019.

[53] T. Rawat and V. Khemchandani, "Feature engineering (fe) tools and techniques for better classification performance," *Int. J. of Innovations in Engineering and Technol.*, vol. 8, no. 2, pp. 169–179, 2017.

[54] J. G. Meyer, R. J. Urbanowicz, P. C. Martin, K. O'Connor, R. Li, P.-C. Peng, T. J. Bright, N. Tatonetti, K. J. Won, G. Gonzalez-Hernandez *et al.*, "Chatgpt and large language models in academia: opportunities and challenges," *BioData Min.*, vol. 16, no. 1, p. 20, 2023.

[55] E. Kasneci, K. Seßler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. Hüllermeier *et al.*, "Chatgpt for good? on opportunities and challenges of large language models for education," *Learning and individual differences*, vol. 103, p. 102274, 2023.

[56] I. A. Murad, N. M. S. Surameery, and M. Y. Shakor, "Adopting chatgpt to enhance educational experiences," *Int. J. of Information Technol. & Computer Engineering (IJITC) ISSN: 2455-5290*, vol. 3, no. 05, pp. 20–25, 2023.

[57] D. S. McNamara, S. A. Crossley, and P. M. McCarthy, "Linguistic features of writing quality," *Written Communication*, vol. 27, no. 1, pp. 57–86, 2010.

[58] B. Verhoeven and W. Daelemans, "CLiPS Stylometry Investigation (CSI) corpus : A Dutch Corpus for the Detection of Age , Gender , Personality , Sentiment and Deception in text," in *The 9th Int. Conf. on Lang. Resour. and Eval. (LREC)*, 2014.

[59] O. P. John, L. P. Naumann, and C. J. Soto, "Paradigm shift to the integrative big five trait taxonomy: History, measurement, and conceptual issues." 2008.

[60] N. Oostdijk, M. Reynaert, V. Hoste, and I. Schuurman, "Sonar user documentation," *Online:¡ https://ticclops. uvt. nl/SoNaR_end-user_documentation_v*, vol. 1, no. 4, 2013.

[61] E. Keuleers, M. Brysbaert, B. New *et al.*, "Subtlex-nl: A new frequency measure for dutch words based on film subtitles," *Behav. Research Methods*, vol. 42, no. 3, pp. 643–650, 2010.

[62] A. Van Den Bosch and P. Berck, "Memory-based machine translation and lang. modeling." *Prague Bull. Math. Linguistics*, vol. 91, pp. 17–26, 2009.

[63] W. Martin and I. Maks, *Referentie Bestand Nederlands*, S. Bopp and M. Groot, Eds., 2005.

[64] K. Levenberg, "A method for the solution of certain non-linear problems in least squares," *Quart. of Appl. Math.*, vol. 2, pp. 164–168, 1944.

[65] D. W. Marquardt, "An algorithm for least-squares estimation of nonlinear parameters," *J. of the Soc. for Industrial and Appl. Math.*, vol. 11, no. 2, pp. 431–441, 1963.

[66] J. Fan and J. Pan, "A note on the Levenberg-Marquardt parameter," *Appl. Math. and Computation*, vol. 207, no. 2, pp. 351–359, 2009.

[67] J. J. More, "The Levenberg-Marquardt Algorithm: Implementation and Theory," *Watson G.A. (eds) Numerical Anal.. Lecture Notes in Math.*, vol. 630, pp. 762–764, 1978.

[68] N. Schmitt, X. Jiang, and W. Grabe, "The Percentage of Words Known in a Text and Reading Comprehension," *Modern Lang. J.*, vol. 95, no. 1, pp. 26–43, 2011.

[69] E. Keuleers, M. Stevens, P. Mandera, and M. Brysbaert, "Word Knowl. in the crowd: Measuring vocabulary size and word prevalence in a massive online experiment," *Quart. J. of Exp. Psychol.*, vol. 68, no. 8, pp. 1665–1692, 2015.

[70] A. Baron, P. Rayson, and D. Archer, "Word frequency and key word statistics in historical corpus linguistics," Tech. Rep., 2009.

[71] G. Leroy and D. Kauchak, "The effect of word familiarity on actual and perceived text difficulty," *J. of the American Medical Inform. Assoc.*, vol. 21, no. E2, pp. 169–172, 2014.

[72] P. C. Gordon, R. Hendrick, and M. Johnson, "Effects of noun phrase type on sentence complexity," *J. of Memory and Lang.*, vol. 51, no. 1, pp. 97–114, 2004.

[73] K. Zablotskaya, M. Abbas, S. Zablotskiy, S. Walter, and W. Minker, "Measuring verbal Intell. using linguistic Anal." in *Proc. - 2011 7th Int. Conf. on Intell. Environments, IE 2011*, 2011, pp. 88–91.

[74] G. A. McCulley, "Writing Quality, Coherence, and Cohesion," *Res. in the Teaching of English*, vol. 19, no. 3, pp. 269–282, 1985.

[75] W. R. Winterowd, "The Grammar of Coherence," *College English*, vol. 31, no. 8, p. 828, 1970.

[76] S. Jarvis, "Capturing the Diversity in Lexical Diversity," *Lang. Learn.*, vol. 63, no. 1, pp. 87–106, 2013.

[77] H. Daller, R. Van Hout, and J. Treffers-D Aller, "Lexical Richness in the Spontaneous Speech of Bilinguals," *Appl. Linguistics*, vol. 24, no. 2, pp. 197–222+267, 2003.

[78] B. Laufer and P. Nation, "Vocabulary size and use: Lexical richness in L2 written production," *Appl. Linguistics*, vol. 16, no. 3, pp. 307–322, 1995.

[79] K. O'loughlin, "Lexical density in candidate output on direct and semi-direct versions of an oral proficiency test," *Lang. Testing*, vol. 12, no. 2, pp. 217–237, 1995.

[80] G. Yu, "Lexical diversity in writing and speaking task Performances," *Appl. Linguistics*, vol. 31, no. 2, pp. 236–259, 2010.

[81] B. Richards, "Type/token ratios: What do they really tell us?" *J. of Child Lang.*, vol. 14, no. 2, pp. 201–209, 1987.

[82] D. Malvern and B. Richards, "Investigating accommodation in Lang. proficiency interviews using a new measure of lexical diversity," *Lang. Testing*, vol. 19, no. 1, pp. 85–104, 2002.

[83] S. F. Chen, D. Beeferman, and R. Rosenfeld, "Eval. metrics for Lang. models," *Proc. of the DARPA Broadcast News Transcription and Understanding Workshop*, 1998.

[84] S. A. Crossley, "Linguistic features in writing quality and development: An overview," *J. of Writing Research*, vol. 11, no. 3, p. 415–443, Feb. 2020. [Online]. Available: https://www.jowr.org/index.php/jowr/article/view/582

[85] A. Riemenschneider, Z. Weiss, P. Schröter, and D. Meurers, "Linguistic complexity in teachers' assessment of german essays in high stakes testing," *Assessing Writing*, vol. 50, p. 100561, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1075293521000507

[86] G. Leroy and D. Kauchak, "The effect of word familiarity on actual and perceived text difficulty," *J. of the American Medical Inform. Assoc.*, vol. 21, no. e1, pp. e169–e172, 10 2013. [Online]. Available: https://doi.org/10.1136/amiajnl-2013-002172

[87] C. Lee, H. Ge, and E. Chung, "What linguistic features distinguish and predict l2 writing quality? a study of examination scripts written by adolescent chin. learners of english in hong kong," *System*, vol. 97, p. 102461, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0346251X21000154

[88] S. F. Beers and W. E. Nagy, "Syntactic complexity as a predictor of adolescent writing quality: Which measures? which genre?" *Reading and Writing*, vol. 22, pp. 185–200, 2 2009.

[89] D. J. Arya, E. H. Hiebert, and P. D. Pearson, "The effects of syntactic and lexical complexity on the comprehension of elementary science texts," *Int. Electronic J. of Elementary Education*, vol. 4, no. 1, pp. 107–125, 2011.

[90] C. W. Jo, "Mapping adolescent literacy across l1 backgrounds: Linguistic and discourse features as predictors of persuasive essay quality," *System*, vol. 104, p. 102698, 2022.

[91] U. Maamuujav, C. B. Olson, and H. Chung, "Syntactic and lexical features of adolescent l2 students' academic writing," *J. of Second Lang. Writing*, vol. 53, p. 100822, 2021.

[92] L. Ouyang, Q. Lv, and J. Liang, "Coh-metrix model-based automatic assessment of interpreting quality," *Testing and assessment of interpreting: Recent developments in China*, pp. 179–200, 2021.

[93] P. M. McCarthy and S. Jarvis, "MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment," *Behav. Res. Methods*, vol. 42, no. 2, pp. 381–392, 2010.

[94] S. Crossley and D. McNamara, "Cohesion, coherence, and expert evaluations of writing proficiency," in *Proc. of the Annu. Meeting of the Cognitive Science Society*, vol. 32, no. 32, 2010.

[95] M.-D. Dascălu, "Assessing writing and student performance using natural lang. processing and a dialogical framing," Ph.D. dissertation, Ludwig Maximilian University of Munich, Germany.

[96] S. Somasundaran, M. Flor, M. Chodorow, H. Molloy, B. Gyawali, and L. McCulla, "Towards evaluating narrative quality in student writing," *Transactions of the Assoc. for Computational Linguistics*, vol. 6, pp. 91–106, 2018.

[97] T.-T. Goh, H. Sun, and B. Yang, "Microfeatures influencing writing quality: The case of chin. students' sat essays," *Computer Assisted Lang. Learn.*, vol. 33, no. 4, pp. 455–481, 2020.

[98] F. J. Breyer, Y. Attali, D. M. Williamson, L. Ridolfi-McCulla, C. Ramineni, M. Duchnowski, and A. Harris, " A Study of the Use of the e-rater ® Scoring Engine for the Analytical Writing Measure of the GRE ® revised General Test ," *ETS Res. Rep. Ser.*, vol. 2014, no. 2, pp. 1–66, 2014.

[99] J. Burstein and M. Chodorow, "Automated essay scoring for nonnative English speakers," p. 68, 1999.

[100] G. Gutu, M. Dascalu, S. Trausan-Matu, and P. Dessus, "Readerbench goes online: a comprehension-centered framework for educational purposes," in *13th Interanational Conference on Human-Computer Interaction (RoCHI 2016)*. MATRIX ROM, 2016, pp. 95–102.

[101] I. Toma, A.-M. Marica, M. Dascalu, and S. Trausan-Matu, "Readerbench–automated feedback generation for essays in romanian," *University Politehnica of Bucharest Scientific Bulletin Series C-Electrical Engineering and Computer Science*, pp. 21–34, 2021.

[102] R.-M. Botarleanu, M. Dascalu, M.-D. Sirbu, S. A. Crossley, and S. Trausan-Matu, "Readme–generating personalized feedback for essay writing using the readerbench framework," in *The Interplay of Data, Technology, Place and People for Smart Learn.: Proc. of the 3rd International Conference on Smart Learn. Ecosystems and Regional Development 3*. Springer, 2019, pp. 133–145.

[103] I. Osakwe, G. Chen, A. Whitelock-Wainwright, D. Gašević, A. P. Cavalcanti, and R. F. Mello, "Towards automated content analysis of educational feedback: A multi-language study," *Computers and Education: Artif. Intelligence*, vol. 3, p. 100059, 2022.

[104] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: Liwc and computerized text analysis methods," *J. of language and social psychology*, vol. 29, no. 1, pp. 24–54, 2010.

[105] C. W. Jo, "Short vs. extended adolescent academic writing: A cross-genre analysis of writing skills in written definitions and persuasive essays," *J. of English for Academic Purposes*, vol. 53, p. 101014, 2021.

[106] B. Nicula, M. Dascalu, N. Newton, E. Orcutt, and D. S. McNamara, "Automated paraphrase quality assessment using recurrent neural networks and language models," in *Intell. Tutoring Syst.: 17th International Conference, ITS 2021, Virtual Event, June 7–11, 2021, Proc. 17*. Springer, 2021, pp. 333–340.

[107] D. Kauchak, O. Mouradi, C. Pentoney, and G. Leroy, "Text simplification tools: Using Mach. Learn. to discover features that identify difficult text," *Proc. of the Annu. Hawaii Int. Conf. on System Sciences*, pp. 2616–2625, 2014.

[108] L. Kohnke, B. L. Moorhouse, and D. Zou, "Chatgpt for language teaching and learning," *RELC J.*, p. 00336882231162868, 2023.

[109] Y. Su, Y. Lin, and C. Lai, "Collaborating with chatgpt in argumentative writing classrooms," *Assessing Writing*, vol. 57, p. 100752, 2023.

[110] D. Kalla and N. Smith, "Study and analysis of chat gpt and its impact on different fields of study," *Int. J. of Innovative Science and Research Technol.*, vol. 8, no. 3, 2023.

[111] A. Haleem, M. Javaid, and R. P. Singh, "An era of chatgpt as a significant futuristic support tool: A study on features, abilities, and challenges," *BenchCouncil transactions on benchmarks, standards and evaluations*, vol. 2, no. 4, p. 100089, 2022.

[112] M. Menekse, "Envisioning the future of learning and teaching engineering in the artificial intelligence era: Opportunities and challenges," *J. of Engineering Educ.*

[113] J. Kocoń, I. Cichecki, O. Kaszyca, M. Kochanek, D. Szydło, J. Baran, J. Bielaniewicz, M. Gruza, A. Janz, K. Kanclerz *et al.*, "Chatgpt: Jack of all trades, master of none," *Information Fusion*, p. 101861, 2023.

[114] J. Huang and M. Tan, "The role of chatgpt in scientific communication: writing better scientific review articles," *American J. of Cancer Research*, vol. 13, no. 4, p. 1148, 2023.

[115] J. M. R. Rodríguez, M. S. R. Montoya, M. B. Fernández, and F. L. Lara, "Use of chatgpt at university as a tool for complex thinking: Students' perceived usefulness," *NAER: J. of New Approaches in Educ. Research*, vol. 12, no. 2, pp. 323–339, 2023.

[116] R. W. McGee, "Is chat gpt biased against conservatives? an empirical study," *SSRN Electronic J.*, 2023.

[117] G. M. Currie, "Academic integrity and artificial intelligence: is chatgpt hype, hero or heresy?" in *Seminars in Nuclear Medicine*. Elsevier, 2023.

[118] M. Farrokhnia, S. K. Banihashem, O. Noroozi, and A. Wals, "A swot analysis of chatgpt: Implications for educational practice and research," *Innovations in Educ. and Teaching Int.*, pp. 1–15, 2023.

[119] C. K. Lo, "What is the impact of chatgpt on education? a rapid review of the literature," *Educ. Sciences*, vol. 13, no. 4, p. 410, 2023.

Mohsin Abbas is as an Assistant Professor at Faculty of Information Technology, University of Central Punjab, Pakistan. He is also perusing his PhD from Open University in the Netherlands. His research areas include Natural Language Processing and Big Data Analysis. He did his Masters in Communications Technology from University of Ulm, Germany. He has more than 16 years of teaching experience in leading universities of Pakistan.

Peter Van Rosmalen is currently working as an Associate Professor, Department of Educational Development and Research, School of Health Professions Education, Faculty of Health, Medicine and Life Sciences, Maastricht University. Peter is chair of the taskforce Instructional Design and E-learning and does research in Educational Technology.

Marco Kalz is full Professor of technology-enhanced learning at the Heidelberg University of Education. He is also affiliated to Chair of Open Education, Faculty of Management Science, Department of Strategic Management, Open University of the Netherlands and Chair of Digital Education, Faculty of Cultural Science and Humanities, Department of Media Education, Heidelberg University of Education,. His research interest lies on the use of open education, pervasive technologies and formative assessment to support learning and knowledge construction. He has published more than 100 peer-reviewed publications.