# Automatic Thesis Grading: Predicting the Grades of Bachelor's and Master's Theses

Jessica Bormann

Department of Psychology, University of Amsterdam

Supervisor: Dr. R.P.P.P. (Raoul) Grasman

23.05.2022

## Abstract

Teachers evaluate essays manually, which is time-consuming and results in only a handful of opportunities for students to practice their writing skills. Over the past 50 years, researchers developed several automatic essay grading (AEG) systems for short essay texts. This study investigated the success of AGE+, an existing AEG system for short essay texts, in predicting the grades of bachelor's and master's theses and identified essential features for domain-specific texts. AGE+ encompasses a wide variety of features of the categories: linguistic (lexical sophistication, grammar, mechanics), content and coherence. Seven regression models were fit on essay texts from the Hewlett datasets and bachelor's and master's theses. Results displayed low quadratic weighted kappa (QWK) scores, low to moderate correlations for thesis texts and high QWK scores and correlations for essay texts. For thesis texts, features of the categories content, mechanics, grammar, and coherence were more decisive than lexical sophistication and most part-of-speech tag features. In conclusion, AGE+ is not as successful in predicting the grades of domain-specific texts as for short essay texts. More features developed explicitly for bachelor and master theses should be tested to establish a successful AEG system for domain-specific texts.


*Keywords*: automated essay grading, automated thesis grading, automated essay scoring, natural language processing, domain-specific text grading, scientific text grading

**Contents**

## Introduction

Teachers assess creative thinking and students' writing skills with essays, which are short argumentative texts about a specific topic. The evaluation of essays is a time-intensive process, resulting in only a handful of opportunities for the students to receive feedback and practice their writing skills (Ramesh & Sanampudi, 2021; Zupanc & Bosnić, 2017). A teacher grades an essay based on rubrics, which are criteria set by the teacher to make the grading process more reliable and valid. The rubrics also serve as a guideline for students when writing the essay. Despite the assumption that rubrics increase both inter-rater and intra-rater reliability and validity of the grading process, without specific training on the construction and proper usage of rubrics, both reliability and validity of human grading are questionable (Rezaei & Lovorn, 2010). Therefore, a different form of essay evaluation that is more reliable and valid than human grading is needed.

### History of Automatic Essay Grading Systems

In 1966 Ellis Page attempted to improve the reliability and validity of essay grading by presenting the first computer-based essay grading system, called the *Project Essay Grader* (PEG). His ground-breaking proposal of letting computers grade essays was met with rejection at first, as many doubted the ability of a computer to deal with qualitative data (Page, 1968). However, Page viewed this as a psychometric problem. He transformed qualitative text data into statistical features that described an essay's qualities and named them "trins" for intrinsic variables. In the training stage of PEG, the correlations of the "trins", for instance, word count, grammar, and punctuation, are determined. Page referred to the correlations as "proxes" and utilised them as the regression equation coefficients. In the final stage, the scoring stage, Page identified the "proxes" for each essay as input for the prediction equation and obtained the final score (Page, 1966).

Even though PEG successfully predicted human essay grades, Chung and O'Neil (1997) criticised the system for lacking features describing the content of essays. At the end of the century, Burstein, Kukich, Wolff, Lu, and Chodorow (1998) published *E-rater* and Foltz, Laham, and Landauer (1999) published *Intelligent Essay Assessor* (IEA), which both considered the content of essays for predicting a grade. Table 1 displays the first automatic essay grading (AEG) systems. E-rater included both style and content features, which utilised prompt-specific vocabulary, and it reached a higher correlation between human and predicted grades than PEG and IEA (Hussein, Hassan, & Nassef, 2019). The IEA system considered only the content of the essay and its similarity with the source text by analysing text with *Latent semantic analysis* (LSA). This machine learning technique enabled IEA to provide feedback and detect plagiarism (Foltz et al., 1999).

**Table 1**

*Comparison of the first AEG systems*

| AEG system | Features | Model | Correlation |
|---|---|---|---|
| PEG (Page, 1966) | Style | Statistical | 0.87 |
| E-rater (Burstein et.al, 1998) | Style & content | NLP | ~0.91 |
| IEA (Foltz et al., 1999) | Content | LSA | 0.90 |

*Note.* From Hussein et al. (2019)

Latent semantic analysis works based on word vectors derived with PCA. Word vectors are numerical representations of text, which, based on the *distributional hypothesis,* provide information about the semantics of text (Sahlgren, 2008). Taghipour and Ng (2016) and Dong and Zhang (2016) were the first to utilise a neural network model in combination

with word vectors and reached higher Quadratic weighted kappa (QWK) scores than models with style-based features. The QWK is an evaluation metric that quantifies the agreement between the computer-predicted and the human rater's grade by considering the ordinal nature of grades (Ramesh & Sanampudi, 2021). The potential of neural network approaches quickly became evident as they produced higher QWK scores than regression models (Hussein et al., 2019). For instance, the LSTM model by Zhao, Zhang, Xiong, Botelho and Heffernan (2017) trained with statistical features reached a QWK of 0.78, while a regression model trained on both statistical and style-based features reached a QWK of 0.69 (Cummins, Zhang, & Briscoe, 2016). Table 2 gives an overview of various AEG systems trained and tested on the Hewlett datasets (The Hewlett Foundation: Automated Essay Scoring, 2012).

**Table 2**

*Comparison of AEG systems trained on the Hewlett datasets*

| AEG system | Features | Model | QWK |
|---|---|---|---|
| Cummins et al. in (2016) | Statistical and style-based | Regression | 0.690 |
| Taghipour and Ng (2016) | Word vectors (one-hot representation) | CNN + LSTM | 0.761 |
| Dong and Zhang (2016) | Word vectors (bag-of-words) | CNN | 0.734 |
| Zhao et al. (2017) | Statistical | LSTM (memory network) | 0.780 |
| AGE+ (Zupanc & Bosnić, 2017) | Linguistic, content and coherence | Random Forest | 0.818 |
| SAGE+ (Zupanc & Bosnić, 2017) | Linguistic, content, coherence, and consistency | Random Forest | 0.828[1] |
| Wang, et al., (2018) | Word embeddings | BLSTM | 0.83 |

*Note.* 1 Dataset with source-based essays. Quadratic weighted Kappa (QWK) is the agreement measure between two graders.

Despite the improvements in AEG systems, Zupanc and Bosnić (2017) criticised the lack of focus on the semantics of essays. They developed AGE+, which included style and coherence features and SAGE+ with additional consistency features. Both systems reached high QWK scores (0.818 and 0.828) and outperformed neural network models employing vector embedding features at the time. However, subsequent research focused on context-aware vector embedding features, including prominent pre-trained vector embeddings such as BERT embeddings (Devlin, Chang, Lee, & Toutanova, 2018). Neural network models which employ meaningful word vectors reach similar QWK scores as AGE+ and SAGE+. For instance, Wang, Liu and Dong (2018) fit a neural network model which adopted context-aware word embeddings and archived state-of-the-art performance. The results of both types of systems highlight the importance of considering semantics for essay grading.

Looking back from PEG until now, overall, the evaluation metrics displayed high correlations and QWK scores, which indicates that AEG systems are performing well in predicting essay grades. Most systems perform better than human graders, who reached a QWK of 0.754 averaged over all Hewlett datasets (Dong & Zhang, 2016). In contrast to human raters, AEG systems have perfect test-retest reliability because they are computer-based systems. However, according to Powers, Burstein, Chodorow, Fowles and Kukich (2002), the validity of most systems has not been fully established yet because the models are evaluated based on the agreement between the human rater and the predicted grade. The problem with this approach is that the validity of human grades is assumed by accepting the grade given by human raters as ground truth.

Powers et al. (2002) evaluated the validity of both 1) the predicted grades of an AEG system and 2) human grades by examining the relationship with independent indicators of the student's writing skills, such as grades of previous essays written by the student. They reasoned that the more similar the relationship of the predicted and the human grade was to

the independent indicators, the more truly each method reflected the student's writing skill. Even though Powers et al. (2002) reported a weaker relationship for predicted scores than for human scores, they concluded that there is some evidence for the validity of the AEG system. To further improve AEG systems and facilitate acceptance, Powers et al. (2002) stressed testing the system's limits, for instance, by determining for which kinds of texts the system produces good results and for which texts it produces poor results.

The types of texts researchers train AEG systems on are somewhat limited, especially AEG systems for more complex, domain-specific texts are scarce. Most AEG systems utilise short essay texts, such as the Hewlett datasets, which 90% of the systems adopt (Ramesh & Sanampudi, 2021; The Hewlett Foundation: Automated Essay Scoring, 2012). The essays contained in the Hewlett datasets were written by high school students and do not include domain-specific knowledge. Ramesh and Sanampudi (2021) stressed the need to test AEG systems with domain-specific texts. In contrast to short essays, student writing assignments such as bachelor's and master's theses in higher levels of education are highly domain-specific by aiming to answer a novel research question. Mosallam, Toma, Adhana, Chiru and Rebedea (2014) developed an AEG system employing linguistic features trained on bachelor and master theses. However, they achieved low accuracy with this set of features and suggested that semantic features would improve the accuracy of thesis grading by capturing the complexity of thesis texts. Therefore, this study aims to apply an existing AEG system that adopts semantic features to domain-specific text.

One successful AEG system which includes a wide variety of statistical and semantic features and performs well on the Hewlett datasets is AGE+, proposed by Zupanc and Bosnić (2017). The system comprises 101 features, of which 67 are linguistic, five content and 29 coherence features. Tables A1, A2 and A3 give an overview of the AGE+ linguistic, content and coherence features. The linguistic features capture basic statistical data about the text,

such as the word count or sentence count, grammar errors, and part-of-speech tags. The content features measure the cosine similarity of a thesis and all other theses by considering the grade given by the human rater. The coherence features capture the semantics of the text based on the assumption that the content in a coherent essay changes throughout the text (Zupanc & Bosnić, 2017).

The present study investigates whether AGE+ is also successful in predicting a grade for bachelor's and master's theses texts. It will be explored whether the same features important for short essay texts are also meaningful for more complex domain-specific texts or whether different features stand out to be more important. It is expected that AGE+ features work similarly well on more complex theses texts as on shorter essay texts. However, different features are expected to be more important for the complex texts than for the short essay texts. An AEG system for scientific texts such as bachelor and master theses would contribute to the validation of AEG systems and potentially provide students with more opportunities to practice and improve scientific writing skills.

## Methods

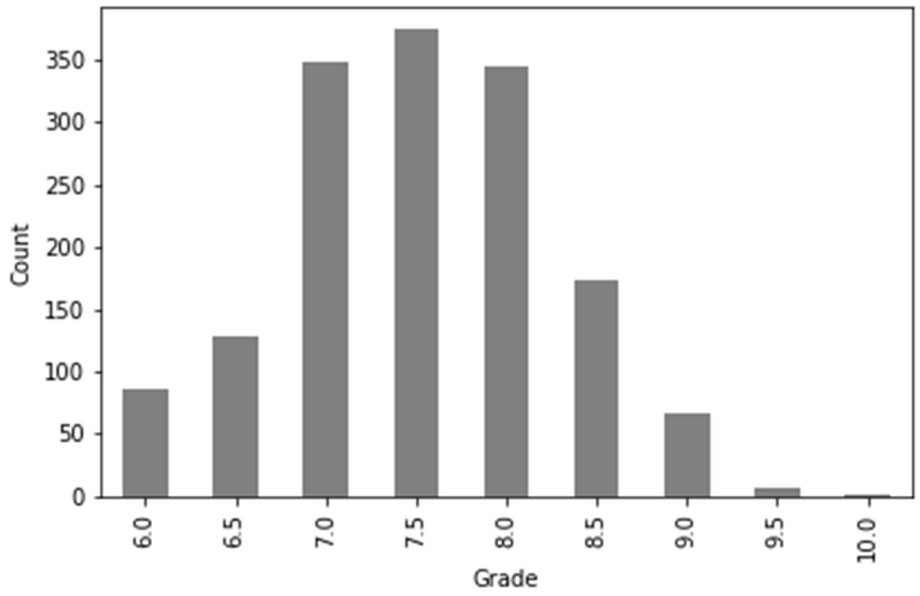
### The Data

In this study, two different datasets were used. The first dataset included bachelor's and master's theses texts and grades from 1366 students of the psychology program of the University of Amsterdam. The thesis files had to be matched with the grade information extracted from the university's student administration systems. Overall, the dataset included 1531 matched grades and theses from which 159 students published both a bachelor and master theses, 758 only a bachelor's and 614 only a master's theses to the University Library website. The publication date of the theses ranged from 2008 to 2019, and most of the theses are from the years 2014, 2015 and 2017. The average number of words per thesis is 6858.3.

Of the theses, 1218 were Dutch, and 312 were English. The dataset included grades ranging from 6.0 to 10.0, where a grade above 5.5 is a pass. The frequency distribution of grades is displayed in figure 1. Each thesis is graded by two independent supervisors based on standardised rubrics. Students gave consent for the usage of their data when enrolling for the study programme. The ethics review board approved the research project, and the researcher kept all information disclosed that could reveal the identity of the students to maintain their anonymity.

**Figure 1**

*Frequencies of grades*




The second set of datasets is the Hewlett essay datasets from which only the training set was used (The Hewlett Foundation: Automated Essay Scoring, 2012). This dataset comprised eight datasets of high school student essays with different scoring ranges. The average number of words of the essays in the datasets ranged from about 120 to about 620. See figures B1 to B4 for the frequency distribution of grades for each dataset. Two

independent raters rated all essays, and the target variable was calculated based on the two rater scores differently for each dataset. For dataset two, two target variables are available. Each dataset included different types of essays: persuasive, narrative, expository, or source-based.

**Materials**

The data extraction, cleaning, pre-processing, feature implementation, and modelling were implemented in a Jupyter notebook (Jupyter 6.4.5; Kluyver et al., 2016) using Python (Python 3.9.7, Van Rossum and Drake, 2009). The thesis text was extracted from PDF files with the python library Apache Tika[1] (Mattmann, 2021) and the language was detected with the library Language detection[2] (Nakatani, 2014). The "text" and "no text" parts of the thesis were classified with a pre-trained multilanguage BERT model[3] (Google, 2018). An English and a Dutch SpaCy model were used for text pre-processing and retrieving the part-of-speech tags (Honnibal & Montani, 2017). Moreover, several python packages were used for the feature implementation, such as Pandas (McKinney, 2010), NumPy (Harris et al., 2020), lexical-diversity[4] (Kyle, 2020) and Matplotlib (Hunter, 2007) for plotting the grade frequencies. The neural network model was fit using Keras (Chollet, 2015) and TensorFlow (Abadi et al., 2016). Furthermore, the library Scikit-learn was used for modelling, feature selection, hyperparameter tuning and evaluation metrics (Pedregosa et al., 2011). The remaining plots were created with the libraries ggplot2 (Wickham, 2016) and ggrepel (Slowikowski et al., 2021) using the programming language R (4.0.5, R Core Team, 2020) in the Rstudio environment (RStudio 1.4.11106; RStudio Team, 2018).

[1] https://github.com/chrismattmann/tika-python
[2] https://github.com/shuyo/language-detection
[3] https://tfhub.dev/tensorflow/bert_multi_cased_L-12_H-768_A-12/4
[4] https://github.com/kristopherkyle/lexical_diversity

***Feature Implementation***

Tables A1, A2 and A3 show the AGE+ features of Zupanc and Bosnić (2017) implemented in this study. Linguistic features included lexical sophistication, readability measures, lexical diversity measures, grammar, and mechanic features. Lexical sophistication consisted of features corresponding to average sentence length, the number of words or the number of stop words calculated with the Dutch and English NLTK stop word list (Bird, Klein, & Loper, 2009). Readability measures indicate the difficulty of a text, and lexical diversity measures the word variation (DuBay, 2007; Mellor, 2010; Smith & Jönsson, 2011). Table B1 lists the formulas and functions used to implement these features. The model for the prediction of the number of syllables of English words and the training data was adapted from Grasman (2021). We fit a similar model to the vocabulary of Dutch thesis texts to detect the number of syllables of Dutch words. A wordlist with the number of syllables per word[5] was created to serve as training data for the Dutch model (*Categorie:Woorden naar aantal lettergrepen in het Nederlands - WikiWoordenboek*, 2016).

To calculate the Dale-Chall readability formula for the English theses, the Dale-Chall word list[6], which contained 3000 of the most frequent English words, was used (Scott, n.d.). For the Dutch theses, a wordlist with 3000 most frequent Dutch words[7] was utilised (Dave, 2016). The first 1000 words of the Japan Association of College English Teachers (JACET) wordlist[8], which contained basic English words, were employed to calculate the advanced words (Ishikawa et al., 2003). For the Dutch theses, a list of 1013 basic Dutch words[9] was utilised instead (*Appendix:1000 basic Dutch words*, 2021).

[5] https://nl.wiktionary.org/wiki/Categorie:Woorden_naar_aantal_lettergrepen_in_het_Nederlands
[6] https://readabilityformulas.com/articles/dale-chall-readability-word-list.php
[7] from https://raw.githubusercontent.com/hermitdave/FrequencyWords/master/content/2016/nl/nl_50k.txt
[8] https://lextutor.ca/freq/lists_download/jacet/jacet1000.txt
[9] https://en.wiktionary.org/wiki/Appendix:1000_basic_Dutch_words

The categories of grammar and mechanics comprise features such as part-of-speech tags and language errors. To tag each word in a sentence with a part of speech (POS) tag, an English and Dutch SpaCy model was used (Honnibal & Montani, 2017). The python library Language tool was employed to detect grammar, spelling, punctuation, and capitalisation errors (Morris, 2020). Lastly, the coherence features comprise basic coherence features, spatial data analysis and spatial autocorrelation. These features were calculated based on TFIDF vectors of overlapping text parts. Some features were calculated twice, with the Euclidean distance and cosine similarity (Zupanc & Bosnić, 2017).

### *Model Evaluation*

The quality of the prediction models was determined with the QWK and Pearson's r. The QWK is a number between 0 and 1, where 1 indicates the highest agreement and 0 indicates random agreement between graders. In case of less agreement than at chance level, the metric takes on values below zero. Pearson's r is the correlation coefficient between the two grades. It ranges between 0 and 1 for predictive models. A value of 0 demonstrates no correlation, and a value of 1 indicates high positive correlations (Ramesh & Sanampudi, 2021).

## Procedure

As the first step, the researcher extracted the text from the PDFs and detected the language for each thesis. The extracted text was in raw text form. Besides the actual thesis text of abstract, introduction, methods, results, and discussion, the raw text also included the title page, the table of content, tables of the results, the reference list, and appendices. Therefore, the actual text first had to be identified before feature implementation. A pre-trained binary classification model was fit to classify paragraphs of the raw text as "text" or

"no text". The training data for this model consisted of a sample from the thesis data. The thesis texts were split into paragraphs, and the "no text" data was labelled by hand. Paragraphs such as the title page, the table of content, tables or figures in the text, the reference list and the appendix received the label "no text", and the remaining paragraphs obtained the label "text". After training, the model classified the remaining data and only the thesis paragraphs classified as "text" were selected and re-combined into one text for word and sentence tokenisation. The tokens were cleaned by removing whitespace and digits.

Several pre-processing steps described by Zupanc and Bosnić (2017) were taken before the features could be implemented. The text was tokenised into word and sentence tokens, and an additional list with stemmed word tokens was created. Before the readability indices could be calculated, the number of syllables was detected with a simple linear regression model for English words and Dutch words separately. For the content features, a bag-of-words vector of each thesis text was determined to calculate the cosine similarity between the texts. For the coherence features, the list of word tokens was separated into several overlapping parts. The length of each part was based on the average word count and a window size of 25% of the average word count. Each thesis part was determined in steps of ten words consisting of all the words within the window size. Finally, the TFIDF vectors were calculated for each thesis part, followed by the coherence features.

For reproducibility of the modelling results, a random seed was set to 42. The Hewlett training datasets and the thesis dataset were split into a train and a test set of 0.8 and 0.2, respectively. The datasets were fit on seven regression models, the five models also tested by Zupanc and Bosnić (2017), random forest (RF), linear regression (LR), a regression tree (RT), extremely randomised tree (ERT), and a neural network (NN). In addition to that, a lasso and a ridge regression model were fit. The hyperparameters of lasso, ridge, RF, RT and ERT were tuned with 10-fold cross-validation. The predictors were standardised for the NN,

Ridge and Lasso model. For the ridge and lasso model, different values for the alpha parameter were tested with a grid search. For the random forest, decision tree and extremely random decision tree, the best hyperparameters were determined by a random grid search. The neural network model consisted of input, hidden, and output layers and was trained on 300 epochs with Adam optimiser and mean squared error as loss function. The hidden and output layer had a linear activation function. Part of the training data was used as validation data (0.1). Early stopping was added to the model, and normal distribution was used to initialise weights. The number of neurons in the hidden layer was tuned from values between 1 and 45 to reach the lowest mean squared error. Finally, the coefficients of the lasso model were plotted against the actual values to compare the selected predictors among the datasets.

## Results

Two data points were removed for the analysis. One because the thesis file was not available and the second one because the text could not be extracted. Thirteen out of 101 AEG+ features were not implemented, and six features were combined in pairs. The "cosine similarity with source text" feature was not implemented because the source texts are the references for bachelor's and master's theses. Downloading all references requires a more complicated implementation than for the Hewlett essay datasets, where the source text is available for the source-based essays. The other features which could not be implemented are the twelve POS tag features 1) wh- determiner, 2) wh-pronoun, 3) wh – adverb, 4) predeterminer, 5) genitive marker, 6) comparative adverb, 7) superlative adverb and 8) existential there, 9) verb – gerund/present participle, 10) verb – past participle, 11) verb – third-person singular present, and 12) ordinal adjective or numeral. The specific tag identifying these word types for these features could not be identified for the Dutch spacy model. Other than for the English POS tags, which differ from the Dutch POS tags, no list is

available that explains which grammatical word type each tag stands for. For the same reason, some of the POS tag features were combined as one feature. For instance, "singular or mass common noun" and "plural common noun" were combined to the feature "common noun", "singular proper noun", and "plural proper noun" to "proper noun" and finally ", personal pronoun" and "possessive pronoun" to "pronouns".

Table 4 shows the QWK and Pearson's r of the thesis data for each model. Overall, the models' QWK and Pearson's r scores are low. Lasso, ridge, RF and the NN model have a moderate positive correlation with the human rater's grade and the remaining models, LR, ERT and RT, have low positive correlations.

**Table 4**

*Quadratic weighted kappa, Pearson's r and confidence interval of the predicted and actual values for the theses data*

| | | | 95% CI | |
|---|---|---|---|---|
| Model | QWK | Pearson's r | LL | UL |
| NN | .23 | $r(302) = .32, p = < .0001$ | .215 | .417 |
| Lasso | .20 | $r(302) = .35, p = < .0001$ | .246 | .444 |
| Ridge | .20 | $r(302) = .21, p = < .0001$ | .103 | .318 |
| LR | .16 | $r(302) = .27, p = < .0001$ | .163 | .372 |
| RF | .17 | $r(302) = .33, p = < .0001$ | .221 | .422 |
| ERT | .11 | $r(302) = .18, p = .002$ | .067 | .284 |
| RT | .11 | $r(302) = .28, p = < .0001$ | .172 | .380 |

*Note.* The seven regression models are random forest (RF), regression tree (RT), extremely randomized trees (ERT), linear regression (LR), lasso regression (Lasso), ridge regression (Ridge) and neural network (NN). All models except for LR were tuned.

Table 5 shows the QWK, and table 6 displays Pearson's r of the Hewlett essay datasets separately for each model and an average over all models. All seven models display high agreement scores. The best model based on the average QWK is the NN model with 39

neurons in the hidden layer, followed by the ridge and lasso regression, LR, RF, RT and ERT model. QWK values indicate substantial to almost perfect agreement. The best model with the highest Pearson's r on average over the Hewlett essay datasets is lasso regression, followed by the NN model, ridge regression, LR, RF, RT and ERT.

**Table 5**

*Quadratic weighted kappa of seven regression fit on the Hewlett essay datasets*

| Model | DS1 | DS2A | DS2B | DS3 | DS4 | DS5 | DS6 | DS7 | DS8 | Average |
|---|---|---|---|---|---|---|---|---|---|---|
| NN | **0.84** | **0.69** | **0.70** | **0.67** | **0.73** | 0.80 | **0.75** | **0.82** | **0.72** | 0.75 |
| Ridge | 0.83 | 0.67 | 0.67 | 0.63 | 0.65 | 0.79 | 0.70 | 0.81 | 0.71 | 0.72 |
| Lasso | 0.84 | 0.66 | 0.65 | 0.59 | 0.68 | **0.80** | 0.73 | 0.80 | 0.68 | 0.72 |
| LR | 0.83 | 0.67 | 0.66 | 0.64 | 0.69 | 0.79 | 0.74 | 0.76 | 0.58 | 0.71 |
| RF | 0.75 | 0.56 | 0.56 | 0.61 | 0.56 | 0.72 | 0.65 | 0.75 | 0.60 | 0.64 |
| RT | 0.80 | 0.54 | 0.54 | 0.58 | 0.65 | 0.75 | 0.63 | 0.69 | 0.53 | 0.64 |
| ERT | 0.75 | 0.58 | 0.56 | 0.53 | 0.57 | 0.72 | 0.63 | 0.60 | 0.49 | 0.60 |

*Note*. For each dataset, the best models result is marked in bold. The seven regression models are random forest (RF), regression tree (RT), extremely randomized trees (ERT), linear regression (LR), lasso regression (Lasso), ridge regression (Ridge) and neural network (NN).

**Table 6**

*Pearson's r of the regression models fit on the Hewlett essay datasets*

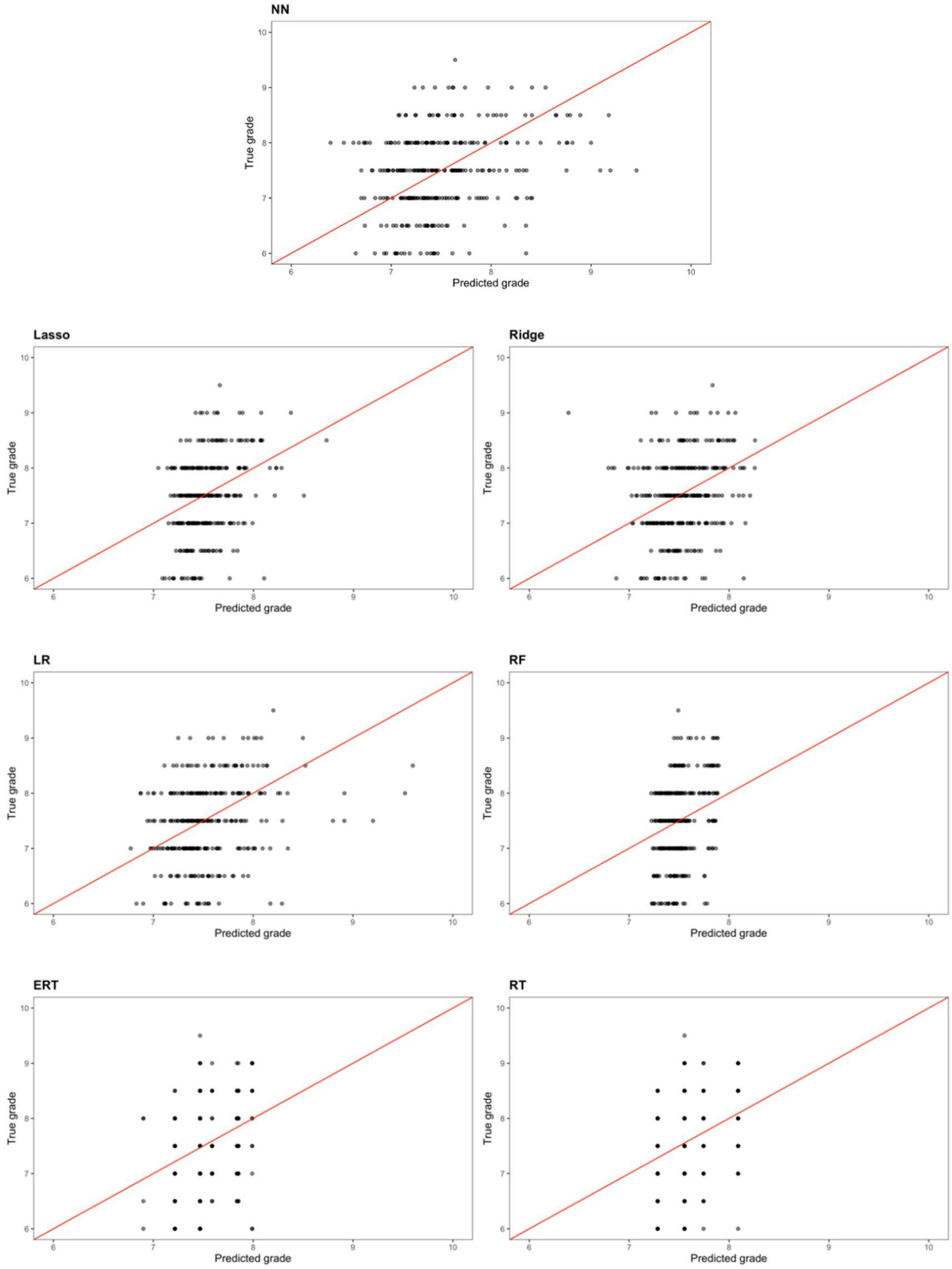
| Model | DS1 | DS2A | DS2B | DS3 | DS4 | DS5 | DS6 | DS7 | DS8 | Average |
|---|---|---|---|---|---|---|---|---|---|---|
| Lasso | **0.87** | **0.74** | **0.75** | 0.69 | 0.76 | **0.83** | **0.80** | 0.82 | 0.72 | 0.77 |
| NN | 0.87 | 0.74 | 0.74 | 0.68 | **0.77** | 0.83 | 0.80 | **0.82** | **0.72** | 0.77 |
| Ridge | 0.86 | 0.73 | 0.74 | 0.69 | 0.77 | 0.83 | 0.79 | 0.81 | 0.71 | 0.77 |
| LR | 0.85 | 0.72 | 0.73 | **0.69** | 0.77 | 0.83 | 0.78 | 0.76 | 0.59 | 0.75 |
| RF | 0.84 | 0.70 | 0.70 | 0.69 | 0.76 | 0.81 | 0.75 | 0.78 | 0.70 | 0.75 |
| RT | 0.82 | 0.62 | 0.62 | 0.68 | 0.74 | 0.79 | 0.70 | 0.72 | 0.58 | 0.70 |
| ERT | 0.75 | 0.58 | 0.56 | 0.53 | 0.57 | 0.73 | 0.63 | 0.60 | 0.49 | 0.61 |

*Note*. For each dataset, the best models result is marked in bold. The seven regression models are random forest (RF), regression tree (RT), extremely randomized trees (ERT), linear regression (LR), lasso regression (Lasso), ridge regression (Ridge) and neural network (NN).

Figure 2 presents the predicted grades of the seven models plotted against the actual grades. The red line displays the actual grades. The NN and linear regression models' predictions are more spread out than the other models. These two models and the lasso and ridge regression are closest to the red line. The RT and ERT models perform poorly as the predictions deviate from the red line the most compared to the other models. Overall, most of the predictions across all models lie between grades 7 and 8. The RF model predictions are limited to this range. The scatterplots of predicted and actual grades for the Hewlett dataset are displayed in figures C1 to C9.

**Figure 2**

*Scatter plots of predicted grades on the x axis and true grades on the y-axis for the thesis dataset*




*Note.* Random forest (RF), regression tree (RT), extremely randomized trees (ERT), linear regression (LR). neural network (NN), ridge regression (Ridge) and lasso regression (Lasso).

Tables D1 displays the coefficients of the lasso regression for thesis texts, and table D2 shows the summed lasso coefficients of the Hewlett datasets. Twenty-three predictors were influential for the thesis data and 79 predictors for the essay data. The feature cosine similarity with essays that have highest score point level is selected for the thesis data but for none of the essay datasets. Out of 85 predictors, the five predictors are not important for both the thesis and the Hewlett datasets: 1) number of different words, 2) number of words, 3) average and 4) minimum cosine similarity between points and 5) index (minimum distance/maximum distance) (basic coherence measure). Table 7 shows the frequency counts of selected features per category. What stands out for the thesis data is that all content features and no spatial autocorrelation features were selected. Overall, coherence, POS tag, lexical sophistication, readability measures and lexical diversity were less often selected for thesis texts than for essay texts. In contrast, grammar and mechanic features were selected equally as much.

**Table 7**

*Frequency count of features determined by lasso regression for theses and essay texts by feature category*

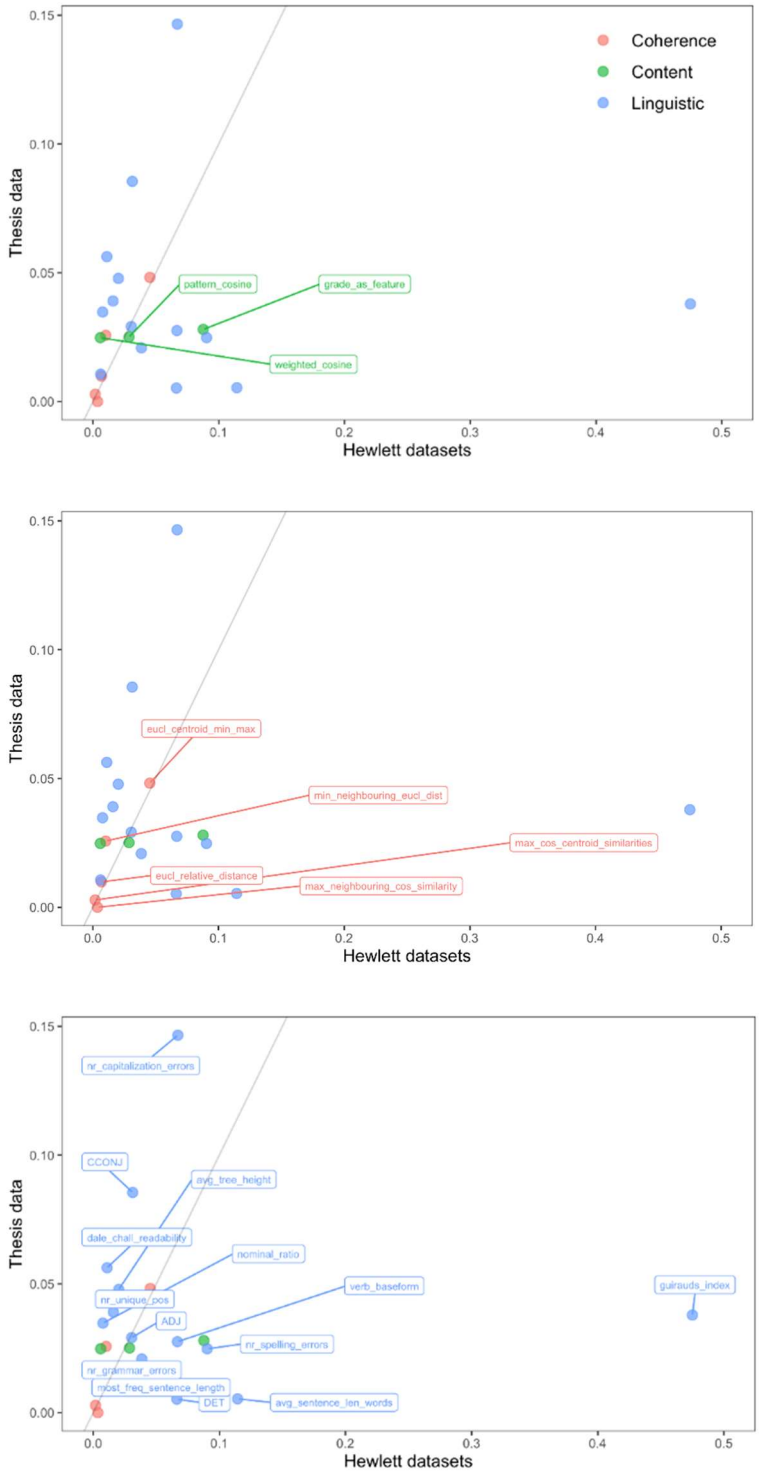
| Feature category | DS 1 | DS 2A | DS 2B | DS 3 | DS 4 | DS 5 | DS 6 | DS 7 | DS 8 | Thesis text | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Linguistic | | | | | | | | | | | |
| Lexical sophistication | 5 | 7 | 7 | 4 | 4 | 8 | 4 | 5 | 4 | 2 | 13 |
| Readability measures | 3 | 0 | 0 | 5 | 4 | 8 | 6 | 5 | 1 | 2 | 9 |
| Lexical diversity | 4 | 2 | 2 | 3 | 2 | 5 | 2 | 3 | 1 | 1 | 6 |
| Grammar | 2 | 3 | 3 | 4 | 3 | 4 | 3 | 4 | 1 | 3 | 4 |
| POS tag | 11 | 11 | 9 | 13 | 13 | 15 | 13 | 11 | 3 | 4 | 17 |
| Mechanic | 3 | 2 | 2 | 2 | 1 | 3 | 3 | 3 | 1 | 2 | 3 |
| Content | | | | | | | | | | | |
| Other | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 4 | 4 |
| Coherence | | | | | | | | | | | |
| Basic coherence measures | 5 | 1 | 1 | 4 | 5 | 7 | 3 | 4 | 3 | 2 | 15 |
| Spatial data analysis | 4 | 4 | 4 | 3 | 3 | 6 | 5 | 3 | 2 | 3 | 11 |
| Spatial autocorrelation | 0 | 1 | 1 | 2 | 2 | 2 | 0 | 1 | 1 | 0 | 3 |

*Note.* Datasets 1 and 2 are persuasive essays, 3 – 6 are source-based essays, 7 expository and 8 narrative essays. The total column displays the total number of features per features category.

Figure 3 visualises the relative importance of the features in the thesis data set vs the essay datasets. The thesis data coefficients plotted with each Hewlett dataset separately are displayed in figures D1 to D9. The coefficients along the diagonal are influential for both thesis and essay texts, where the ones further away from 0 along the diagonal are more important for both datasets than the ones close to zero. Seven features are equally important for thesis and essay texts. The content feature pattern cosine and two linguistic features (number of adjectives and most frequent sentence length), as well as three spatial data analysis features (relative distance, index (minimum distance/maximum distance), maximum cosine similarity between points and centroid) and the basic coherence measure maximum cosine similarity between neighbouring points. The linguistic features: number of capitalisation errors, number of coordinating conjunction, the dale-chall readability measure, and the height of the tree presenting sentence structure are the most important coefficient for thesis texts that are also important for essay texts. Guiraud's index and average sentence length are the most influential feature for essay texts that are also important for thesis texts.

**Figure 3**

*Coefficients of the lasso regression both important for the thesis data and all Hewlett essay datasets*




*Note.* The coefficients of the Hewlett datasets were summed and divided by the number of datasets.

## Discussion

The present study sought to investigate whether AGE+ successfully predicts a grade for bachelor's and master's theses texts and specifically tried to identify which features are important for predicting domain-specific texts. It was found that AGE+ was successful for the Hewlett datasets, as reported by Zupanc and Bosnić (2017) but it was not successful in predicting the grades of domain-specific theses texts. Furthermore, the results show that only a few features are important to predict both types of text. For thesis texts, features of the categories content, mechanics, grammar, and coherence were more determinative than most lexical sophistication and most POS tag features.

Based on the average QWK of all Hewlett datasets, the NN model performed better (.75), and the LR model (.71) performed only slightly worse than reported by Zupanc and Bosnić (2017). They reported an average QWK of .7039 and .7280, respectively. However, the other models, RF, RT and ERT, performed worse than Zupanc and Bosnić (2017) reported. The differences may be explained by the smaller testing set applied in this study, as the training data had to be split into train and test sets because the original test sets of the Hewlett datasets are no longer available. Moreover, the models were trained with fewer features than Zupanc and Bosnić (2017). The ontological consistency features were not implemented in this study, and not all the linguistic and content features were successfully implemented. Therefore, it could not be determined whether these features are as important for thesis texts as they are for determining the grade for essays. However, because the content features are important for predicting the grades of thesis texts, it is suspected that the feature cosine similarity with source text might be equally meaningful. Some of the unimplemented linguistic features might also be important for thesis texts because the first eight of the not implemented POS tag features and the feature number of possessive pronouns are among the top 50 most important features mentioned in Zupanc and Bosnić (2017).

There are several possible explanations for the system performing better for essay texts than thesis texts. One explanation might be that the differences in writing do not vary as much for university students in this dataset as for high school students in the Hewlett datasets. Additionally, the content and language of the thesis texts were probably extensively checked, and the university student received feedback multiple times while writing their thesis to improve the quality and flow of the text. In contrast, the essay texts were written by high school students, which have more varying writing skills among each student and lower writing skills in general than university students. Furthermore, the high school students probably did not receive feedback to improve their essays. Therefore, there might be more variation in the flow of the essay texts, captured by the coherence features and more variation in the language quality, captured by the linguistic features. The possible higher variance in the features may have led to better predictions. In practice, the university students would probably apply the AEG systems during earlier stages of the writing process and not only on the final product.

A limitation of this study is that the thesis dataset only includes thesis texts that count as passed and none that count as failed. The distinction is therefore never learned by the system. The theses texts in this dataset are more similar to each other than the essay texts, which also include failed essays. Another point to consider is the quality of features which leaves room for improvement. For instance, features such as capitalisation errors, number of sentences, and sentence length are based on the quality of SpaCy's sentence tokenisation, which could not detect all the sentences and cut some sentences in half. Another example is the features involving the number of syllables. Two linear regression models predicted the number of syllables for each word. The models achieved r-square values of 0.94 for the English model and 0.84 for the Dutch model. Moreover, some of the readability and lexical

diversity measures were calculated based on English word lists like the dale-chall wordlist, for which no exact Dutch equivalent was available.

Further research is necessary for the evaluation of domain-specific texts. A start for future research could be to test the features left out in this study on a dataset that includes failed theses. Moreover, it is necessary to develop new features tailored explicitly to domain-specific texts. For instance, in the case of bachelor and master thesis texts, the number of references, the number of APA mistakes, paragraph length, and the coherence within each paragraph and among the paragraphs could be tested. Moreover, especially the content features were determinative for thesis texts. Therefore, in addition to cosine similarity with the references of a thesis, the content features could be applied explicitly to other theses with similar topics rather than to theses with varying topics. The features may be even more influential when applied to theses in a similar domain.

An alternative approach to predicting the grade for the whole thesis is to apply the system to each part of the thesis (abstract, introduction, methods, results and discussion) separately and receive a grade for each section to provide student with more feedback. This approach was attempted in this study. However, separating the different parts of the thesis was unsuccessful because of the formatting differences in the raw text. Furthermore, additional independent evaluation criteria, such as grades from previous writing assignments as tested by Powers et al. (2002), should be considered to increase the validity of AEG systems instead of only aiming for high agreement with teacher grades. In conclusion, the AGE+ system cannot predict grades for bachelor and master theses as successfully as for short essay texts. However, valuable insights were gained about determinative features for domain-specific texts, and a direction for further research was set to establish an AEG system that successfully predicts grades of bachelor and master theses.

## References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, Kaiser, L., Kudlur, M., … & Zheng, X. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*.

*Appendix:1000 basic Dutch words*. (2021, December 31). Wiktionary, the free dictionary. https://en.wiktionary.org/wiki/Appendix:1000_basic_Dutch_words

Bird, S., Klein, E., & Loper, E. (2009). Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit (1st ed.). O'Reilly Media. https://www.nltk.org/book/

Burstein, J., Kukich, K., Wolff, S., Lu, C., & Chodorow, M. (1998, April). Computer analysis of essays. In *NCME Symposium on automated Scoring*.

*Categorie:Woorden naar aantal lettergrepen in het Nederlands - WikiWoordenboek*. (2016, May 16). WikiWoordenboek. Retrieved May 21, 2022, from https://nl.wiktionary.org/wiki/Categorie:Woorden_naar_aantal_lettergrepen_in_het_Nederlands

Chollet, F. (2015). Keras. GitHub. Retrieved from https://github.com/fchollet/keras

Chung, K. W. K. & O'Neil, H. F. (1997). Methodological approaches to online scoring of essays (ERIC reproduction service no ED 418 101).

Cummins, R., Zhang, M., & Briscoe, E. (2016, August). Constrained multi-task learning for automated essay scoring. Association for Computational Linguistics.

Dave, H. (2016, August 16). *FrequencyWords/content/2016/nl/nl_50k.txt*. GitHub. Retrieved April 29, 2022, from https://raw.githubusercontent.com/hermitdave/FrequencyWords/master/content/

2016/nl/nl_50k.txt

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dong, F., & Zhang, Y. (2016, November). Automatic Features for Essay Scoring-An Empirical Study. In *EMNLP* (Vol. 435, pp. 1072-1077).

DuBay, W. H. (2007). Smart Language: Readers, Readability, and the Grading of Text. http://www.impact-information.com/impactinfo/newsletter/smartlanguage02.pdf

Foltz, P. W., Laham, D., & Landauer, T. K. (1999). The intelligent essay assessor: Applications to educational technology. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning*, *1*(2), 939-944.

Google. (2018). *bert_multi_cased_L-12_H-768_A-12*. TensorFlow Hub. Retrieved April 18, 2022, from https://tfhub.dev/tensorflow/bert_multi_cased_L-12_H-768_A-12/4

Grasman, R. (2021, September 1). *A syllable counting model exercise with caret*. Kaggle: Your Machine Learning and Data Science Community. https://www.kaggle.com/code/datasniffer/a-syllable-counting-model-exercise-with-caret

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., ... & Berg, S. (2020). Smith 474 nj. *Kern R, Picus M, Hoyer S, van Kerkwijk MH, Brett M, Haldane A, del R'ıo JF, Wiebe M, Peterson P, G'erard-475 Marchant P, et al. Array programming with NumPy. Nature*, *585*(7825), 357-362.

Honnibal, M., & Montani, I. (2017). *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing*.

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. Computing in Science & Engineering, 9(3), 90–95.

Hussein, M. A., Hassan, H., & Nassef, M. (2019). Automated language essay scoring systems: A literature review. *PeerJ Computer Science*, *5*, e208.

Ishikawa, S., Uemura, T., Kaneda, M., Shimizu, S., Sugimori, N., & Tono, Y.（2003）. JACET8000: JACET list of 8000 basic words. Tokyo: JACET.

Kakkonen, T., Myller, N., & Sutinen, E. (2006). Applying Part-of-Seech Enhanced LSA to Automatic Essay Grading. *arXiv preprint cs/0610118*.

Kluyver, T., Ragan-Kelley, B., Perez, F., Granger, B., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J., Corlay, S., Ivanov, P., Avila, D., Abdalla, S., Willing, C., and Jupyter development team, (2016) Jupyter notebooks – a publishing format for reproducible computational workflows. In F. Loizides &B. Schmidt (Eds.),Positioning and power in academic publishing: Players, agents and agendas(p. 87- 90).

Kyle, K. (2020, March 4). *lexical diversity*. GitHub. Retrieved May 21, 2022, from https://github.com/kristopherkyle/lexical_diversity

Mattmann, C. (2021, June 7). *Tika-Python*. GitHub. Retrieved February 28, 2022, from https://github.com/chrismattmann/tika-python

McKinney, W. (2010, June). Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference* (Vol. 445, No. 1, pp. 51-56).

Mellor, A. (2010). Essay Length, Lexical Diversity and Automatic Essay Scoring. In Technology, Series B (Vol. 55, Issue 2).

Morris, J. (2020, April). *language_tool_python: A free python grammar checker*. GitHub. Retrieved April 12, 2022, from https://github.com/jxmorris12/language_tool_python

Mosallam, Y., Toma, L., Adhana, M. W., Chiru, C. G., & Rebedea, T. (2014, January). Unsupervised system for automatic grading of bachelor and master thesis. In *Proceedings of the International Conference on eLearning and Software for*

*Education (eLSE 2014)* (pp. 165-171).

Nakatani, S. (2014, March 3). *language-detection*. GitHub. Retrieved February 28, 2022, from https://github.com/shuyo/language-detection

Nissel, M. (2015, June 5). *Yule's K and Yule's I for lexical diversity in Python 3*. GitHubGist. https://gist.github.com/magnusnissel/d9521cb78b9ae0b2c7d6

Page, E. B. (1966). The imminence of... grading essays by computer. *The Phi Delta Kappan*, *47*(5), 238-243.

Page, E. B. (1968). The use of the computer in analyzing student essays. *International review of education*, 210-225.

Pedregosa, F., Varoquaux, Ga"el, Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Müller, A., Nothman, J., Louppe, G., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Coumapean, D., Brucher, M., Perrot, M., Duchesnav, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*(Oct), 2825–2830.

Powers, D. E., Burstein, J. C., Chodorow, M. S., Fowles, M. E., & Kukich, K. (2002). Comparing the validity of automated and human scoring of essays. *Journal of Educational Computing Research*, *26*(4), 407-425.

Ramesh, D., & Sanampudi, S. K. (2021). An automated essay scoring systems: a systematic literature review. *Artificial Intelligence Review*, 1-33.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rezaei, A. R., & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing writing*, *15*(1), 18-39.

RStudio Team. (2018). Rstudio: Integrated development environment for R [Computer software manual].Boston, MA. Retrieved fromhttp://www.rstudio.com/

Sahlgren, M. (2008). The distributional hypothesis. *Italian Journal of Disability Studies*, *20*, 33-53.

Scott, B. (n.d.). *The Dale-Chall 3,000 Word List for Readability Formulas*. Readability Formulas. Retrieved April 14, 2022, from https://readabilityformulas.com/articles/dale-chall-readability-word-list.php

Slowikowski, K., Schep, A., Hughes, S., Dang, T. K., Lukauskas, S., Irisson, J. O., Kamvar, Z. N., Ryan, T., Christophe, D., Hiroaki, Y., Gramme, P., Abdol, A. M., Barrett, M., Cannoodt, R., Krassowski, M., Chirico, M., & Aphalo, P. (2021, January 15). *CRAN - Package ggrepel*. CRAN. Retrieved May 14, 2022, from https://cran.r-project.org/web/packages/ggrepel/index.html

Smith, C., & Jönsson, A. (2011). Automatic Summarization As Means Of Simplifying Texts, An Evaluation For Swedish. 198–205.

Taghipour, K., & Ng, H. T. (2016, November). A neural approach to automated essay scoring. In *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 1882-1891).

*The Hewlett Foundation: Automated Essay Scoring*. (2012, February 10). Kaggle. Retrieved February 16, 2022, from https://www.kaggle.com/c/asap-aes

Van Rossum, G., & Drake, F. L. (2009). *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace.

Wang, Z., Liu, J., & Dong, R. (2018, November). Intelligent Auto-grading System. In *2018 5th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS)* (pp. 430-435). IEEE.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. Retrieved from https://ggplot2.tidyverse.org

Zhao, S., Zhang, Y., Xiong, X., Botelho, A., & Heffernan, N. (2017, April). A memory-

augmented neural model for automated grading. In *Proceedings of the fourth (2017) ACM conference on learning@ scale* (pp. 189-192).

Zupanc, K., & Bosnić, Z. (2017). Automated essay evaluation with semantic analysis. *Knowledge-Based Systems*, *120*, 118-132.

# Appendix A

## Table A1

*AGE+ linguistic features from Zupanc and Bosnić (2017)*

| Lexical sophistication | Readability measures | Lexical diversity | Grammar | Mechanics |
|---|---|---|---|---|
| 1. Number of characters | 14. Gunning Fox index | 23. Type-token-ratio | 29. Number of different PoS tags | 50. Number of spell-checking errors |
| 2. Number of words | 15. Flesch reading ease | 24. Guiraud's index | 30. Height of the tree presenting sentence structure | 51. Number of capitalization errors |
| 3. Number of long words | 16. Flesch Kincaid grade level | 25. Yule's K | 31. Correct verb form | 52. Number of punctuation errors |
| 4. Number of short words | 17. Dale-Chall readability formula | 26. The D estimate | 32. Number of grammar errors | |
| 5. Most frequent word length | 18. Automated readability index | 27. Hapax legomena | 33 - 49. Number of each POS tag | |
| 6. Average word length | 19. Simple measure of Gobbledygook | 28. Advanced Guiraud | 33. Coordinating conjunction | |
| 7. Number of sentences | 20. LIX | | 34. Preposition / subordinating conjunction | |
| 8. Number of long sentences | 21.Word variation index | | 35. Preposition | |
| 9. Number of short sentences | 22. Nominal ratio | | 36. Numeral | |
| 10. Most frequent sentence length | | | 37. Determiner | |
| 11. Average sentence length | | | 38. Adjective | |
| 12. Number of different words | | | 39. Comparative adjective | |
| 13. Number of stop words | | | 40. Superlative adjective | |
| | | | 41. Modal auxiliary | |
| | | | 42. Participle | |

| | |
|---|---|
| | 43. Infinitive marker<br>44. Verb- base form<br>45. Verb- past tense<br>46. Adverb<br>*47. Common noun (singular or mass and plural)*<br>*48. Proper noun (singular and plural)*<br>*49. Pronoun (personal and possessive)* |

*Note.* The features in italics are a combination of two features. The POS tag features not implemented: Existential there; Comparative adverb; Superlative adverb; Verb-gerund/present participle, Verb-past participle; Verb-3rd person singular present; Wh-determiner; Wh-pronoun; Wh-adverb; Ordinal adjective or numeral; Predeterminer; Genitive marker

**Table A2**

*AGE+ content features from Zupanc and Bosnić (2017)*

| Content |
|---|
| 1. Score point level for maximum cosine similarity over all score points |
| 2. Cosine similarity with essays that have highest score point level |
| 3. Pattern cosine |
| 4. Weighted sum of all cosine correlation values |

*Note.* The feature Cosine similarity with source text was not implemented.

**Table A3**

*AGE+ coherence features from Zupanc and Bosnić (2017)*

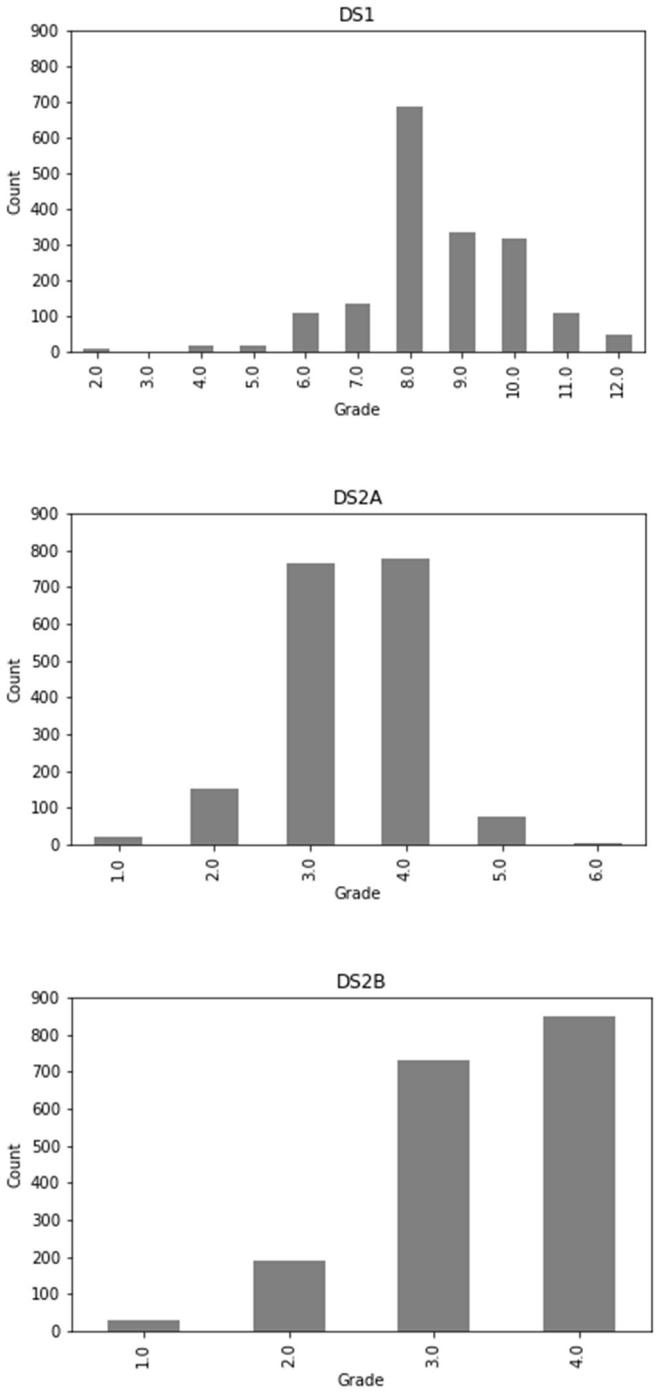
| Basic coherence measures | Spatial data analysis | Spatial autocorrelation |
|---|---|---|
| 1. - 2. Average distance between neighbouring points (2x) | 16. – 17. Average distance between points and centroid (2x) | 27. Moran's I |
| 3. – 4. Minimum distance between neighbouring points (2x) | 18. – 19. Minimum distance between points and centroid (2x) | 28. Geary's C |
| 5. – 6. Maximum distance between neighbouring points (2x) | 20. – 21. Maximum distance between points and centroid (2x) | 29. Getis's G |
| 7. – 8. Index (minimum distance/maximum distance) (2x) | 22. – 23. Index (minimum distance/maximum distance) (2x) | |
| 9. – 10. Average distance between any two points (2x) | 24. Standard distance | |
| 11. – 12. Maximum difference between any two points (2x) | 25. Relative distance | |
| 13. Clark's and Evans' distance to nearest neighbour | 26. Determinant of distance matrix | |
| 14. Average distance to nearest neighbour | | |
| 15. Cumulative frequency distribution | | |

*Note.* Most of the coherence features are measured twice, once with Euclidian distance and once with cosine similarity.

# Appendix B

## Figure B1

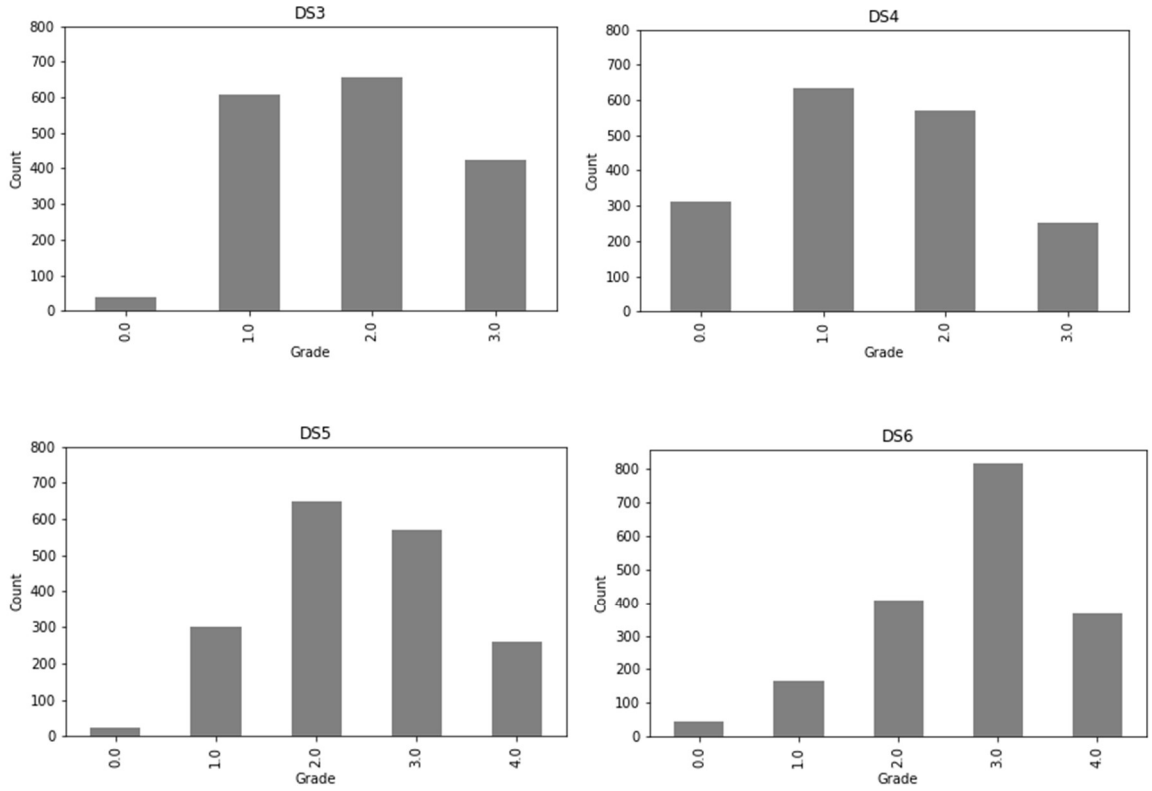
*Distribution of grades for persuasive essays*




*Note*. Dataset 1 contains 1783 essays with an average number of 371.04 words. Dataset 2 contains 1800 essays with an average number of 387.28 words.

**Figure B2**

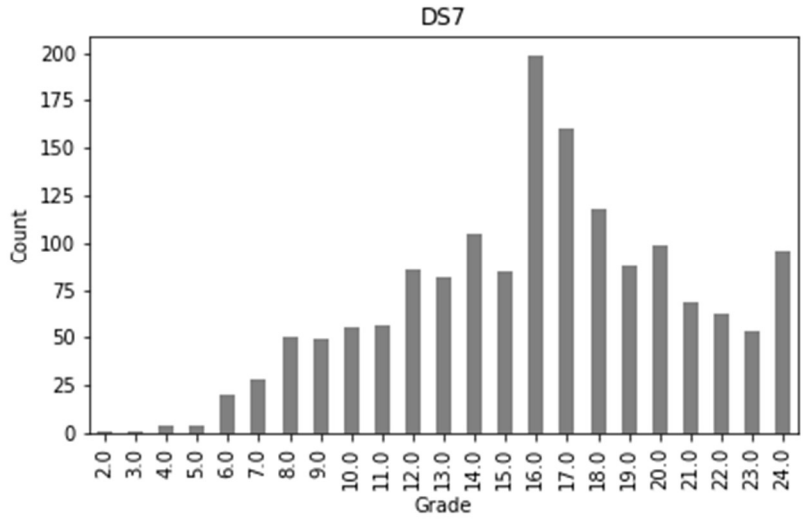
*Distribution of grades for source-based essays*




*Note*. Dataset 3 contains 1726 essays with an average number of 111.86 words. Dataset 4 contains 1772 essays with an average number of 96.49 words. Dataset 5 contains 1805 essays with an average number of 123.26 words. Dataset 6 contains 1800 essays with an average number of 154.56 words.

**Figure B3**

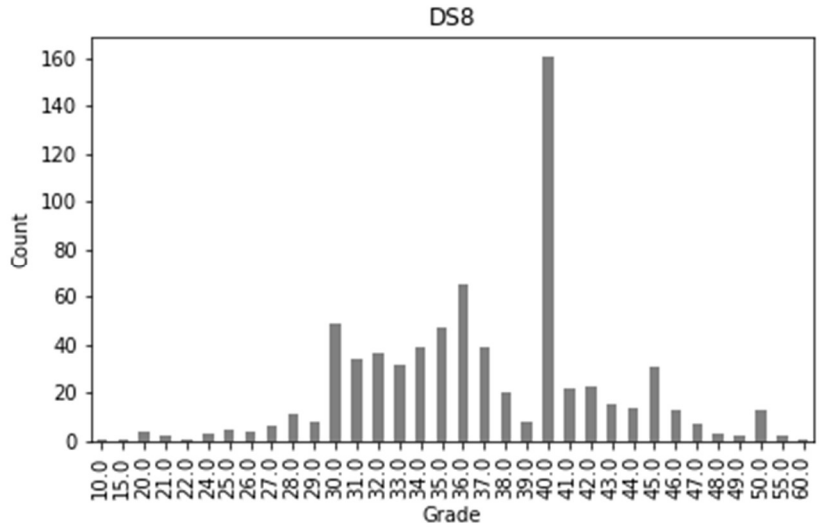
*Distribution of grades for expository essays*




*Note.* Dataset 7 contains 1569 essays with an average number of 171.94 words.

**Figure B4**

*Distribution of grades for Narrative essays*




*Note.* Dataset 8 contains 723 essays with an average number of 614.41 words.

**Table B1**

*Formulas of Readability indices and Lexical diversity features.*

| Feature | Formula | Source |
|---|---|---|
| | Readability Measures | |
| Gunning Fog Index | 0.4 * (average sentence length + hard words) | (DuBay, 2007) |
| Flesch reading ease | 206.835 − 1.015 * average number of words per sentence − 84.6 * average number of syllables | (Smith & Jönsson, 2011) |
| Flesch Kincaid grade level | 0.39 * average number words per sentence + 11.8 average number syllables per word - 15.59 | (DuBay, 2007) |
| Automated readability index | 0.50 * average words per sentence + 4.71 * average chars per word - 21.43 | (DuBay, 2007) |
| Simple measure of Gobbledygook | 3 + sqrt(number of polysyllables) | (DuBay, 2007) |
| LIX | average nr of words per sentence + ((number of words > 6 chars) / number of words) * 100 | (Smith & Jönsson, 2011) |
| Word variation index | log(number of words) / log( 2 - (log(number of unique words) / log(number of words))) | (Smith & Jönsson, 2011) |
| Nominal ratio | (number of nouns + number of prepositions + number of participles) / (number of pronouns + number of adverbs + number of verbs) | (Smith & Jönsson, 2011) |
| Dale-Chall readability formula | 0.1579 * PDW + 0.0496 * average words per sentence + 3.6365 | (DuBay, 2007) |
| | Lexical Diversity | |

| | | |
|---|---|---|
| Hapax legomena | number of words occurring only once in a text | (Mellor, 2010) |
| Type token ratio | number of different words / total number of words (lexical-diversity library) | (Kyle, 2020) |
| Guiraud's index | number of unique words / sqrt(N) | (Mellor, 2010) |
| Yule's K | 10^4 * sum(r^2 * Wr - N) / N^2 (r = 1, 2, ...) | (Mellor, 2010) (Nissel, 2015) |
| The D estimate | hypergeometric distribution D (lexical-diversity library) | (Kyle, 2020) |
| Advanced Guiraud | number advanced word types / sqrt(N) | (Mellor, 2010) |

*Note.* Hard words are the number of words with more than two syllables. Polysyllables are words with more than three syllables. N is the number of words in the document. Wr is the number of word types occurring r times in a text consisting of N token words. Advanced word types are words that do not occur in the JACET wordlist[6] (Ishikawa et al., 2003). For the Dutch theses, a list of 1013 basic Dutch words[7] was utilised (*Appendix:1000 basic Dutch words*, 2021). PWD is the percentage of words not in the Dale-Chall list[4] (Scott, n.d.). For the Dutch theses, a wordlist with 3000 most frequent Dutch words[5] was utilised (Dave, 2016). For the Dale-Chall readability formula a sample size of 100 was chosen for the Hewlett essay data and a sample size of 300 for the bachelor and master theses.

# Appendix C

**Table C1**

*Quadratic weighted kappa, Pearson's r and confidence interval of the regression models fit on DS1 of the Hewlett essay datasets*

| Model | QWK | 95% CI | | | p value |
|---|---|---|---|---|---|
| | | Pearson's r | LL | UL | |
| LR | .83 | .85 | .82 | .88 | <.0001 |
| RF | .75 | .84 | .80 | .87 | <.0001 |
| RT | .80 | .82 | .78 | .85 | <.0001 |
| ERT | .75 | .75 | .70 | .79 | <.0001 |
| NN | .84 | .87 | .83 | .89 | <.0001 |
| Ridge | .83 | .86 | .82 | .88 | <.0001 |
| Lasso | .84 | .87 | .84 | .89 | <.0001 |

*Note.* The seven regression models are random forest (RF), regression tree (RT), extremely randomized trees (ERT), linear regression (LR), lasso regression (Lasso), ridge regression (Ridge) and neural network (NN). All models except for LR were tuned.

**Table C2**

*Quadratic weighted kappa, Pearson's r and confidence interval of the regression models fit on DS2A of the Hewlett essay datasets*

| Model | QWK | Pearson's r | 95% CI | | p value |
|---|---|---|---|---|---|
| | | | LL | UL | |
| LR | .67 | .72 | .67 | .77 | <.0001 |
| RF | .56 | .70 | .64 | .75 | <.0001 |
| RT | .54 | .62 | .55 | .68 | <.0001 |
| ERT | .58 | .58 | .51 | .65 | <.0001 |
| NN | .69 | .73 | .68 | .77 | <.0001 |
| Ridge | .67 | .73 | .68 | .78 | <.0001 |
| Lasso | .66 | .74 | .69 | .79 | <.0001 |

*Note.* The seven regression models are random forest (RF), regression tree (RT), extremely randomized trees (ERT), linear regression (LR), lasso regression (Lasso), ridge regression (Ridge) and neural network (NN). All models except for LR were tuned.

**Table C3**

*Quadratic weighted kappa, Pearson's r and confidence interval of the regression models fit on DS2B of the Hewlett essay datasets*

| Model | QWK | Pearson's r | 95% CI | | p value |
|---|---|---|---|---|---|
| | | | LL | UL | |
| LR | .66 | .73 | .68 | .78 | <.0001 |
| RF | .56 | .70 | .64 | .75 | <.0001 |
| RT | .54 | .62 | .55 | .68 | <.0001 |
| ERT | .56 | .56 | .49 | .63 | <.0001 |
| NN | .70 | .74 | .69 | .79 | <.0001 |
| Ridge | .67 | .74 | .69 | .78 | <.0001 |
| Lasso | .65 | .75 | .70 | .79 | <.0001 |

*Note.* The seven regression models are random forest (RF), regression tree (RT), extremely randomized trees (ERT), linear regression (LR), lasso regression (Lasso), ridge regression (Ridge) and neural network (NN). All models except for LR were tuned.

**Table C4**

*Quadratic weighted kappa, Pearson's r and confidence interval of the regression models fit on DS3 of the Hewlett essay datasets*

| Model | QWK | Pearson's r | 95% CI | | p value |
|---|---|---|---|---|---|
| | | | LL | UL | |
| LR | .64 | .69 | .63 | .74 | <.0001 |
| RF | .61 | .69 | .63 | .74 | <.0001 |
| RT | .58 | .68 | .62 | .73 | <.0001 |
| ERT | .53 | .53 | .45 | .60 | <.0001 |
| NN | .67 | .68 | .62 | .73 | <.0001 |
| Ridge | .63 | .69 | .63 | .74 | <.0001 |
| Lasso | .59 | .69 | .63 | .74 | <.0001 |

*Note.* The seven regression models are random forest (RF), regression tree (RT), extremely randomized trees (ERT), linear regression (LR), lasso regression (Lasso), ridge regression (Ridge) and neural network (NN). All models except for LR were tuned.

**Table C5**

*Quadratic weighted kappa, Pearson's r and confidence interval of the regression models fit on DS4 of the Hewlett essay datasets*

| Model | QWK | Pearson's r | 95% CI | | p value |
|---|---|---|---|---|---|
| | | | LL | UL | |
| LR | .69 | .77 | .73 | .81 | <.0001 |
| RF | .56 | .76 | .71 | .80 | <.0001 |
| RT | .65 | .74 | .69 | .78 | <.0001 |
| ERT | .57 | .57 | .49 | .64 | <.0001 |
| NN | .73 | .77 | .73 | .81 | <.0001 |
| Ridge | .65 | .77 | .72 | .81 | <.0001 |
| Lasso | .68 | .76 | .71 | .80 | <.0001 |

*Note*. The seven regression models are random forest (RF), regression tree (RT), extremely randomized trees (ERT), linear regression (LR), lasso regression (Lasso), ridge regression (Ridge) and neural network (NN). All models except for LR were tuned.

**Table C6**

*Quadratic weighted kappa, Pearson's r and confidence interval of the regression models fit on DS5 of the Hewlett essay datasets*

| Model | QWK | Pearson's r | 95% CI | | p value |
|---|---|---|---|---|---|
| | | | LL | UL | |
| LR | .79 | .83 | .79 | .86 | <.0001 |
| RF | .72 | .81 | .77 | .84 | <.0001 |
| RT | .75 | .79 | .75 | .83 | <.0001 |
| ERT | .72 | .73 | .67 | .77 | <.0001 |
| NN | .80 | .83 | .80 | .86 | <.0001 |
| Ridge | .79 | .83 | .80 | .86 | <.0001 |
| Lasso | .80 | .83 | .79 | .86 | <.0001 |

*Note.* The seven regression models are random forest (RF), regression tree (RT), extremely randomized trees (ERT), linear regression (LR), lasso regression (Lasso), ridge regression (Ridge) and neural network (NN). All models except for LR were tuned.

**Table C7**

*Quadratic weighted kappa, Pearson's r and confidence interval of the regression models fit on DS6 of the Hewlett essay datasets*

| Model | QWK | Pearson's r | 95% CI | | p value |
|---|---|---|---|---|---|
| | | | LL | UL | |
| LR | .74 | .78 | .73 | .81 | <.0001 |
| RF | .65 | .75 | .70 | .79 | <.0001 |
| RT | .63 | .70 | .64 | .75 | <.0001 |
| ERT | .63 | .63 | .57 | .69 | <.0001 |
| NN | .75 | .80 | .75 | .83 | <.0001 |
| Ridge | .70 | .79 | .74 | .82 | <.0001 |
| Lasso | .73 | .80 | .76 | .83 | <.0001 |

*Note.* The seven regression models are random forest (RF), regression tree (RT), extremely randomized trees (ERT), linear regression (LR), lasso regression (Lasso), ridge regression (Ridge) and neural network (NN). All models except for LR were tuned.

**Table C8**

*Quadratic weighted kappa, Pearson's r and confidence interval of the regression models fit on DS7 of the Hewlett essay datasets*

| Model | QWK | Pearson's r | 95% CI | | p value |
|---|---|---|---|---|---|
| | | | LL | UL | |
| LR | .76 | .76 | .71 | .80 | <.0001 |
| RF | .75 | .78 | .73 | .82 | <.0001 |
| RT | .69 | .72 | .67 | .77 | <.0001 |
| ERT | .60 | .60 | .52 | .67 | <.0001 |
| NN | .82 | .82 | .78 | .85 | <.0001 |
| Ridge | .81 | .81 | .77 | .85 | <.0001 |
| Lasso | .80 | .82 | .78 | .85 | <.0001 |

*Note.* The seven regression models are random forest (RF), regression tree (RT), extremely randomized trees (ERT), linear regression (LR), lasso regression (Lasso), ridge regression (Ridge) and neural network (NN). All models except for LR were tuned.

**Table C9**

*Quadratic weighted kappa, Pearson's r and confidence interval of the regression models fit on DS8 of the Hewlett essay datasets*

| Model | QWK | Pearson's r | 95% CI | | p value |
|---|---|---|---|---|---|
| | | | LL | UL | |
| LR | .58 | .59 | .47 | .68 | <.0001 |
| RF | .60 | .70 | .60 | .77 | <.0001 |
| RT | .53 | .58 | .47 | .68 | <.0001 |
| ERT | .49 | .49 | .36 | .61 | <.0001 |
| NN | .72 | .72 | .63 | .79 | <.0001 |
| Ridge | .71 | .71 | .62 | .78 | <.0001 |
| Lasso | .71 | .72 | .64 | .79 | <.0001 |

*Note.* The seven regression models are random forest (RF), regression tree (RT), extremely randomized trees (ERT), linear regression (LR), lasso regression (Lasso), ridge regression (Ridge) and neural network (NN). All models except for LR were tuned.

**Figure C1**

*Scatter plots of predicted grades on the x axis and true grades on the y-axis for the DS1 of the Hewlett datasets*

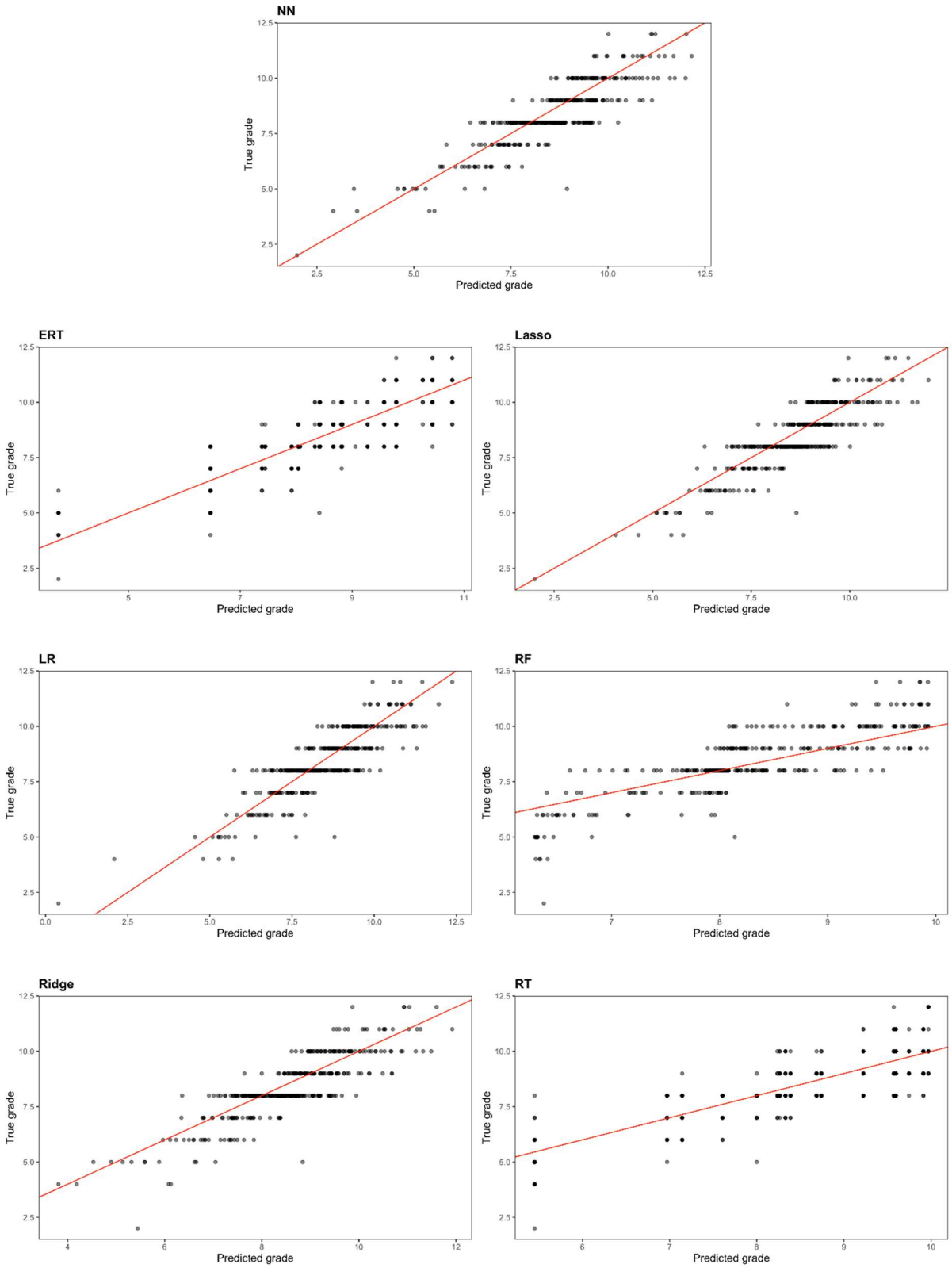


**Figure C2**

*Scatter plots of predicted grades on the x axis and true grades on the y- axis for the DS2A of the Hewlett datasets*

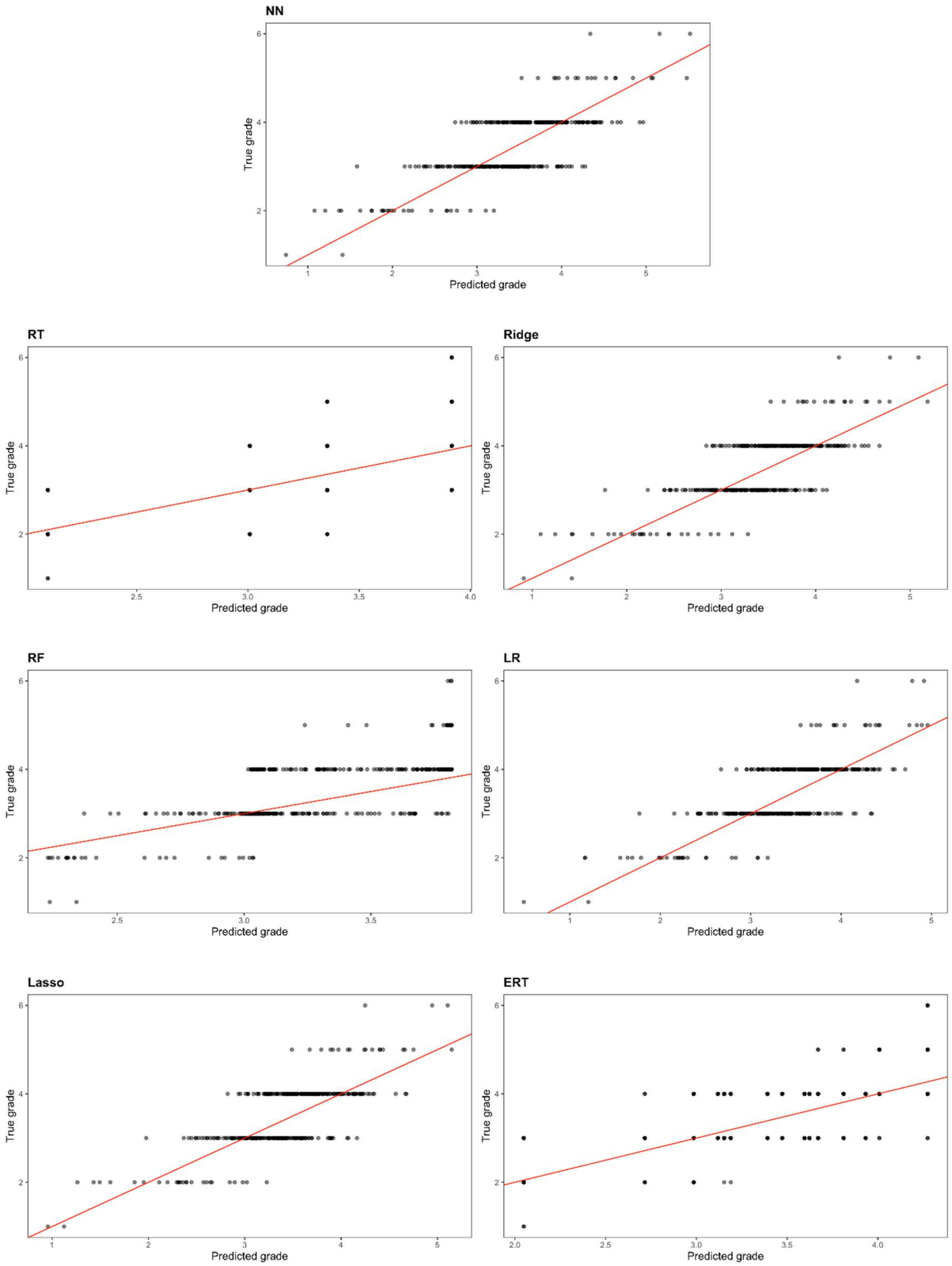


**Figure C3**

*Scatter plots of predicted grades on the x axis and true grades on the y-axis axis for the DS2B of the Hewlett datasets*

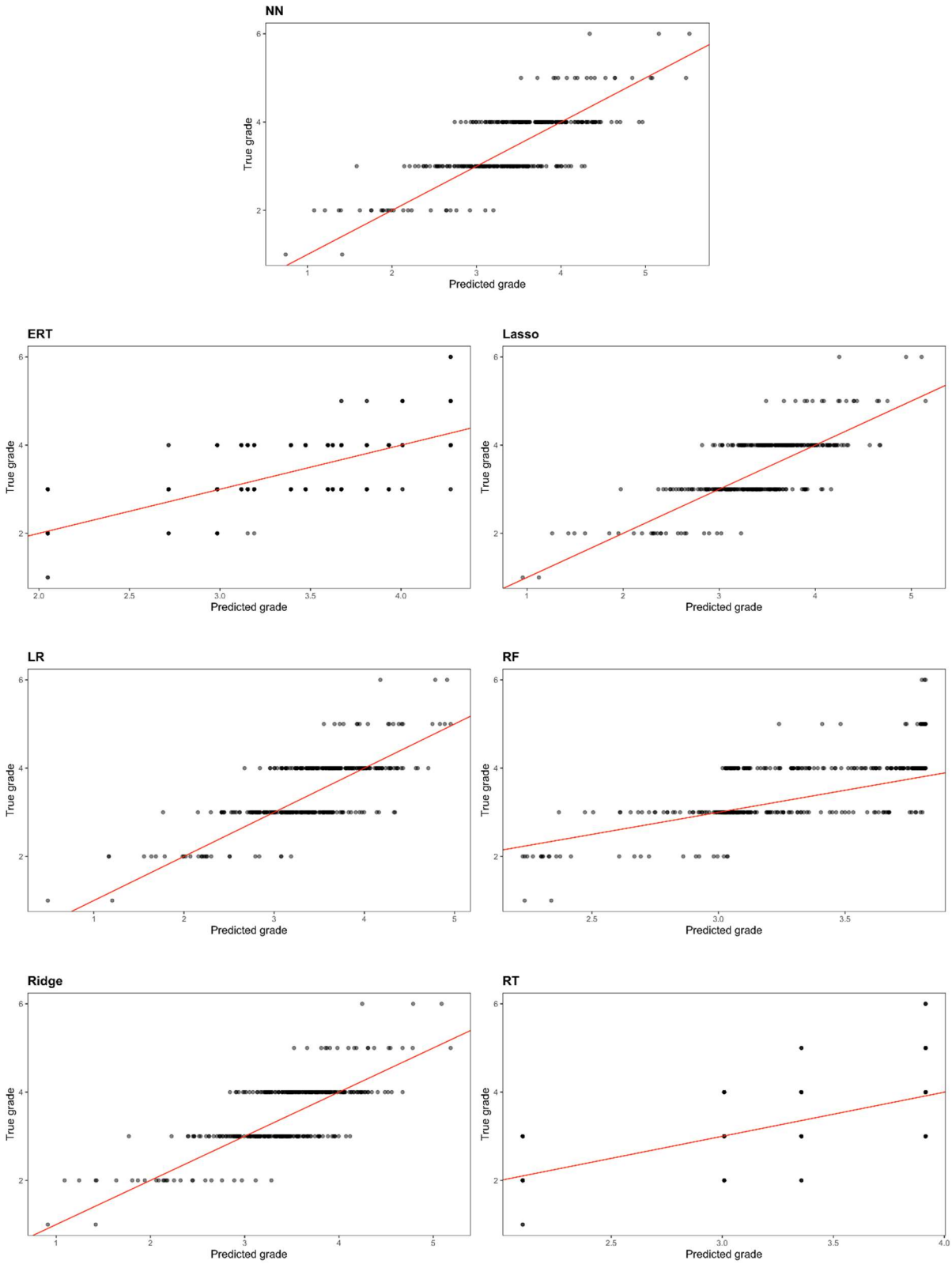


**Figure C4**

*Scatter plots of predicted grades on the x axis and true grades on the y-axis axis for the DS3 of the Hewlett datasets*

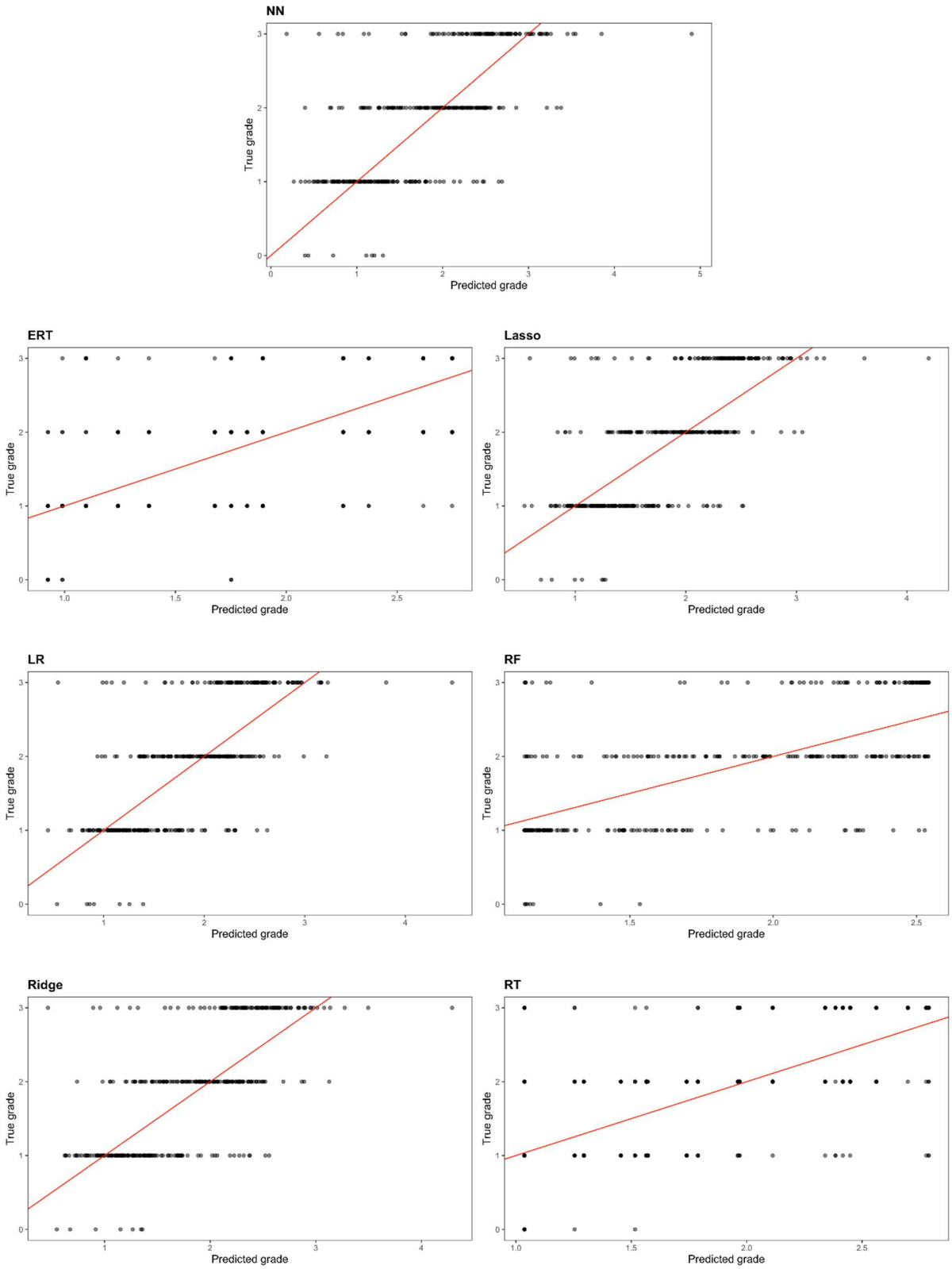


**Figure C5**

*Scatter plots of predicted grades on the x axis and true grades on the y-axis for the DS4 of the Hewlett datasets*

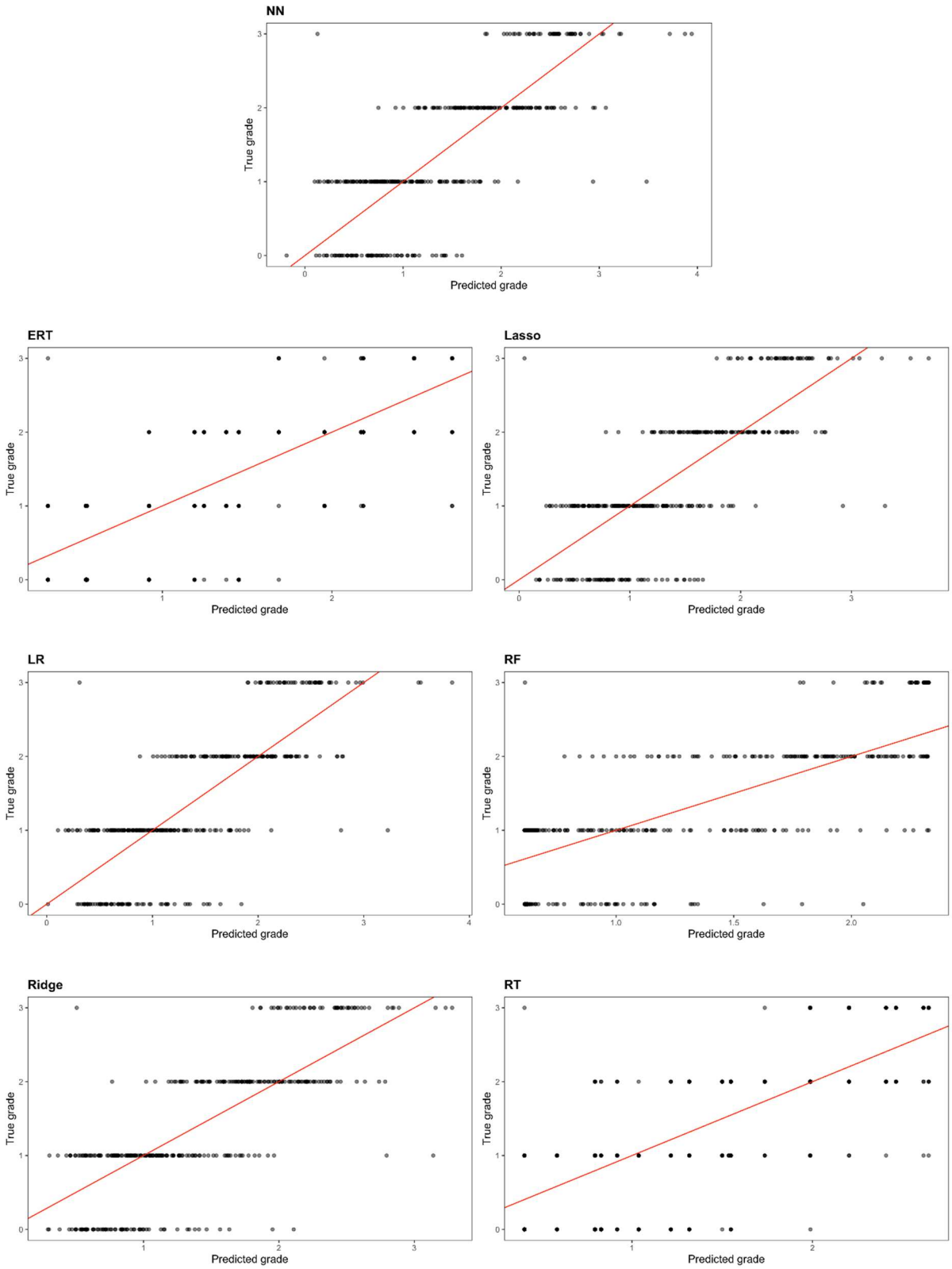


**Figure C6**

*Scatter plots of predicted grades on the x axis and true grades on the y-axis for the DS5 of the Hewlett datasets*

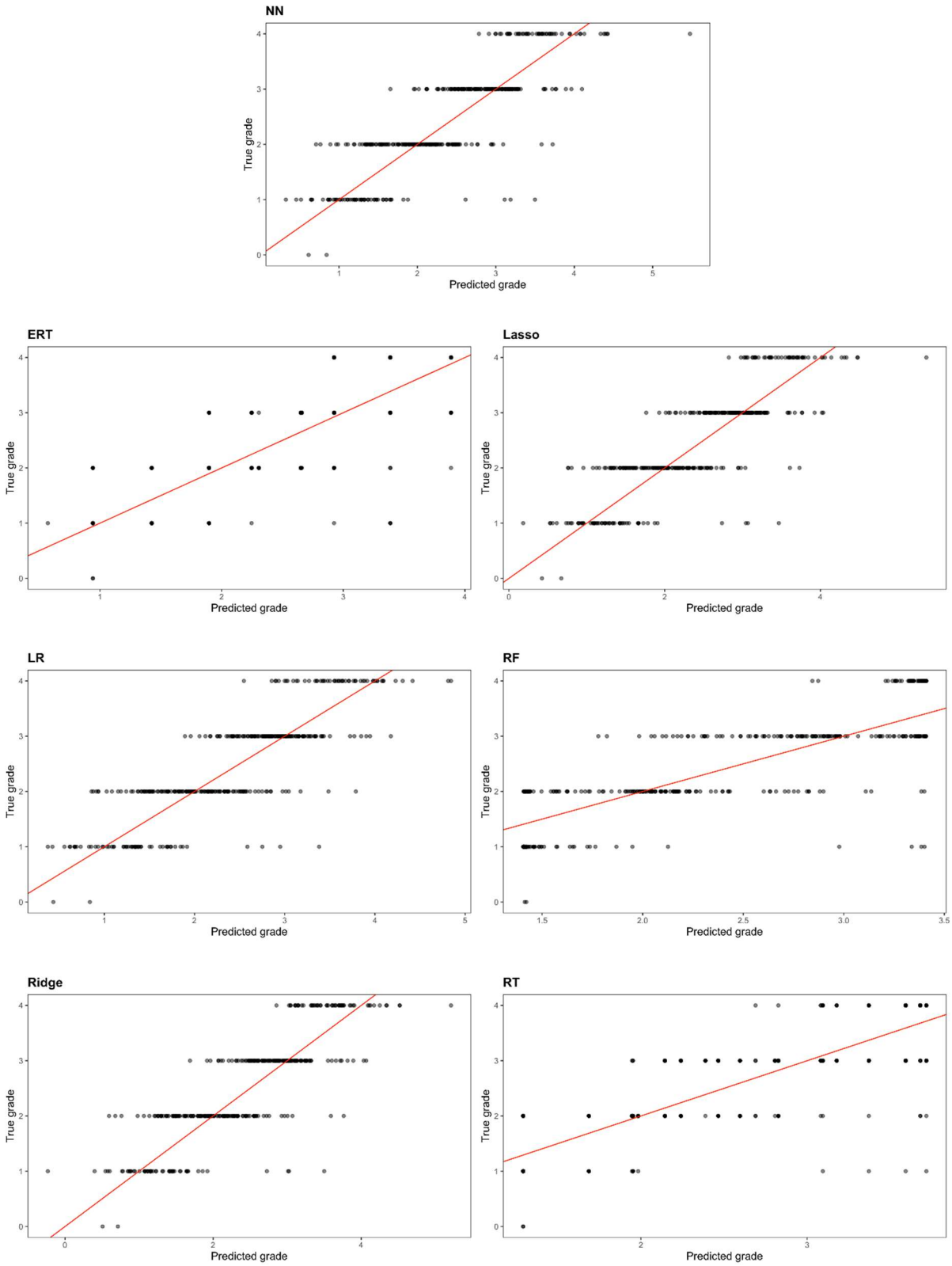


**Figure C7**

*Scatter plots of predicted grades on the x axis and true grades on the y-axis for the DS6 of the Hewlett datasets*

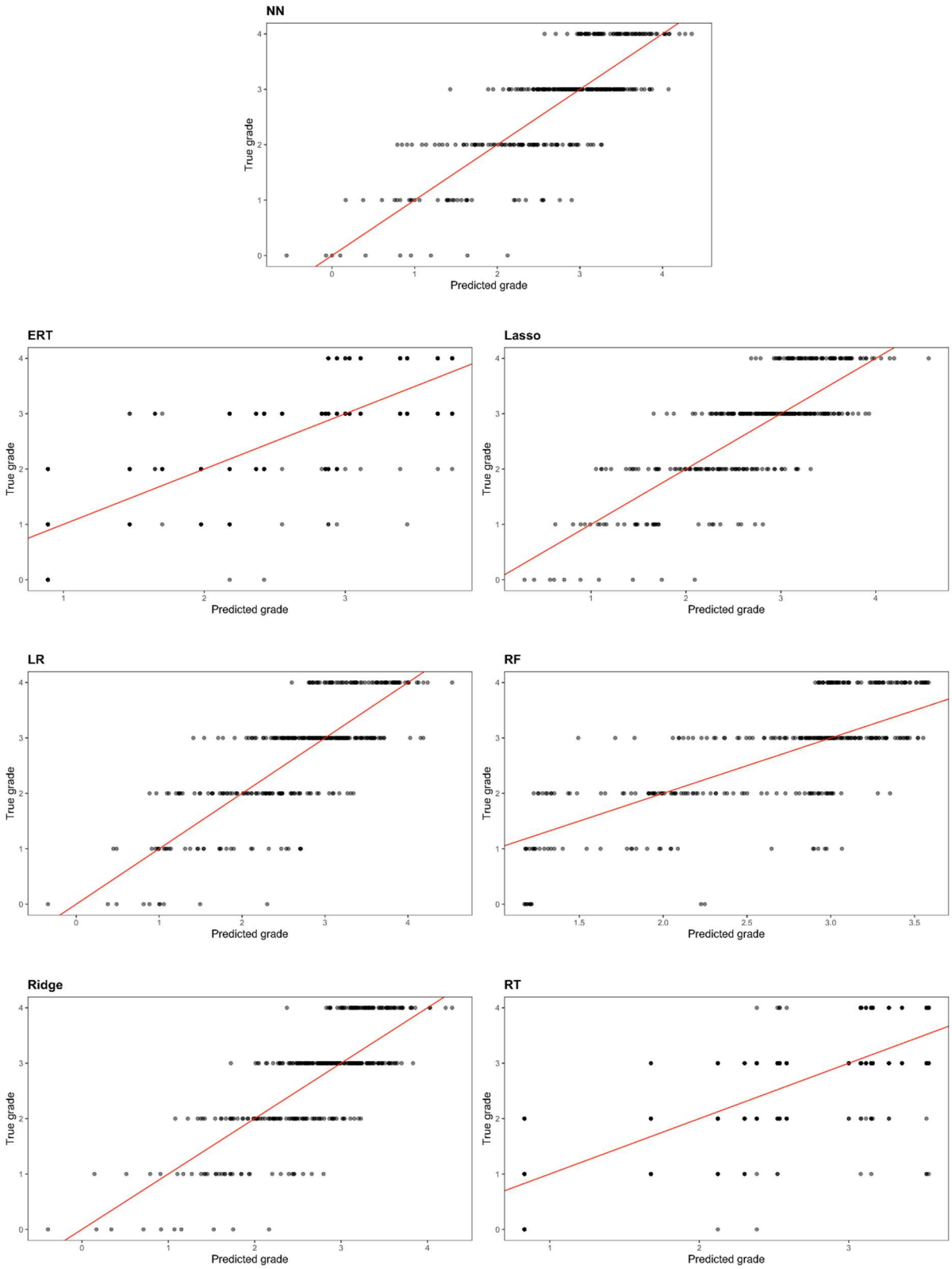


**Figure C8**

*Scatter plots of predicted grades on the x axis and true grades on the y-axis for the DS7 of the Hewlett datasets*

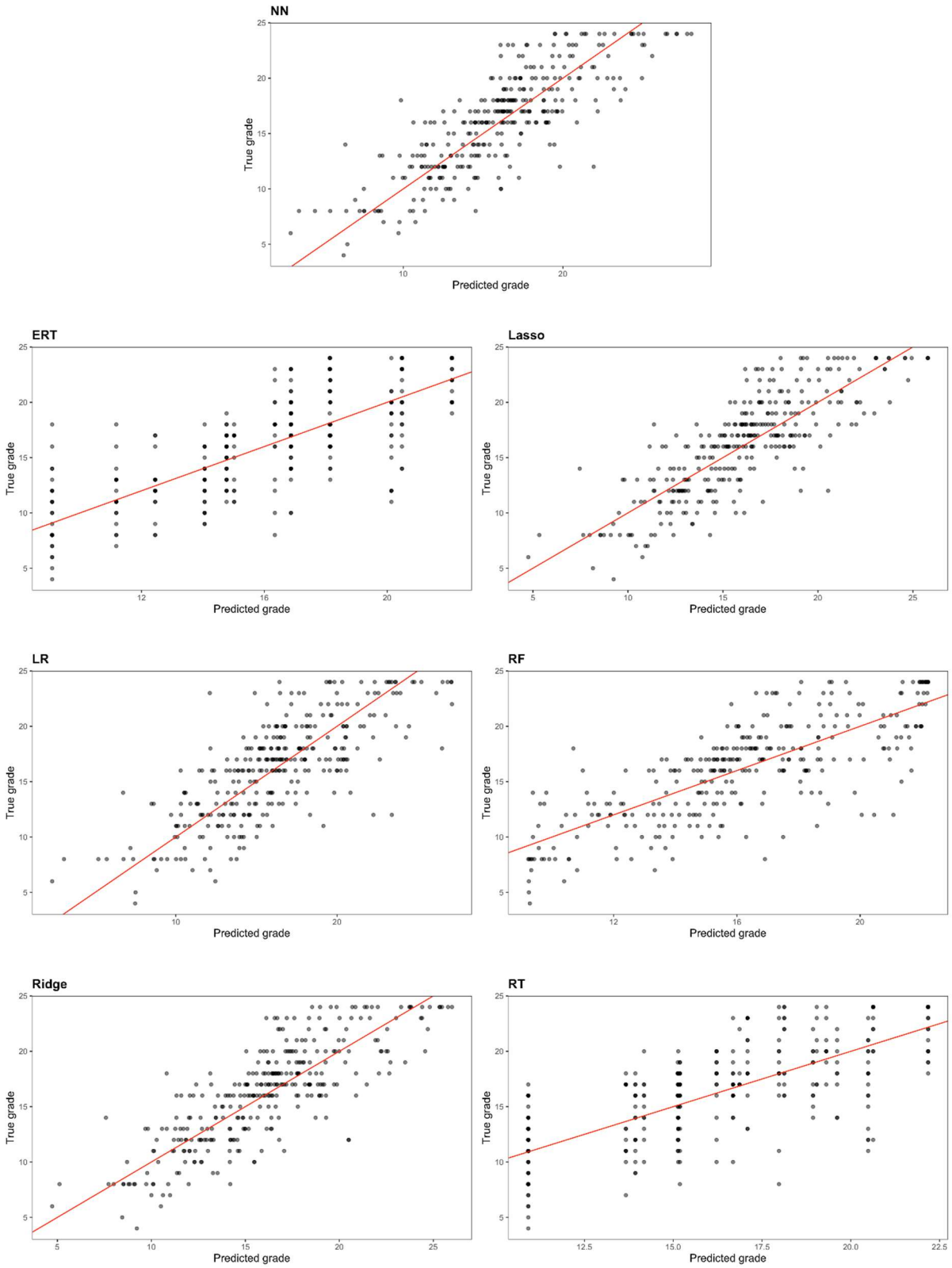
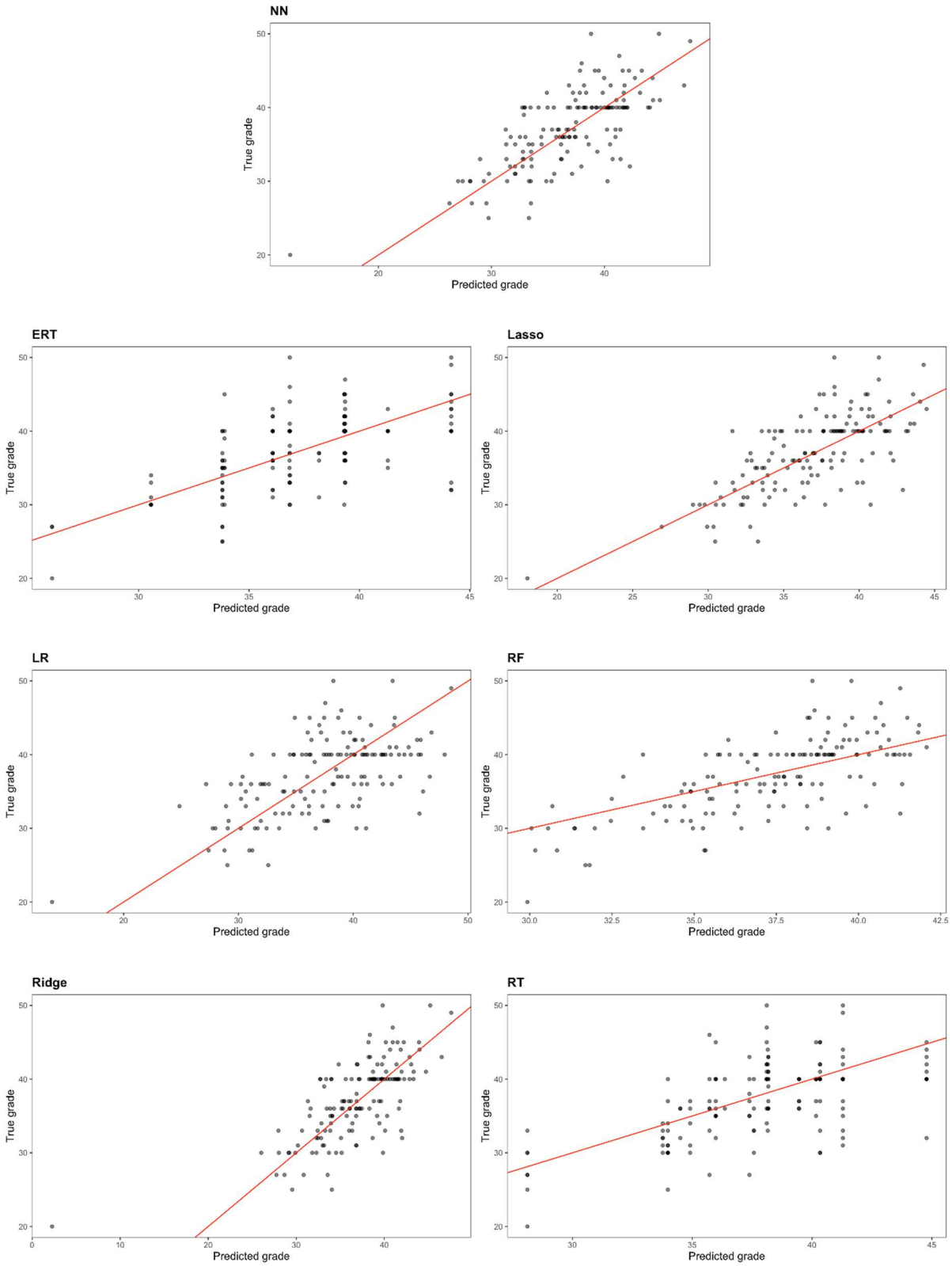


**Figure C9**

*Scatter plots of predicted grades on the x axis and true grades on the y-axis for the DS8 of the Hewlett datasets*



# Appendix D

**Table D1**

*Features selected by the lasso regression for the thesis data*

| Number | Features | Coefficients |
|---|---|---|
| 1 | nr_capitalization_errors | 0,146539792 |
| 2 | CCONJ | 0,085497253 |
| 3 | dale_chall_readability | 0,056264512 |
| 4 | eucl_centroid_min_max | 0,048200883 |
| 5 | avg_tree_height | 0,047796059 |
| 6 | nr_unique_pos | 0,039043643 |
| 7 | guirauds_index | 0,037881885 |
| 8 | nominal_ratio | 0,034748893 |
| 9 | ADJ | 0,029139645 |
| 10 | *avg_cosine_similariy_high_grade* | 0,0290356 |
| 11 | grade_as_feature | 0,027999893 |
| 12 | verb_baseform | 0,027548252 |
| 13 | min_neighbouring_eucl_dist | 0,025712762 |
| 14 | pattern_cosine | 0,025101326 |
| 15 | weighted_cosine | 0,024787387 |
| 16 | nr_spelling_errors | 0,024784088 |
| 17 | nr_grammar_errors | 0,020836769 |
| 18 | most_freq_sentence_length | 0,010603601 |
| 19 | eucl_relative_distance | 0,009843298 |
| 20 | avg_sentence_len_words | 0,005362628 |
| 21 | DET | 0,005253732 |
| 22 | max_cos_centroid_similarities | 0,002845667 |
| 23 | max_neighbouring_cos_similarity | 1,06994E-06 |
| 24 | min_neighbouring_cos_similarity | 0 |
| 25 | avg_eucl_dist | 0 |
| 26 | cos_min_max_quotient | 0 |
| **27** | **eucl_min_max_quotient** | **0** |
| 28 | max_eucl_dist | 0 |
| 29 | max_neighbouring_eucl_dist | 0 |
| 30 | avg_neighbouring_eucl_dist | 0 |
| 31 | avg_neighbouring_cos_similarity | 0 |
| 32 | nr_punctuation_errors | 0 |
| 33 | NOUN | 0 |
| 34 | PROPN | 0 |
| 35 | PRON | 0 |
| 36 | ADV | 0 |
| 37 | avg_cos_similarity | 0 |
| 38 | standard_distance | 0 |
| 39 | max_cos_similarity | 0 |
| 40 | eucl_clark_evan_dist_nn | 0 |
| 41 | avg_eucl_dist_nn | 0 |

| | | |
|---|---|---|
| 42 | avg_eucl_cumulative_freq_distribution | 0 |
| 43 | avg_eucl_centroid_distances | 0 |
| **44** | **avg_cos_centroid_similarities** | **0** |
| 45 | min_eucl_centroid_distances | 0 |
| **46** | **min_cos_centroid_similarities** | **0** |
| 47 | max_eucl_centroid_distances | 0 |
| 48 | cos_centroid_min_max | 0 |
| 49 | NUM | 0 |
| 50 | det_eucl_dist_matrix | 0 |
| 51 | morans_i | 0 |
| 52 | gearys_c | 0 |
| 53 | ADP | 0 |
| 54 | yules_k | 0 |
| 55 | SCONJ | 0 |
| 56 | flesh_kincaid_grade_level | 0 |
| 57 | nr_stopwords | 0 |
| **58** | **unique_words** | **0** |
| **59** | **word_count** | **0** |
| 60 | char_count | 0 |
| 61 | avg_word_len | 0 |
| 62 | nr_long_words | 0 |
| 63 | nr_short_words | 0 |
| 64 | most_freq_word_length | 0 |
| 65 | nr_sentences | 0 |
| 66 | nr_long_sentences | 0 |
| 67 | nr_short_sentences | 0 |
| 68 | gunning_fox_index | 0 |
| 69 | flesh_reading_ease | 0 |
| 70 | automated_readability_index | 0 |
| 71 | verb_past_tense | 0 |
| 72 | SMOG | 0 |
| 73 | LIX | 0 |
| 74 | OVIX | 0 |
| 75 | TTR | 0 |
| 76 | d_estimate | 0 |
| 77 | hapax_legomena | 0 |
| 78 | advanced_guiraud | 0 |
| 79 | correct_verb_form | 0 |
| 80 | comparative_adj | 0 |
| 81 | superlative_adj | 0 |
| 82 | modal_aux | 0 |
| 83 | participle | 0 |
| 84 | infinitive_marker | 0 |
| 85 | gettis_g | 0 |

*Note.* The features not selected for thesis and essay texts are marked in bold and the feature only selected for thesis texts is italicised.

**Table D2**

*Features selected by the lasso regression for all Hewlett datasets combined*

| Number | Features | Summed coefficients |
|---|---|---|
| 1 | guirauds_index | 4,27491251 |
| 2 | nr_sentences | 1,801038638 |
| 3 | nr_stopwords | 1,289374318 |
| 4 | standard_distance | 1,26967989 |
| 5 | avg_word_len | 1,227555893 |
| 6 | verb_past_tense | 1,198818608 |
| 7 | gettis_g | 1,168258596 |
| 8 | gunning_fox_index | 1,111274797 |
| 9 | avg_sentence_len_words | 1,029604883 |
| 10 | nr_spelling_errors | 0,813276916 |
| 11 | grade_as_feature | 0,789021182 |
| 12 | nr_capitalization_errors | 0,603298325 |
| 13 | verb_baseform | 0,600464101 |
| 14 | DET | 0,597007358 |
| 15 | infinitive_marker | 0,562268628 |
| 16 | avg_eucl_centroid_distances | 0,555576801 |
| 17 | avg_eucl_dist | 0,536371861 |
| 18 | char_count | 0,506025612 |
| 19 | avg_neighbouring_cos_similarity | 0,504623462 |
| 20 | PROPN | 0,474391917 |
| 21 | participle | 0,468404565 |
| 22 | flesh_reading_ease | 0,441235986 |
| 23 | eucl_centroid_min_max | 0,407464245 |
| 24 | hapax_legomena | 0,405687552 |
| 25 | flesh_kincaid_grade_level | 0,377963066 |
| 26 | nr_grammar_errors | 0,346476135 |
| 27 | SCONJ | 0,286045009 |
| 28 | CCONJ | 0,28162607 |
| 29 | SMOG | 0,277519387 |
| 30 | OVIX | 0,276805092 |
| 31 | ADJ | 0,274281574 |
| 32 | NUM | 0,259473234 |
| 33 | pattern_cosine | 0,259046576 |
| 34 | d_estimate | 0,245194641 |
| 35 | nr_long_sentences | 0,236815069 |
| 36 | correct_verb_form | 0,235058537 |
| 37 | ADV | 0,230374952 |
| 38 | max_eucl_dist | 0,219697164 |
| 39 | avg_eucl_dist_nn | 0,21147874 |
| 40 | yules_k | 0,208879784 |
| 41 | most_freq_word_length | 0,194798276 |
| 42 | modal_aux | 0,190760128 |
| 43 | comparative_adj | 0,18225004 |

| | | |
|---|---|---|
| 44 | avg_tree_height | 0,181730792 |
| 45 | TTR | 0,179472422 |
| 46 | morans_i | 0,17910827 |
| 47 | nr_long_words | 0,176215194 |
| 48 | nr_punctuation_errors | 0,176101658 |
| 49 | NOUN | 0,17037328 |
| 50 | min_eucl_centroid_distances | 0,16976461 |
| 51 | max_eucl_centroid_distances | 0,162555136 |
| 52 | nr_short_words | 0,161936233 |
| 53 | det_eucl_dist_matrix | 0,160842963 |
| 54 | advanced_guiraud | 0,145705322 |
| 55 | nr_short_sentences | 0,144936606 |
| 56 | nr_unique_pos | 0,143503157 |
| 57 | LIX | 0,132421594 |
| 58 | avg_neighbouring_eucl_dist | 0,114197858 |
| 59 | eucl_clark_evan_dist_nn | 0,108080253 |
| 60 | max_neighbouring_eucl_dist | 0,101902744 |
| 61 | dale_chall_readability | 0,099258035 |
| 62 | superlative_adj | 0,098628366 |
| 63 | min_neighbouring_eucl_dist | 0,091996901 |
| 64 | PRON | 0,081856073 |
| 65 | nominal_ratio | 0,069213956 |
| 66 | gearys_c | 0,063445368 |
| 67 | automated_readability_index | 0,061274648 |
| 68 | eucl_relative_distance | 0,060735259 |
| 69 | most_freq_sentence_length | 0,055102469 |
| 70 | weighted_cosine | 0,052888476 |
| 71 | avg_cos_similarity | 0,051105709 |
| 72 | avg_eucl_cumulative_freq_distribution | 0,045724611 |
| 73 | max_cos_similarity | 0,040858102 |
| 74 | ADP | 0,039593637 |
| 75 | cos_centroid_min_max | 0,038803296 |
| 76 | max_neighbouring_cos_similarity | 0,032742191 |
| 77 | cos_min_max_quotient | 0,026330965 |
| 78 | max_cos_centroid_similarities | 0,0162227 |
| 79 | min_neighbouring_cos_similarity | 0,002292817 |
| **80** | **unique_words** | **0** |
| **81** | **word_count** | **0** |
| 82 | avg_cosine_similariy_high_grade | 0 |
| **83** | **eucl_min_max_quotient** | **0** |
| **84** | **avg_cos_centroid_similarities** | **0** |
| **85** | **min_cos_centroid_similarities** | **0** |

*Note*. Summed coefficients of the Hewlett datasets. The features not selected for thesis and essay texts are marked in bold.

**Figure D1**

*Coefficients of the lasso regression for the thesis data and DS1 of the Hewlett datasets*

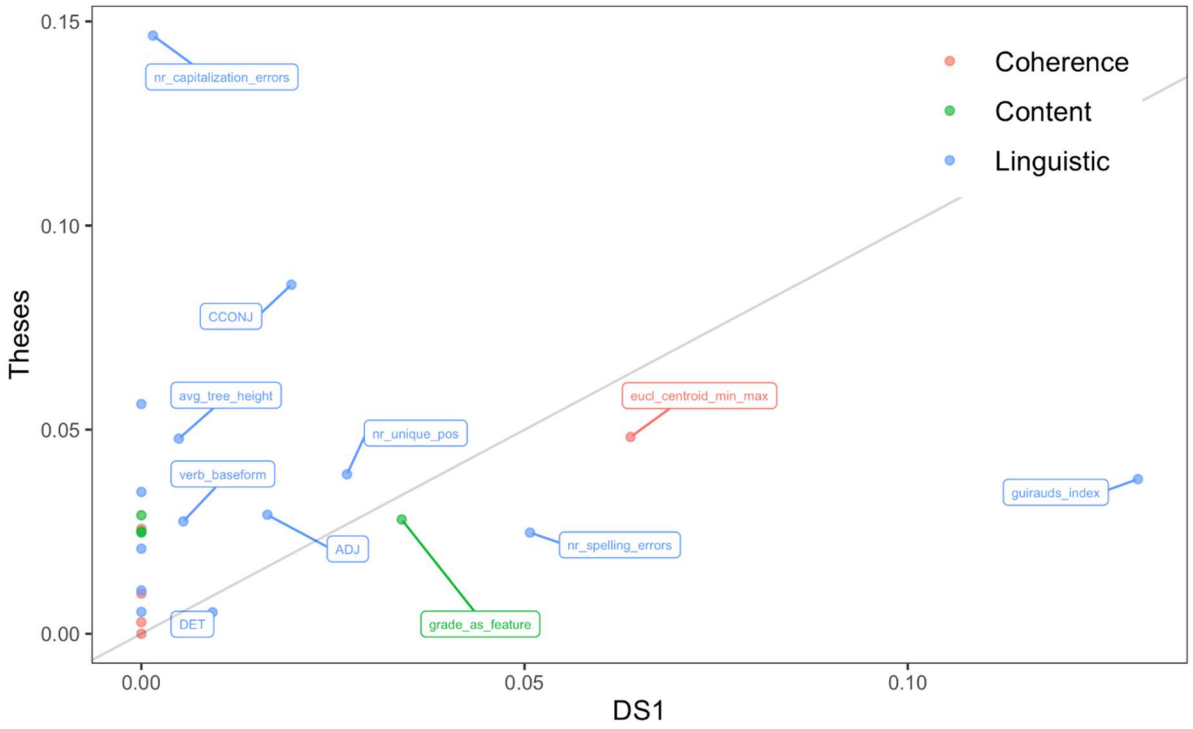



**Figure D2**

*Coefficients of the lasso regression for the thesis data and DS2A of the Hewlett datasets*

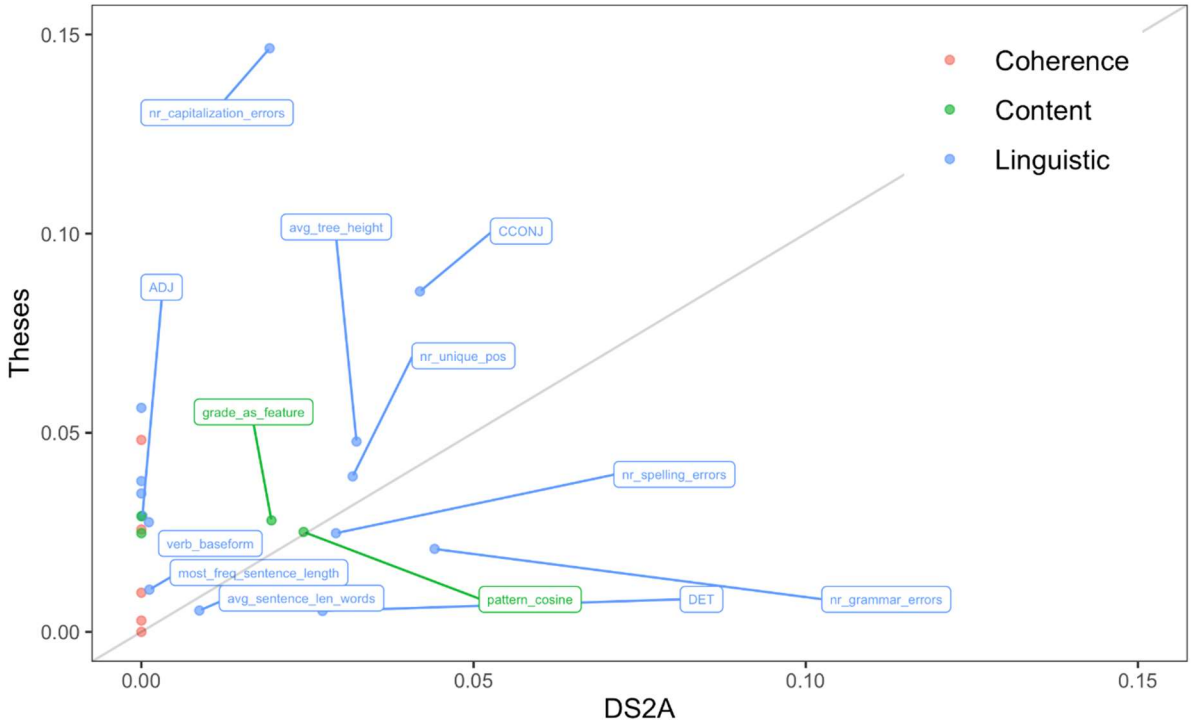


**Figure D3**

*Coefficients of the lasso regression for the thesis data and DS2B of the Hewlett datasets*

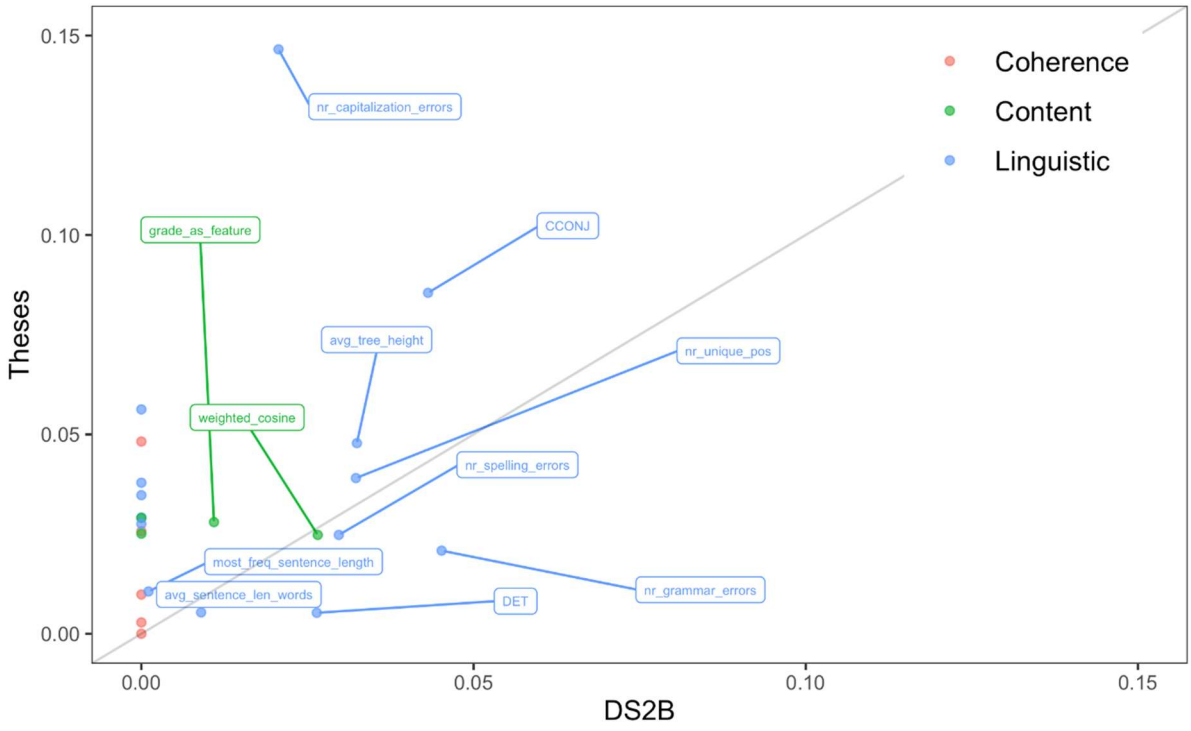



**Figure D4**

*Coefficients of the lasso regression for the thesis data and DS3 of the Hewlett datasets*

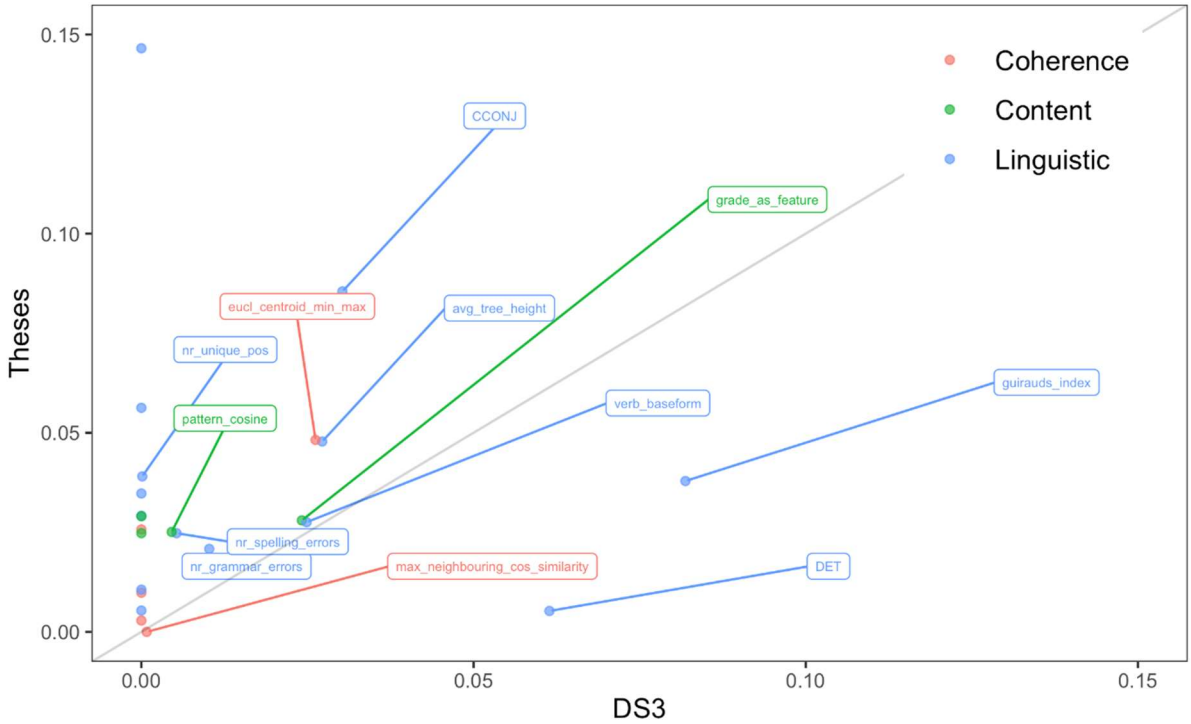


**Figure D5**

*Coefficients of the lasso regression for the thesis data and DS4 of the Hewlett datasets*

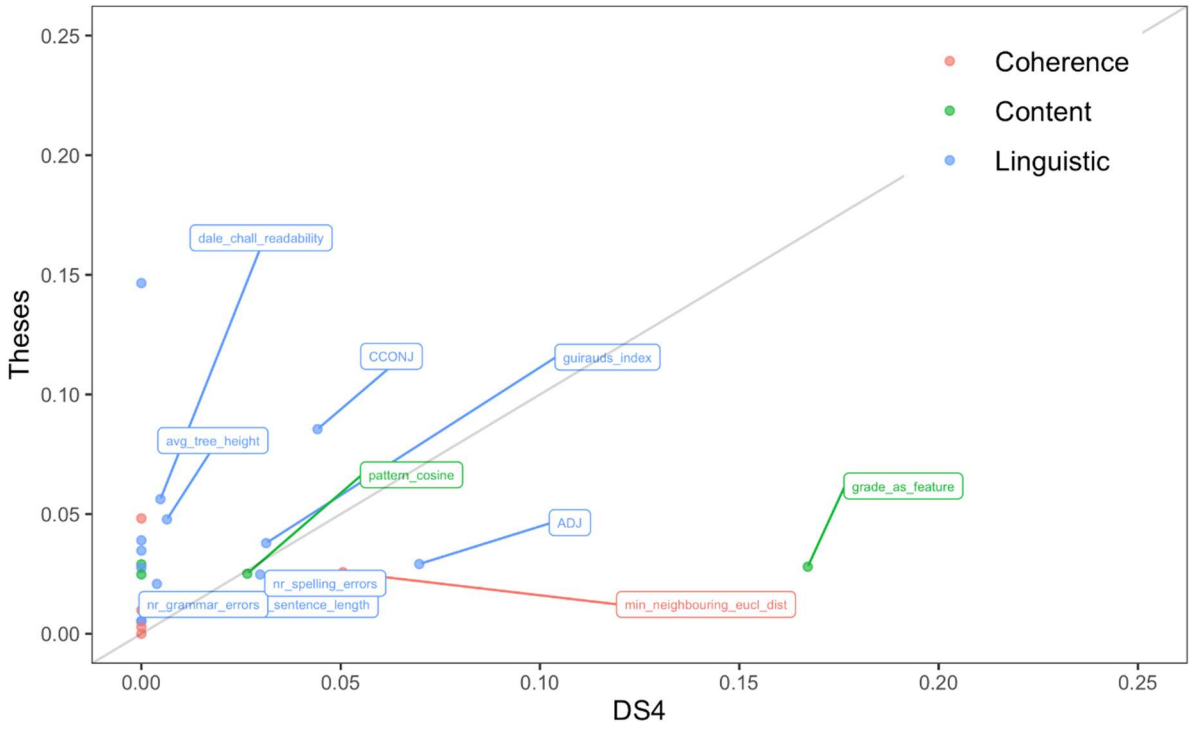



**Figure D6**

*Coefficients of the lasso regression for the thesis data and DS5 of the Hewlett datasets*

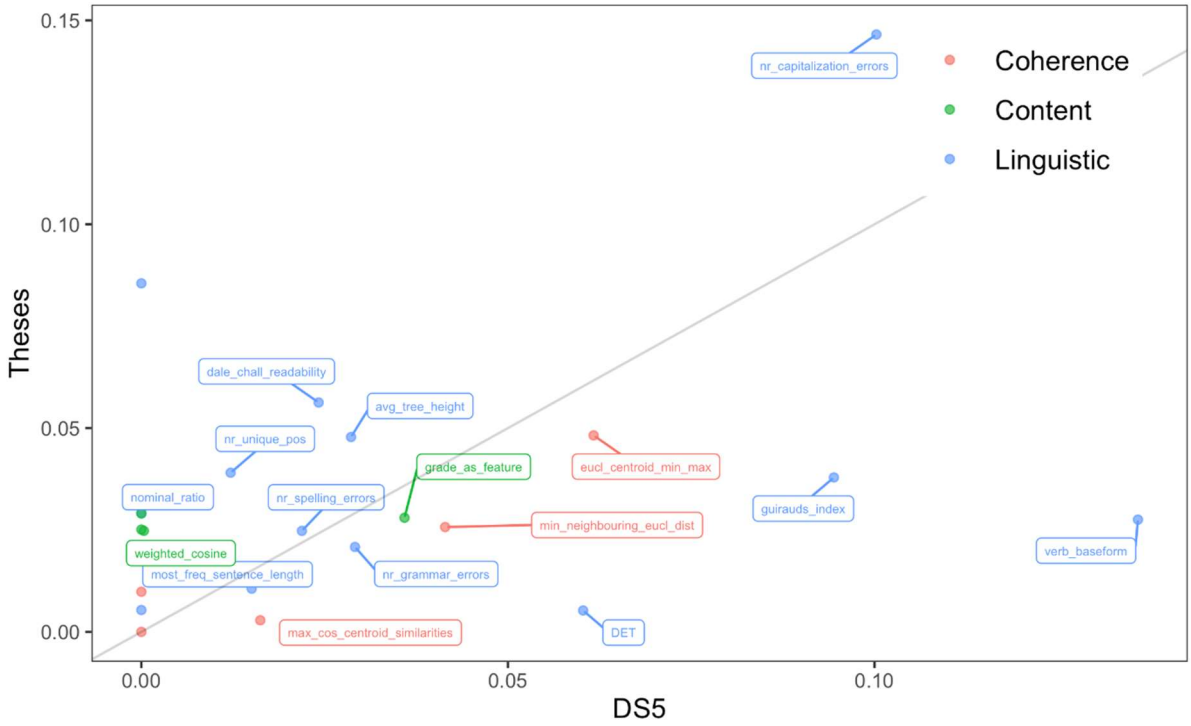


**Figure D7**

*Coefficients of the lasso regression for the thesis data and DS6 of the Hewlett datasets*

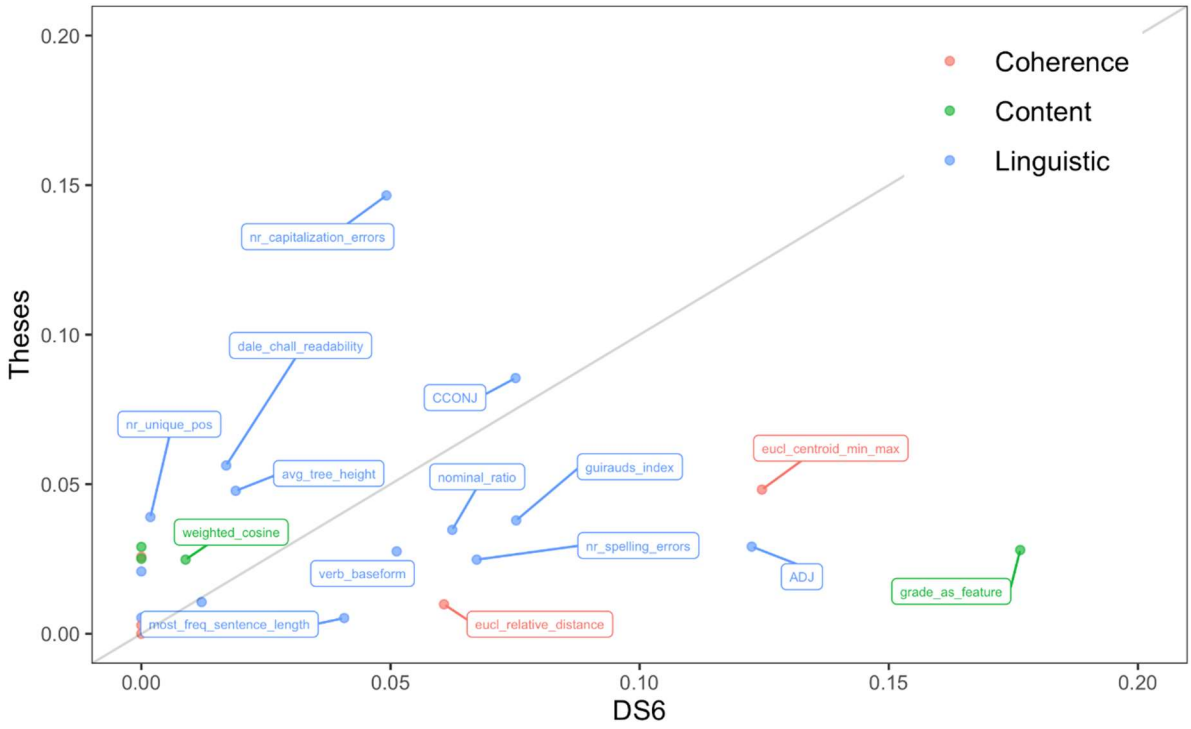



**Figure D8**

*Coefficients of the lasso regression for the thesis data and DS7 of the Hewlett datasets*

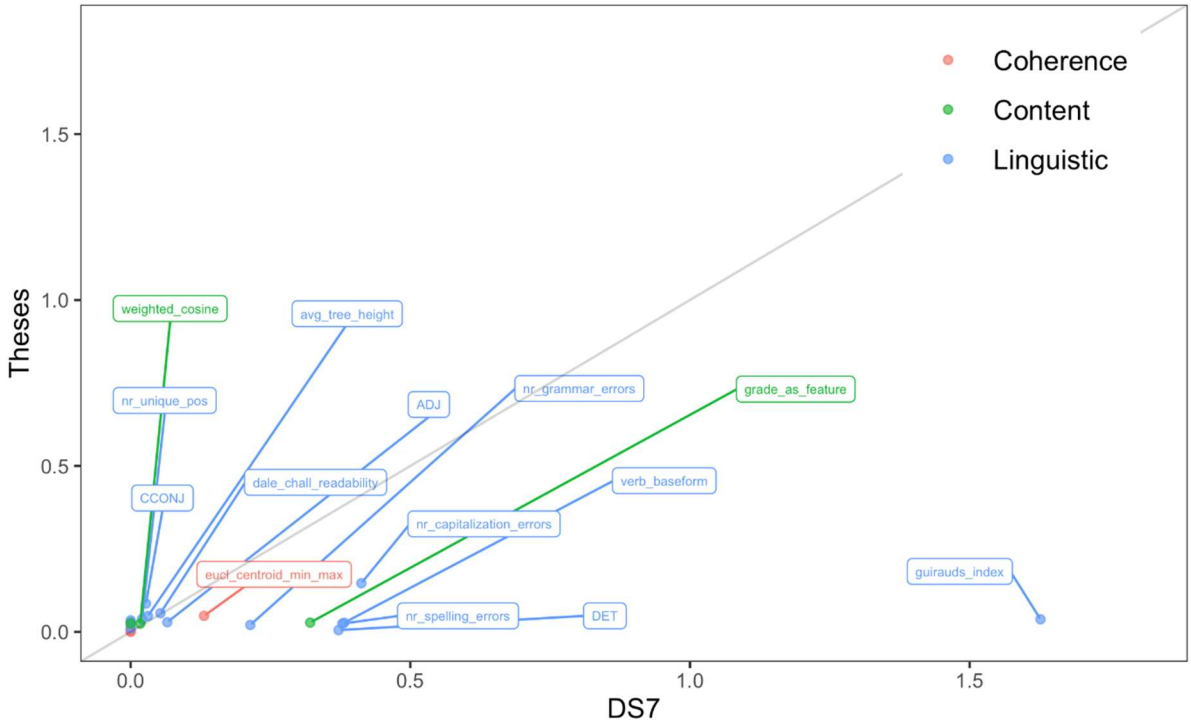


**Figure D9**

*Coefficients of the lasso regression for the thesis data and DS8 of the Hewlett datasets*