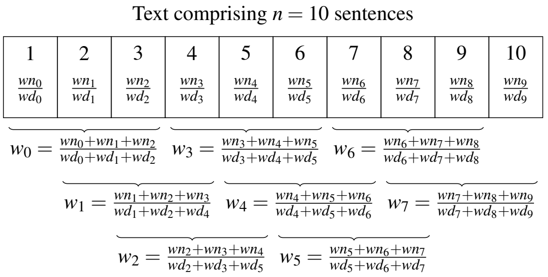
Appendix

```
Algorithm 1: Feature ablation algorithm
  Input: $N$ training instances with feature set
         $F = \{f_1, \dots, f_D\}$
  Input: $m$ features to remove at each step
  Result: $list$ containing the feature
          importance rank order
begin
    $t \leftarrow 0$
    $list \leftarrow []$
    while $|F| > 0$ do
        Train a classifier with $|F|$ input
          features;
        Compute $S_{i,t}, i \in F$;
        Find $f_{i_1}, \dots, f_{i_m}$, where
          $S_{i_1,t}, \dots, S_{i_m,t}$ are m largest
          among all $S_{i,t} (i \in F)$ in
          descending order;
        $list \leftarrow list.append([f_{i_1}, \dots, f_{i_m}])$;
        $F \leftarrow F - \{f_{i_1}, \dots, f_{i_m}\}$;
        $t \leftarrow t + 1$;
    return $list$
```



Text comprising $n = 10$ sentences

1	2	3	4	5	6	7	8	9	10
$\frac{wn_0}{wd_0}$	$\frac{wn_1}{wd_1}$	$\frac{wn_2}{wd_2}$	$\frac{wn_3}{wd_3}$	$\frac{wn_4}{wd_4}$	$\frac{wn_5}{wd_5}$	$\frac{wn_6}{wd_6}$	$\frac{wn_7}{wd_7}$	$\frac{wn_8}{wd_8}$	$\frac{wn_9}{wd_9}$

$w_0 = \frac{wn_0+wn_1+wn_2}{wd_0+wd_1+wd_2}$ $w_3 = \frac{wn_3+wn_4+wn_5}{wd_3+wd_4+wd_5}$ $w_6 = \frac{wn_6+wn_7+wn_8}{wd_6+wd_7+wd_8}$

$w_1 = \frac{wn_1+wn_2+wn_3}{wd_1+wd_2+wd_4}$ $w_4 = \frac{wn_4+wn_5+wn_6}{wd_4+wd_5+wd_6}$ $w_7 = \frac{wn_7+wn_8+wn_9}{wd_7+wd_8+wd_9}$

$w_2 = \frac{wn_2+wn_3+wn_4}{wd_2+wd_3+wd_5}$ $w_5 = \frac{wn_5+wn_6+wn_7}{wd_5+wd_6+wd_7}$

Figure 3: Schematic illustration of how complexity measurements are obtained in CoCoGen for a text comprising ten sentences with a window size of three sentences.

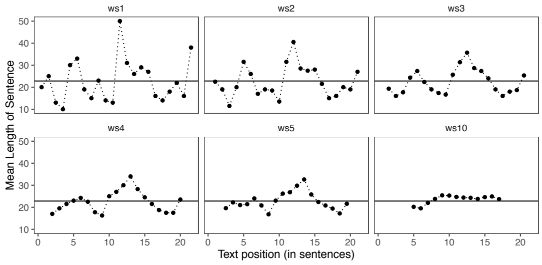


Figure 4: Complexity contours of the same text with window size settings ranging between 1-10. A window size of one yields a sentence-by-sentence assessment of text complexity. With increasing window size the contours approaches the text average complexity value, represented here by the horizontal line.

	Baseline Model	
	M	SD
Accuracy	0.28	0.03
Precision 2	0.26	0.04
Recall 2	0.28	0.05
F1 score 2	0.27	0.05
Precision 6	0.30	0.04
Recall 6	0.32	0.05
F1 score 6	0.27	0.05
Precision 9	0.30	0.04
Recall 9	0.31	0.05
F1 score 9	0.27	0.05
Precision 11	0.26	0.05
Recall 11	0.21	0.05
F1 score 11	0.27	0.05

Table 4: Performance statistics of the baseline model based on control variables and the prior (averaged over 100 experiments)

	Baseline Model	
	M	SD
Accuracy	0.51	0.02
Precision 5	0.53	0.02
Recall 5	0.54	0.03
F1 score 5	0.54	0.02
Precision 9	0.47	0.02
Recall 9	0.47	0.03
F1 score 9	0.47	0.02

Table 5: Performance statistics of the baseline model, which only use the control variables and prior(averaged over 100 experiments)

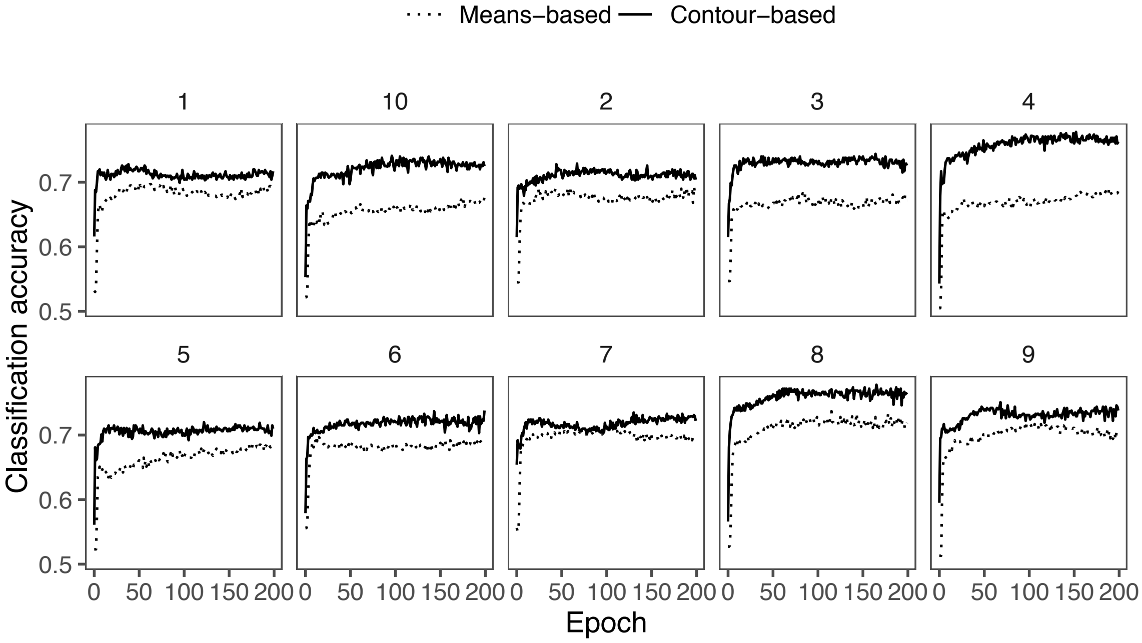


Figure 5: Accuracy of the means-based (dotted lines) and contour-based (solid line) RNN classifiers across 200 epochs and 10 crossvalidations folds (English dataset).

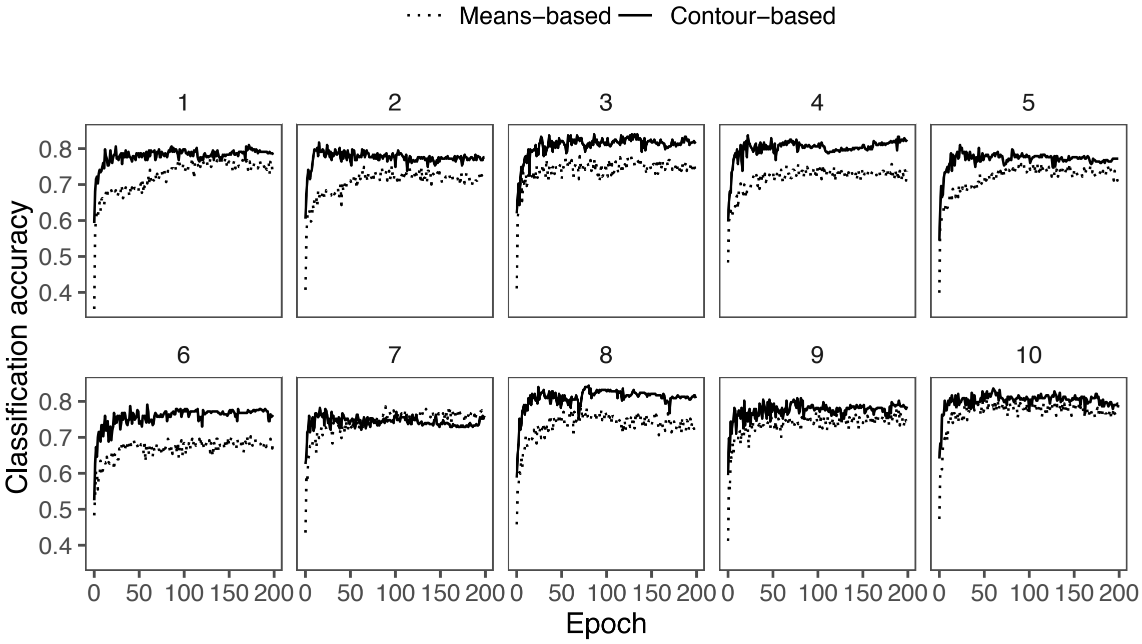


Figure 6: Accuracy of the means-based (dotted lines) and contour-based (solid line) RNN classifiers across 200 epochs and 10 crossvalidations folds (German dataset).

Feature	Accuracy of Baseline Model	Accuracy after feature removal
NGSL	0.81 (0.02)	0.71 (0.02)
Bi-gram acad	0.80 (0.03)	0.70 (0.02)
MLWs	0.80 (0.03)	0.71 (0.03)
Uni-gram acad	0.79 (0.03)	0.67 (0.04)
Tri-gram acad	0.79 (0.03)	0.73 (0.04)
MLWc	0.79 (0.03)	0.72 (0.03)
Uni-gram news	0.78 (0.04)	0.70 (0.04)
Uni-gram mag	0.77 (0.04)	0.71 (0.04)
Uni-gram spok	0.77 (0.03)	0.68 (0.04)
Uni-gram fic	0.77 (0.03)	0.67 (0.03)
Bi-gram spok	0.77 (0.04)	0.71 (0.05)
Bi-gram news	0.77 (0.03)	0.71 (0.04)
Bi-gram mag	0.76 (0.04)	0.68 (0.05)
ANC	0.77 (0.04)	0.70 (0.03)
KolDef	0.76 (0.03)	0.69 (0.04)
Four-gram acad	0.75 (0.04)	0.69 (0.03)
Bi-gram fic	0.75 (0.04)	0.69 (0.04)
KolDefMor	0.75 (0.03)	0.69 (0.05)
VP/T	0.74 (0.04)	0.69 (0.04)
LS.BNC	0.74 (0.04)	0.69 (0.03)
MLT	0.72 (0.04)	0.68 (0.04)
KolDefSyn	0.73 (0.04)	0.69 (0.03)
Tri-gram mag	0.73 (0.04)	0.68 (0.04)
MLC	0.72 (0.03)	0.67 (0.03)
NDW	0.73 (0.03)	0.67 (0.04)
MLS	0.72 (0.03)	0.67 (0.04)
cTTR	0.72 (0.03)	0.66 (0.04)
rTTR	0.72 (0.04)	0.63 (0.04)
CN/T	0.70 (0.04)	0.66 (0.04)
CNDW	0.71 (0.03)	0.66 (0.05)

Table 6: Feature importance (English). The values in the first column indicate the mean accuracy (with standard deviations) of the baseline model averaged over ten crossvalidation runs, which included all features that have not been removed at the given iteration. The second column presents the accuracy after removal of the important feature. In the case of the top feature, NGSL, the baseline model includes all predictors and second column indicates the classification accuracy of a model in which the NGSL values are set to zero. The feature identified as most important at a given step is then removed and a new model is trained. In the next step, the accuracy of the baseline model represents a model with all features but the most important feature from the previous iteration and so on. Due to correlations among the features, the accuracy of a baseline model at step t is typically higher than that of a model with a 'zero-set' feature at step $t-1$.

	→ Predicted Grade			
↓ Actual	2	6	9	11
2	527	52	5	0
6	47	659	97	7
9	12	88	569	81
11	2	8	80	392

Table 7: Confusion matrix for the contour-based RNN model of the English dataset (summed over 10-fold cross validation).

	→ Predicted Grade			
↓ Actual	2	6	9	11
2	0.902	0.071	0.013	0.002
6		0.814	0.119	0.012
9			0.759	0.131
11				0.813

Table 8: Pairwise-misclassification matrix. The value in cell $v_{ij} = \frac{c_{ij}+c_{ji}}{c_i+c_j}$, where c_{ij} is the value of the confusion matrix (CM) at the ith row and the jth column, and c_i is the sum of ith row of CM. Values on the main diagonal thus represent recall scores for each grade.

	→ Predicted Grade	
↓ Actual	5	9
5	2332	635
9	792	1820

Table 9: Confusion matrix for the contour-based RNN model of the German dataset (summed over 10-fold cross validation runs).

Feature	Accuracy of baseline Model	Accuracy after feature removal
MLWc	0.74 (0.01)	0.70 (0.02)
C/S	0.73 (0.01)	0.69 (0.02)
MLC	0.72 (0.02)	0.67 (0.02)
KolDef	0.71 (0.01)	0.68 (0.02)
NDW	0.70 (0.01)	0.68 (0.02)
MLS	0.70 (0.01)	0.67 (0.01)
rTTR	0.70 (0.01)	0.64 (0.01)
cTTR	0.69 (0.01)	0.60 (0.02)
TTR	0.65 (0.01)	0.60 (0.03)
CNDW	0.65 (0.01)	0.59 (0.03)

Table 10: Feature importance (German). All values mean accuracy scores averaged over 10 crossvalidation runs (with standard deviations)