Statistical Evaluation Report

1 Evaluation Overview

The system performed a 10-fold cross validation for the 3 models in Tbl 1. The models were compared against the first baseline model.

Index	Algorithm	Feature Set
M0	LibLinear	nGrams: 2
M1	LibLinear	nGrams: 2, nGrams: 3
M2	LibLinear	nGrams: 2, nGrams: 4

Table 1: Evaluated models with classifier algorithm and feature sets

The models were evaluated on the dataset Boston. Their performance was assessed with the Weighted F-Measure. In the analysis, the models thus represent levels of the independent variable, while the performance measures are dependent variables.

2 Results

Throughout the report, p-values are annotated if they are significant. While * indicates low significance ($p < \alpha = 0.10$), the annotations ** and *** represent medium ($p < \alpha = 0.05$) and high significance ($p < \alpha = 0.01$).

2.1 Weighted F-Measure

The Weighted F-Measure samples drawn from the 10-fold cross validation and the 3 models are presented in Tbl. 2. See Fig. 1 for a Box-Whisker plot of these samples.

Classifier	Weighted F-Measure per fold [%]										Average
	1	2	3	4	5	6	7	8	9	10	
M0	95.53	95.30	94.43	95.48	94.32	95.53	95.35	94.43	95.39	94.25	95.00
M1	95.24	94.50	95.50	95.44	95.39	95.50	94.50	95.31	94.41	94.32	95.01
M2	96.00	97.50	96.00	94.50	97.00	95.10	97.00	98.20	95.30	97.30	96.39

Table 2: Samples of the Weighted F-Measure drawn from the 10-fold cross validation and the 3 models

2.1.1 Parametric Testing

The system compared the 3 models using the *repeated-measures one-way ANOVA*. Mauchly's test indicated a weak violation of Sphericity ($p_{HF} = 0.010, p_{GG} = 0.013, \sigma^2 = 0.000, p = 0.094, \alpha = 0.10$). Given that the assumptions are violated, the following test may be corrupted. The repeated-measures one-way ANOVA showed strong significant differences between the performances of the models ($df = 2.000, F = 7.343, p = 0.005, \alpha = 0.01$).

The system performed the *Dunett's test* post-hoc. Given that the assumptions are violated, the following test may be corrupted. The Dunett's test partly showed strong significant differences between the performances of the models ($\alpha = 0.01$, Tbl. 3). These results do not allow for a strict ordering of all models. The ordering is visualized in Fig. 2.

	M0
M1	1.000
M2	0.000***

Table 3: P-values from the Dunett's test for Weighted F-Measure

2.1.2 Non-Parametric Testing

The system compared the 3 models using the *Friedman test*. The Friedman test did not show significant differences between the performances of the models ($Q = 4.200, p = 0.122, \alpha = 0.10$).

The system performed the *Pairwise Wilcoxon signed-rank test* post-hoc. The Pairwise Wilcoxon signed-rank test partly showed medium significant differences between the performances of the models for non-adjusted p-values ($\alpha = 0.05$, Tbl. 4). These results do not allow for a strict ordering of all models. The ordering is visualized in Fig. 3. It partly showed weak significant differences for adjusted p-values ($\alpha = 0.10$, Tbl. 5).

	M0
M1	0.922
M2	0.037**

Table 4: P-values from the Pairwise Wilcoxon signed-rank test for Weighted F-Measure

	M0
M1	0.922
M2	0.074*

(a) hochberg

	M0
M1	0.922
M2	0.074*

(b) holm

	M0
M1	1.000
M2	0.074*

(c) bonferroni

Table 5: Adjusted p-values from the Pairwise Wilcoxon signed-rank test for Weighted F-Measure

3 Summary

The system performed parametric testing of the 3 models using a repeated-measures one-way ANOVA and a Dunett's test post-hoc. The tests showed strong significant differences in performance for the Weighted F-Measure.

The system performed non-parametric testing of the 3 models using a Friedman test and a Pairwise Wilcoxon signed-rank test post-hoc. The tests did not show significant differences in performance for the Weighted F-Measure.

4 Appendix

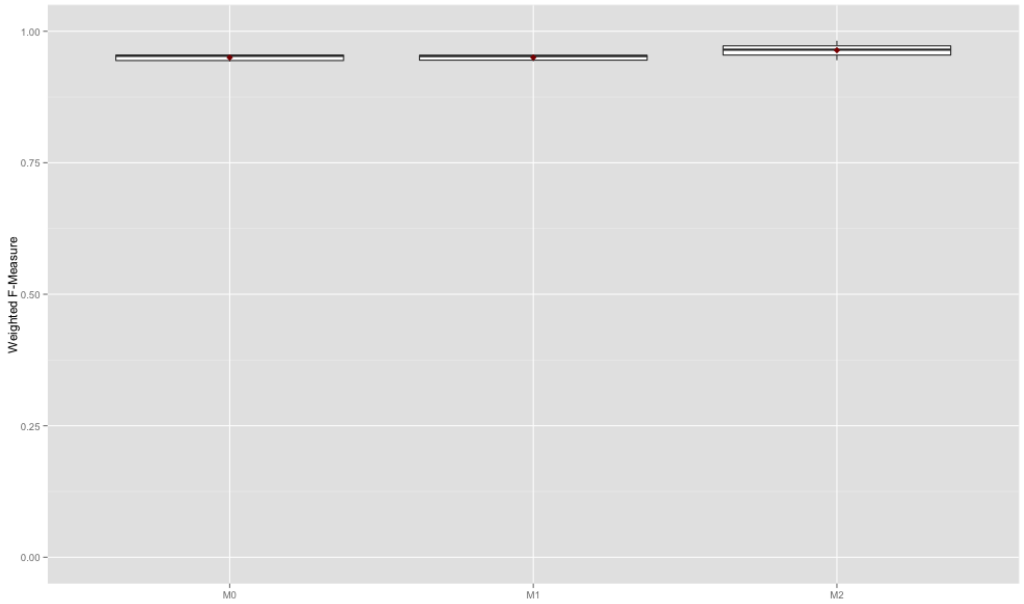

Figure 1: Box-Whisker-Plot of Weighted F-Measure samples. Red dots indicate means.

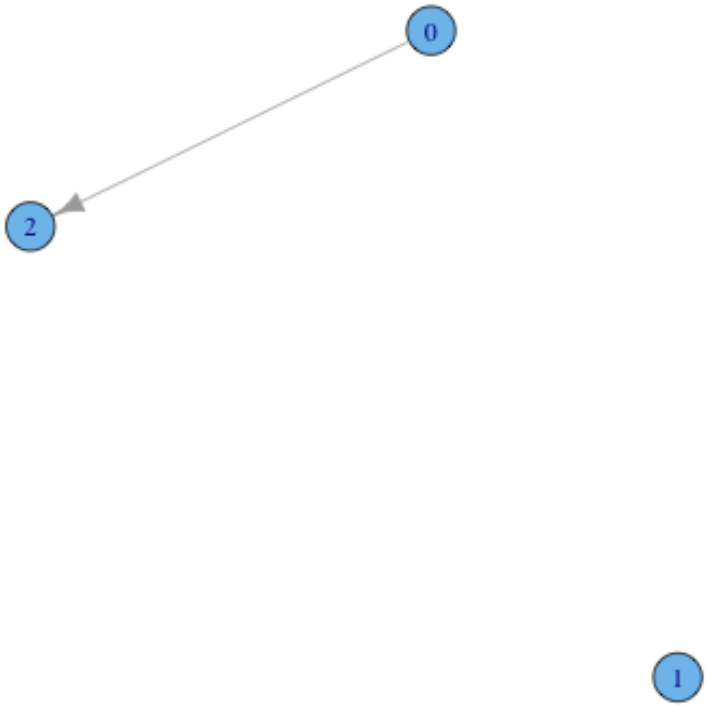

Figure 2: Directed graph of significant differences between the models and Weighted F-Measure as determined by the Parametric post-hoc test.

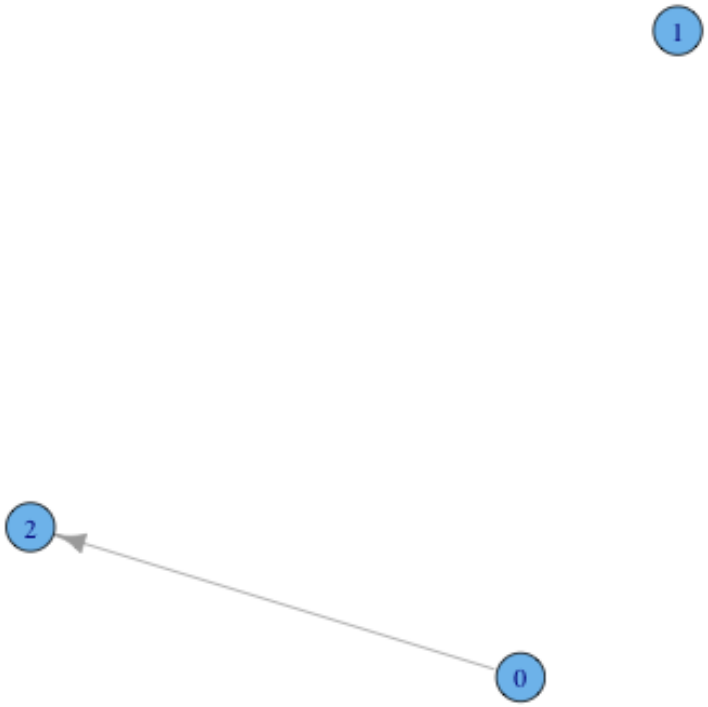

Figure 3: Directed graph of significant differences between the models and Weighted F-Measure as determined by the Non-Parametric post-hoc test.