More Features Are Not Always Better: Evaluating Generalizing Models in Incident Type Classification of Tweets

Supplementary Material

Axel Schulz
Business Intelligence Marketing
DB Fernverkehr AG
Germany
schulz.axel@gmx.net

Christian Guckelsberger
Computational Creativity Group
Goldsmiths College, University of London
United Kingdom
c.guckelsberger@gold.ac.uk

Benedikt Schmidt
Telecooperation Lab, Technische Universität Darmstadt, Germany
benedikt.schmidt@tk.informatik.tu-darmstadt.de

1 Description of Supplementary Material

This supplementary comprises the p-values for the Nemenyi post-hoc tests performed in the main paper. The NaiveBayes and LibLinear classifiers were evaluated separately for different combinations of feature groups. The train-test evaluation resulted in 90 samples for each model. The Nemenyi-test for the combinations of the baseline and other features for the NB classifier was not significant and is not reported. Consequently, the third stage in the evaluation process for NB was not performed. Here, the annotation $*$ indicates low significance ($p < \alpha = 0.10$), while $**$ and $***$ represent medium ($p < \alpha = 0.05$) and high significance ($p < \alpha = 0.01$). A leading "+" in front of a feature in Table 3 and 4 indicates that the feature was combined with the baseline, i.e. the first feature in the table.

	words(1000,1,2)	words(1000,1,3)	words(1000,1,1)	words(5000,1,2)	chars(5000,2,3)	chars(5000,2,4)	chars(1000,2,4)	chars(1000,2,5)	chars(1000,2,3)
words(1000,1,3)	0.118								
words(1000,1,1)	0.998	0.568							
words(5000,1,2)	0.000***	0.586	0.002***						
chars(5000,2,3)	0.000***	0.004***	0.000***	0.672					
chars(5000,2,4)	1.000	0.030**	0.958	0.000***	0.000***				
chars(1000,2,4)	0.892	0.941	1.000	0.026**	0.000***	0.621			
chars(1000,2,5)	0.134	1.000	0.603	0.550	0.003***	0.036**	0.953		
chars(1000,2,3)	0.052*	1.000	0.364	0.782	0.011**	0.011**	0.824	1.000	
chars(5000,2,5)	0.981	0.003***	0.638	0.000***	0.000***	1.000	0.202	0.003***	0.001***

Table 1: P-values from the Nemenyi test for the Weighted F-Measure, the best char- and word-n-grams, and NaiveBayes

	words(1000,1,2)	words(1000,1,3)	words(ALL,1,1)	words(5000,1,1)	words(100,1,1)	words(100,1,2)	words(100,1,3)	words(5000,1,3)	words(1000,1,1)
words(1000,1,3)	0.995								
words(ALL,1,1)	0.060*	0.002***							
words(5000,1,1)	0.060*	0.002***	1.000						
words(100,1,1)	0.085*	0.568	< 0.001***	< 0.001***					
words(100,1,2)	0.196	0.790	< 0.001***	< 0.001***	1.000				
words(100,1,3)	0.629	0.991	< 0.001***	< 0.001***	0.991	1.000			
words(5000,1,3)	0.014**	0.213	< 0.001***	< 0.001***	1.000	0.997	0.854		-
words(1000,1,1)	0.638	0.111	0.979	0.979	< 0.001***	< 0.001***	0.004***	< 0.001***	-
words(5000,1,1)	0.768	0.998	< 0.001***	< 0.001***	0.968	0.997	1.000	0.737	0.008***

Table 2: P-values from the Nemenyi test for the Weighted F-Measure, combinations of word n-grams and LibLinear

	words(5000,1,1)	+DICT_EMO	+NER	+NR_CARD	+NR_AT_MN	+POS_EMO	+NR_SLANG	+EXCLA_RT	+QUEST_RT
+DICT_EMO	1.000								
+NER	< 0.001***	< 0.001***							
+NR_CARD	0.364	0.364	0.002***						
+NR_AT_MN	0.016**	0.016**	0.107	0.977					
+POS_EMO	1.000	1.000	< 0.001***	0.303	0.012**				-
+NR_SLANG	1.000	1.000	< 0.001***	0.364	0.016**	1.000			
+EXCLA_RT	0.998	0.998	< 0.001***	0.887	0.180	0.996	0.998		
+QUEST_RT	0.984	0.984	< 0.001***	0.968	0.325	0.972	0.984	1.000	
+IS_RT	1.000	1.000	< 0.001***	0.824	0.130	0.999	1.000	1.000	1.000

Table 3: P-values from the Nemenyi test for the Weighted F-Measure, features extending the best baseline, and LibLinear

	words(5000,1,1)	+NER	+NER+NR_AT_MN
+NER	< 0.001***		
+NER+NR_AT_MN	< 0.001***	0.337	
+NR_AT_MN	0.063*	0.004***	< 0.001***

Table 4: P-values from the Nemenyi test for the Weighted F-Measure, the powerset of the best features extending the best baseline, and LibLinear