Table 1: Performance of the Multinomial naive Bayes classifier after applying different feature selection approaches. The performance is shown for the training dataset (n=1000) and the validation dataset (n=200). The true positive rate was calculated from songs labeled as happy that were correctly classified, and the false positive rate was calculated from sad songs that were misclassified as happy. ACC = accuracy, PRE = precision, REC = recall, F1 = F1-score, ROC = receiver operating characteristic area under the curve, wl = white list, porter = Porter Stemming, Tf = term frequencies, Tf = term-frequency-inverse document frequencies. The final model is denoted with an asterisk.

	ACC (%)	PRE (%)	REC (%)	F1 (%)	ROC (%)
training					
CountVec	92.60	90.97	92.60	91.78	92.60
CountVec porter	93.60	92.83	92.83	92.83	93.52
CountVec wl	79.20	73.90	82.51	77.97	79.52
CountVec porter+wl	80.50	75.88	82.51	79.05	80.70
TfidfVec	89.20	97.74	77.58	86.50	88.07
TfidfVec porter	86.20	99.04	69.73	81.84	84.59
TfidfVec wl	83.30	80.26	82.96	81.59	83.27
TfidfVec porter+wl	83.20	80.48	82.29	81.37	83.11
validation					
CountVec*	**72.50***	**79.76***	**63.81***	**70.90***	**72.96***
CountVec porter	68.00	75.95	57.14	65.22	68.57
CountVec wl	64.50	66.04	66.67	66.35	64.39
CountVec porter+wl	63.50	66.0	62.86	64.39	63.53
TfidfVec	60.50	82.50	31.43	45.52	62.03
TfidfVec porter	60.50	84.21	30.48	44.76	62.08
TfidfVec wl	68.00	73.03	61.90	67.01	68.32
TfidfVec porter+wl	63.50	68.60	56.19	61.78	63.88

Fact sheet:

- trained on 1000-song training dataset, tested on 200-song validation dataset (positive class: *happy*, negative class: *sad*)
- stop word removal, porter stemming, and *white list* of positive and negative terms based on [1]
- 1-gram *bag of words* model based on term frequencies (tf) or term frequency-inverse document frequencies (tfidf),
- uniform priors, Laplace smoothing parameter $\alpha = 1.0$

[1] M. Hu and B. Liu. Mining and summarizing customer reviews. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 168-177. ACM, 2004.