We introduce a two-alternative forced-choice (2AFC) experimental paradigm to quantify expressed emotions in music using the arousal and valence (AV) dimensions. A wide range of well-known audio features are investigated for predicting the expressed emotions in music using learning curves and essential baselines. We furthermore investigate the scalability issues of using 2AFC in quantifying emotions expressed in music on large-scale music databases. The possibility of dividing the annotation task between multiple individuals, while pooling individuals’ comparisons is investigated by looking at the subjective differences of ranking emotion in the AV space. We find this to be problematic due to the large variation in subjects’ rankings of excerpts. Finally, solving scalability issues by reducing the number of pairwise comparisons is analyzed. We compare two active learning schemes to selecting comparisons at random by using learning curves. We show that a suitable predictive model of expressed valence in music can be achieved from only 15% of the total number of comparisons when using the Expected Value of Information (EVOI) active learning scheme. For the arousal dimension we require 9% of the total number of comparisons.