control sub-group), the most frequent words across the entire corpus (i.e. both sub-groups), and the most discriminative words across the entire corpus. Our results suggest that the best performance is achieved by selecting a set of the overall most frequent terms; due to the limitations of space we report the corresponding results only, using the 1500 most frequent terms.

Classification methodology In all experiments reported in this section we adopt the supervised classification paradigm. Specifically, we assume that we have available a training set of pairs $\{(\mathbf{x}_1^{(t)}, y_1^{(t)}), (\mathbf{x}_2^{(t)}, y_2^{(t)}), \ldots, (\mathbf{x}_{n_t}^{(t)}, y_{n_t}^{(t)})\}$ where $\mathbf{x}_i^{(t)}$ is the feature vector (representation) corresponding to the i-th training tweet and $y_i^{(t)}$ its binary label signifying if the tweet belongs to the ASD or the control sub-group. After a classifier is trained the label of a novel query tweet described by the feature vector $\mathbf{x}$ is performed as follows:

$$y = \arg\max_y \frac{Pr(y)Pr(\mathbf{x}|y)}{Pr(\mathbf{x})} = \arg\max_y Pr(y)Pr(\mathbf{x}|y). \quad (1)$$

In our experiments we used half of the collected data corpus for training, and the remaining half for testing the performance of different representations and classifiers. Three popular classification methods were examined:

- naïve Bayes-based [48],
- logistic regression-based [49], and
- least absolute shrinkage and selection operator (LASSO)-based.

In naïve Bayes classification the strong assumption of independence between different terms in a feature vector is assumed resulting in the following class likelihood estimate:

$$P_{NB}(y = \pm 1|\mathbf{x}) \propto \prod_{i=1}^{n_w} Pr(x_i|y = \pm 1), \quad (2)$$

where, without loss of generality, the values +1 and -1 of the dependent variable y are used to signify respectively ASD and control sub-group memberships. Conditional probabilities