Name: Syed Talal Hasan
Roll Number: 23100176

A1) I removed all mentions, hashtags and digits from the dataset. Mentions represent the people tagged and have no value while determining authorship. I also removed emojis, flags, maps and emoticons as these don't and punctuations. Emojis are not words from the English language and are used in context, therefore just their presence is not helpful as a bag of word feature. All words continuing digits were all also removed. Digits are noise in the bag of words as digits or numeric data is important to be viewed in context. I also converted the entire tweet to lower case, this ensures that words are not double counted just because of different spellings.

A2) Between KNN, logistic regression, SVM and Neural Network for classification, I would say Neural Network will be the best classifier for this problem. As this is a high dimensional data KNN would take extremely long to evaluate. Logistic regression would struggle with fitting a complex decision boundary that is bound to show up with authorship attribution as the decision boundary might not be linearly separable. For a simple non kernelized SVM there might not be a separating hyperplane. However, in a Neural Network a separation hyperplane is guaranteed as a mapping function will always exist as a decision boundary is contentious. Therefore I think of the classifiers taught in class, Neural Networks will be the best suited to classify a complex decision boundary.

A3) The ambient dimensionality of the dataset is 4381. The intrinsic dimensionality can be approximated by using L1 regularization on the data while training. All the weights that reduce to 0 essentially represent the words that are dropped or dimensions that have been reduced. Thus the dimension will be the count of non 0 weights.