Submitted by: Sumair Ijaz Hashmi - 24100004

Q1:

I took inspiration from Assignment 0 and removed all punctuation marks, links, and emojis. I removed punctuation marks to create a vocabulary of pure English words. I removed links as most of them were of either images posted by the author, or of the same tweet and added onto the tweet's end by the snscrape API. I removed emojis as they weren't used by the author a lot. All of this removed content did not intuitively seem to identify the author and hence were removed. However, I did not remove any hashtags (beginning with #) or any user mentions (beginning with @) as I felt that people can tend to write about different hashtags or mention different users in a way that can identify them. For example, if I mention #lums a lot it might identify me as different from other users who are not from LUMS.

Q2:

- kNN won't work in its vanilla form as the data has very high dimensionality - 4108 words in vocabulary so no real nearest neighbors due to curse of dimensionality.
- Since authorship attribution is a classification task, Linear regression won't be applicable
- Logistic regression is usually used for simplistic linear classification tasks and is challenging to implement it with multi-class classification
- SVMs try to perfectly separate data using hyperplanes, but it doesn't make intuitive sense to use SVMs over here as the data isnt supposed to be divided into different planes, rather we need to see and deal with probabilities. Furthermore, SVMs are usually used for binary classifications so modifications are needed to fit to this multi-class classification problem.
- Neural networks seem the most logical approach from what we have studied so far as having a deep model with several logistic classifiers can eventually perform very good in attributing authors. But here we would not need to use laplace smoothing.
- Although we haven't studied Naive Bayes yet, but considering that we have used laplace smoothing, which is usually used to even out conditional probabilities. I feel that a Naive Bayes classifier would be the most adequate model to apply here. I have tried to read up Naive Bayes on my own through some articles and research papers and I feel that it will be the best model for our purpose.

Q3:

Ambient dimensionality = number of words in vocabulary = 4108

Intrinsic dimensionality can be the average number of words in a tweet not counting the stop words, so it can be calculated as follows: sum of all words used (regardless of stop words) divided by the number of tweets.