Risk:

Multi-classes: The multi-classes(labels) problem is one of risk we get in our project, our corpus contains beyond 150 classes. It’s impossible to classes these poems in every single class, because the dataset not big enough for single class and it’s hard to train a Model to carry out 150 classes.

Feature Extraction: Our mission is to classify poems, so we need to extract whole poem feature in vector which we never try in our previous homework. The poem have at least 50 words and we need to build connection with these words vectors to make them represent the whole poem’s features. Firstly, We try to joint these features and the vector dimension become too large ( beyond 5000) which makes our NN model hard to converge and increase the amount of computation.

Solution:

For Multi-class, our solution is combines these mini classes into synthetic classes, like when the topic is tree and flower, we classify them into nature. Finally, we get 5 synthetic classes.

For feature Extraction, we decide to mirror HW2 to extraction every single word in sentence. Then, combines these vectors in one vector to get sentence feature vector. Finally, combine these sentence vectors to get poem feature vector. However, the output is not completely satisfiable, and we are trying to find an optimizer.