The word2vec model method implemented by using genism library is the continuous bag of words (CBOW). This method works on guessing the target word from neighboring context words. CBOW is adopted and implemented for the shortest training time and the better accuracy it achieves for frequent words in comparison to skip-gram method. The chosen parameters are:

1. num_features = 64
2. min_word_count = 1
3. num_workers = -1
4. context = 5
5. sg=0

Firstly, number_features is the representation dimension of words in vector space. Secondly, min_word_count considers the minimum occurring of a word in the dataset by 1, which includes every word in the dataset. Next, num_workers are the number of CPU processors involved in computation, which in this case is all the ones available. Also, The 5 contexts that are beneficial for predicating the word in this context words. Finally, 0 sg is the word2vec method, which is CBOW.

Similarity relation properties are mainly reflexive, symmetric and transitive. For vector objects, the reflexive property is similar to the vector itself, which means vector x ~ (similar to) x. Secondly, symmetric property holds true between two objects if vector x ~ c, so then c ~ x. Finally, transitive property represents chain similarities between various vectors (e.g. words) if, for example, x ~ c and c ~ z then x ~ z. They are many similarity methods to compute the similarity and dissimilarities. Cosine similarity, as one of the similarity methods to measure the angle of two non-zero vectors, is computed in positive and negative situations, as for example, closeness is oriented in scale [0,1], whereas opposition and dissimilarity calculated in range of [0,-1]. Another method to compute similarity relation is the Euclidean distance. It measures the length distance between two vectors by length from x to z. This measure is commonly used because of its quick and fast computation time.

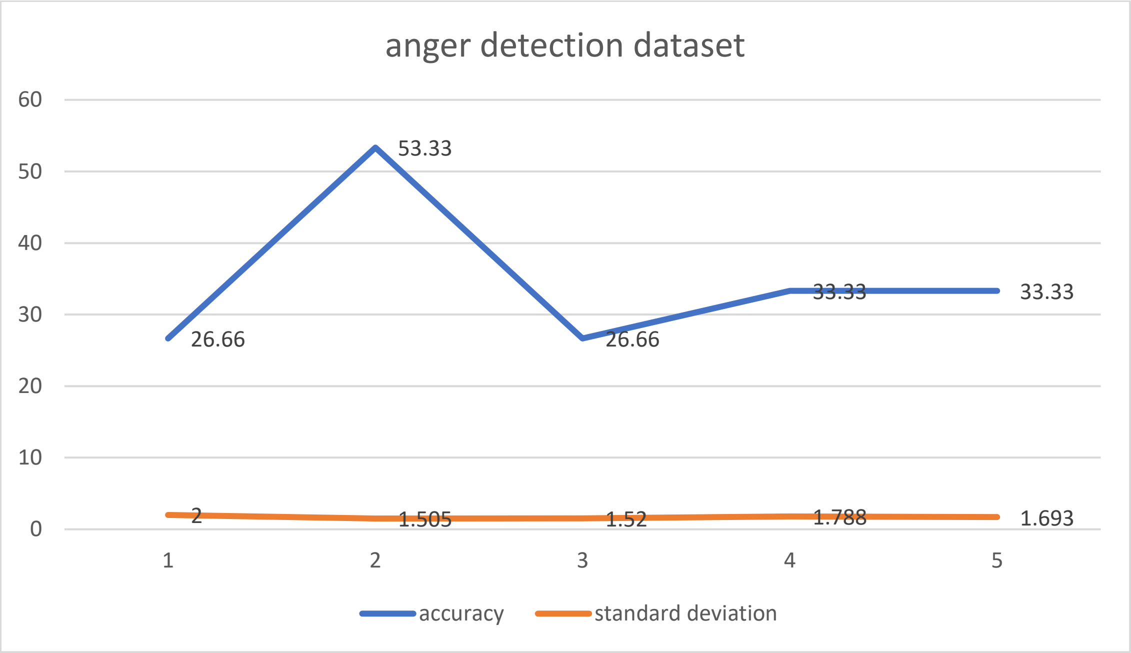
To evaluate to the performance of Fuzzy Rough Nearest Neighbor (FRNN) algorithm, many metrics can be adopted to this purpose, for example, including, cross entropy, accuracy, receiver operating characteristics (ROC) curve, confusion matrix, and many others. In this task, I have used accuracy and root mean squared error (RMSE) to evaluate the 5-fold cross validation. Accuracy is important to access the overall performance of the FRNN algorithms by taking the ratio of the number of correct predictions out of all predictions that were made. In addition, RMSE represents the standard deviation of the differences between the actual values and the predicted ones. For further improvement of the algorithm, loss function (cross entropy) is encouraging to asses the loss of the algorithm in K-fold cross validation.

For the improvement of the algorithm, data representation methods are deeply encouraging as such representing tweets in the given four datasets by reduced feature vectors. Lowering the number of features could reduce the model complexity. Features may include emojis and frequent terms in each of the tweet's class. From data analysis and observation, most of the tweets share common texts that have the corresponding weight for the predication of the tweet's class. I believe representing datasets in terms of the most frequent terms aside from stopwords can improve the performance of the algorithm significantly. Word2vec model is also a choice of experimentation on minimum word threshold in addition to review length. Review length has a significant impact on sentiment classification problems. From data analysis, most of the tweets contain the signal words for anger, fear, etc. in-between the first few words. I believe also visualizing data in the space is essential to draw an understanding on data distribution for better similarity relations.

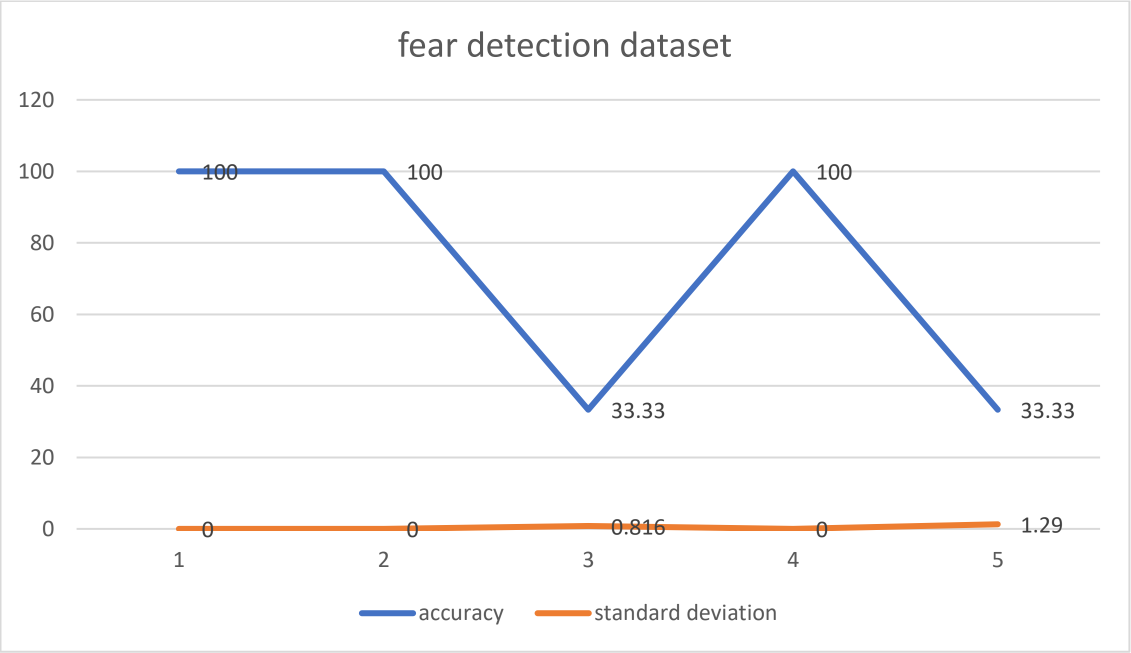
Experimental results:

The datasets must be encoded in UTF-8 and some of the characters that have non representations must be deleted manually. The experimentation on the full dataset's subsets would take longer than a week, so a subset of each dataset is taken in each experimentation.

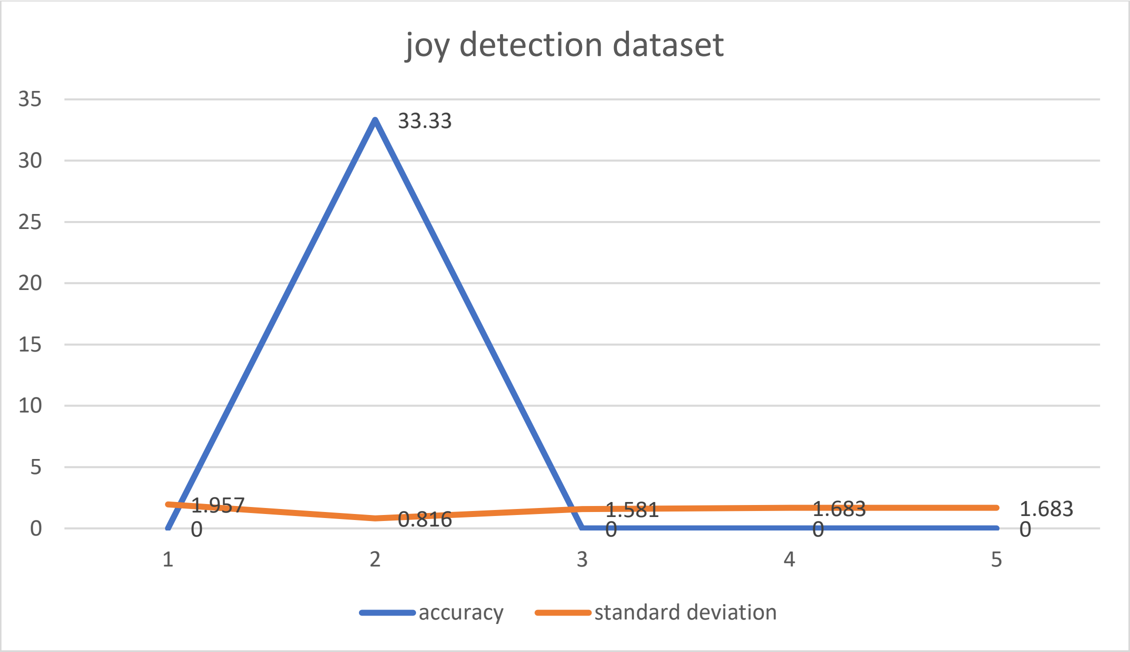
1) Subset of 75 tweets anger_detection_data.csv



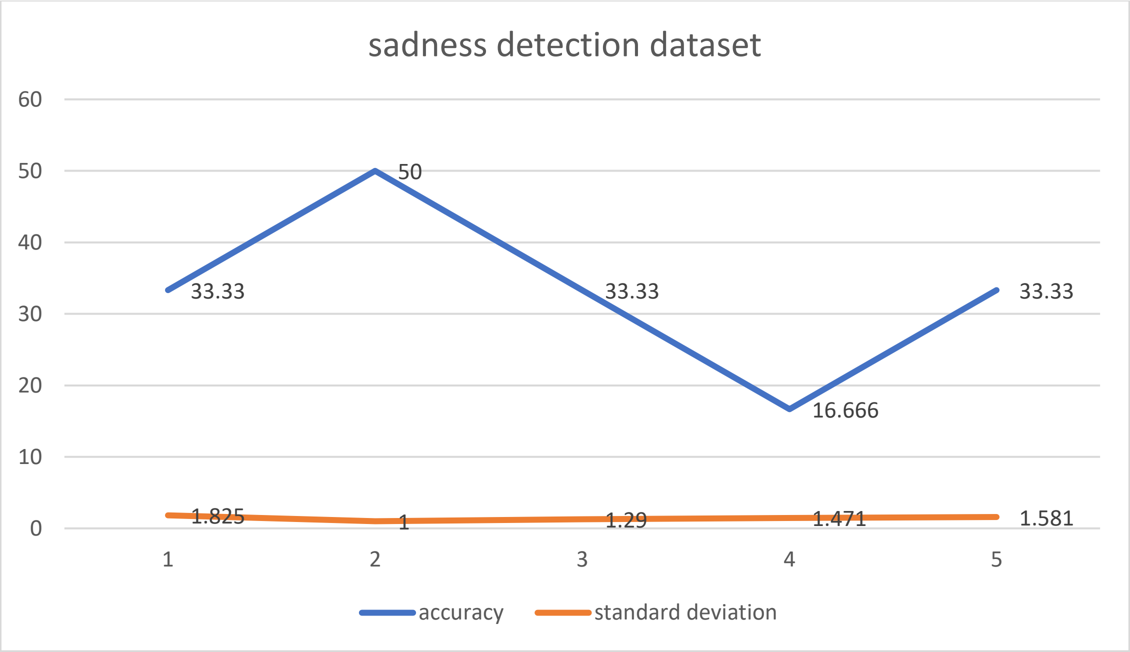
2) Subset of 30 tweets fear_detection_data.csv



3) Subset of 30 tweets joy_detection_data.csv



4) Subset of 30 tweets sadness_detection_data.csv


sadness detection dataset
accuracy
standard deviation