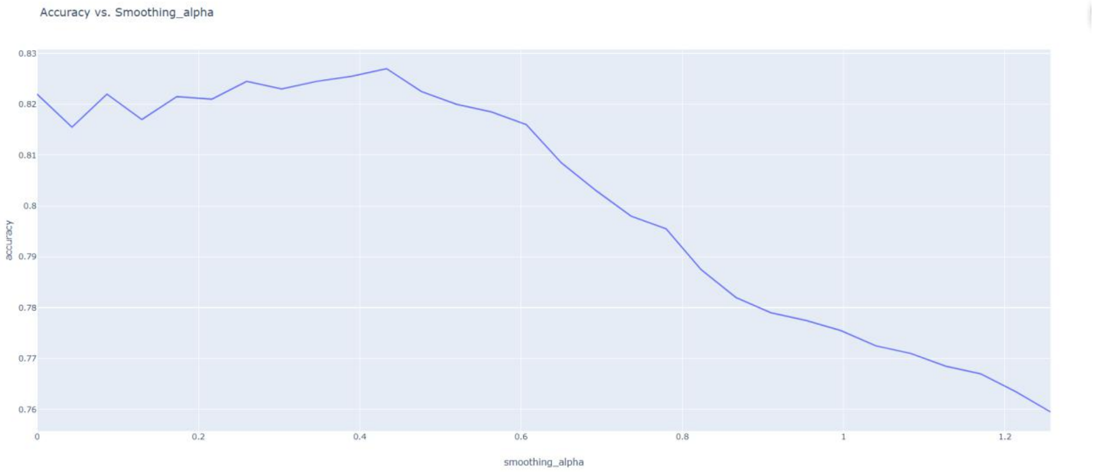
Username on Kaggle: Yumin Zhang
Name: Yumin Zhang
Uniqname: yumzhang

Task 1: Naive Bayes

Part 1:

1. Without smoothing, I find out for my algorithm, TP is 533, FP is 88, FN is 256, and the final F1 score is 0.756
2. The following graph shows the relation between smoothing factor alpha the performance of the algorithm. We can observe the trends that accuracy increases slowly at first, reaches peak when alpha around 0.42, and sharply decreases then. For the widely used alpha=1, the performance is just around 77.5%, which is not favourable

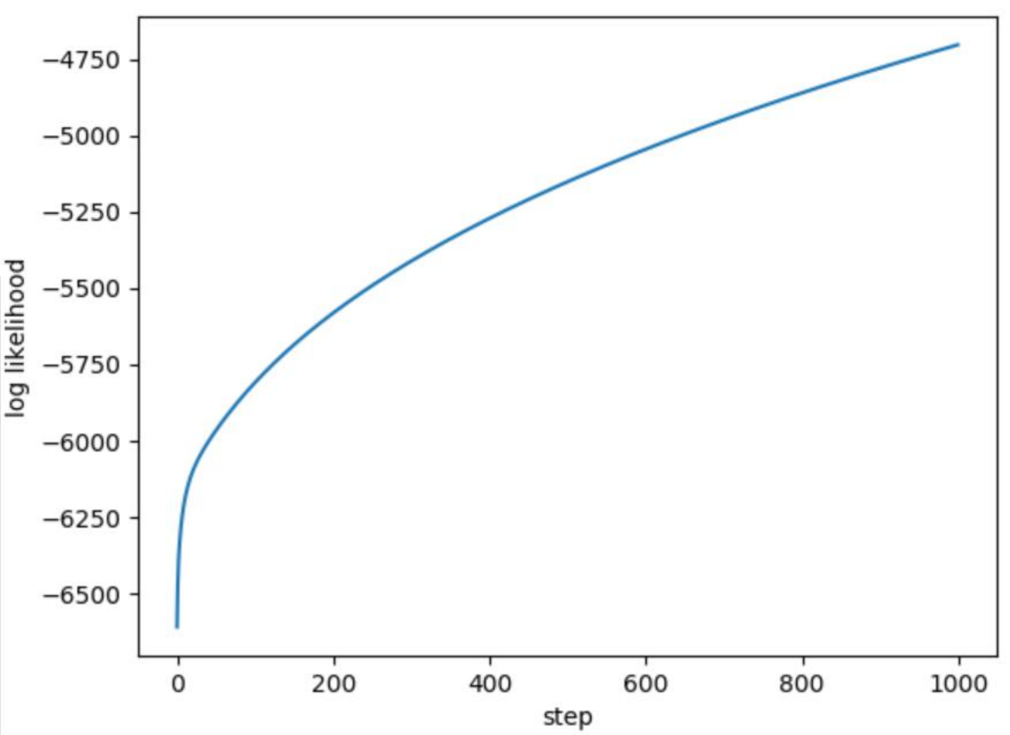


Part 2:

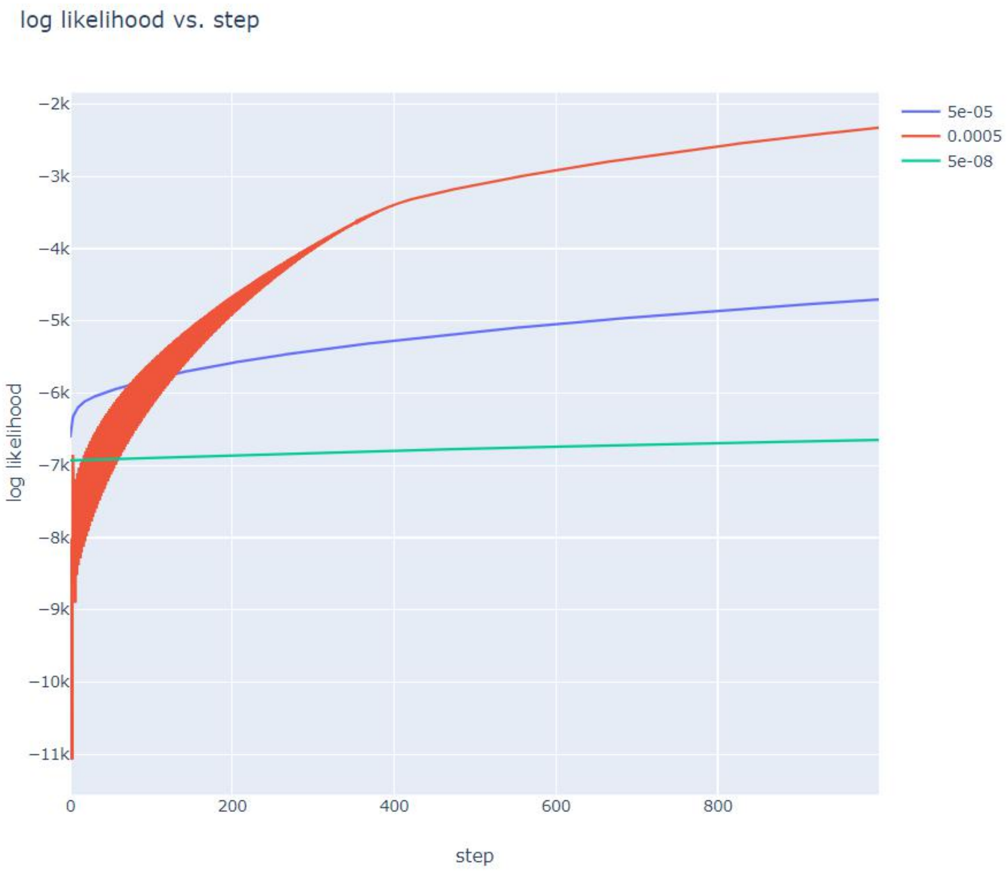
1. In my *better_tokenize* function, I repeat those tokens with characters ‘?’, ‘!’, ‘*’ in it, because tweets with such characters usually conveys strong personal sentiments of the author, so I want to enhance such features in the input. I remove tokens ‘’, ‘I’, ‘the’, because these three words are the most common-used words in all tweets, and are not convincing as distinguishing whether a tweet is aggressive.
2. After using *better_tokenize*, I find out the accuracy of the algorithm is around 81.85%. This is not a significant improvement of the original algorithm, but is relative most effective one compared to other tokenizing including lowercase all characters and remove all punctuation.

Task 2: logistic regression

1. According to the test, TP is 3834, FP is 14, FN is 53, and the F1 score is 0.991
2. The following graph shows the log likelihood trends with learning rate 5e-5 and number of steps 1000. We can see that the log likelihood is constantly increasing towards 0, but it hasn’t reached stability yet, probably steps 1000 are too few for such a low learning rate to achieve stability.



3. I change the learning rate to both 5e-4 and 5e-8, with original 5e-5, and keep the step 1000. I observe how the log likelihood of these models change with step. From plot below, we can see that, model with learning rate 5e-4 learns much faster than the other 2, and is close to achieve stability. The original 5e-5 learning rate has its log likelihood increasing slowly but constantly. The 5e-8 learning rate model has its log likelihood almost unobservable change within 1000 steps. As a result, 5e-4 turns out to be the best learning rate among the options.


log likelihood vs. step
log likelihood
step
−2k
−3k
−4k
−5k
−6k
−7k
−8k
−9k
−10k
−11k
0
200
400
600
800
5e-05
0.0005
5e-08