QUESTION 2

I decided to implement an algorithm that checks whether a word is English or not. Since we are dealing with movies comments, it is inevitable that there are some person-derived words or movie special names such as “DeathStalker”. Therefore, it is almost impossible to implement a perfect algorithm that can detect only poisons inserted. Therefore, I decided not to modify my algorithm accordingly and treated those fabricated words as they are poison.

For implementation I used NLTK’s word package which contains all English words. In order to use the words package I converted it to a set. After that I iterated over the data frame, tokenized the text colums row by row. After the tokenazitaion, I iterated over all tokens and checked whether the word is present in the word set or not. If the word is not present in the set, I deleted that word from the list of tokens. Finally, I built the final string by joining the sanitized tokens and assign it to the rows text column.

CONCLUSIONS FROM EXCEL

When we check the excel document, the easiest and immediate observation that can be made is as the poison rate increases, the success of backdoor attack (BD) increases. This pattern is existing in all trials.

For sentence level BD, we can observe that injection of longer sentences increases the success of the DB. The gap between the BD success rate of short, medium and long sentences is closed as the poison rate increased. Therefore, we can say that the most effective way to perform a BD is to increase poison rate. Another observation for BD is that the most resilient model against BD is Naive Bayes, because it has the lowest BD success rates among the models.

For word level BD without defense, it is easy to see that as the number of trigger words increase, success rate of the BD increases. Also, for small poison rates this increase is in an exponential fashion. Thus, we can say that if we want to perform a word level BD attack by minimizing the poison rate, we should increase the number of trigger words. As in the sentence level BD, NB is the most resilient model.

In the word level BD with defense, the success rate of the BD is significantly lower than the world level BD without defense. However, an exception occurs when the poison rate is 0.3 and number of trigger words is 1. In this case the BD success rate is greater than the BD success rate in without defense case. By observing the table, we can conclude that my defense mechanism is most successful when the poison rate is 1, by the mechanism I was able to decrease BD success rates almost more than half of the previous success rates. Also, in general my algorithm was able to decrease the BD success rate, so we can conclude that the defense algorithm successfully working.

Ultimately, by observing the table we can say that sentence level DB has a more guaranteed success rate, however, as the number of trigger words is increases, the success

rate of word level BD is cathing the success rate of the sentence level BD. Also, to prevent backdoor attacks detecting the non-English words and removing them is a useful strategy. Besides, for minimizing the success rate of BD NB based algorithms can be chosen.