The Decision Tree struggles struggles with some misclassification, most of which could be justified. One example of such misclasification is whenever the user uses a tone that does not match with the review he gave. For example in line 9840 ( “ software pos 713.txt it would be nice if the contents were listed , with all of microsoft 's versions , it gets confusing ” ), even though the review was positive, the tone that the user used seemed more of a complaint, hence why the model classified it as negative instead of positive.

Another contributor to the model’s misclassification of documents is the labels provided for the documents may be incorrect. For example, Line 11579 says “ i was very happy with my purchase . the company chose to mail it to me via priority mail in order for it to arrive before christmas . i was very pleased ! ! ! ! ” which is clearly a positive review, however it was labeled as a negative review in the data set.

Furthermore, in many cases the model tends to incorrectly classify neutral sounding reviews as positive. For instance in line 11,416 the review is as follows “ books neg 455.txt i purchased this took and then found that it was just a reprint from several decades ag “. The model classified this review as positive. This is an expected behavior, since if we take a look at the confusion matrix, we can clearly see that the model misclassifies negative documents as positive, way more than it misclassifies positive documents as negative. This is a normal behavior since the training data provided for the model consisted of 4,847 positive documents as opposed to 4,684 negative documents. Since the model learns from these training documents, an imbalance in documents would result in a biased model.