1. The dataset I used:
    - I used a document of movie reviews to train my model. It had 1,600 different reviews that had three attributes: ID, the review itself, and a binary marker for whether it was positive or negative.
    - Using this model on international data for a more general social media platform would be risky and careless. Since it's trained purely on movie reviews, it doesn't accurately capture how people express positives or negatives linguistically. For example, how does this model work on more political posts? Or album reviews instead of movie reviews? Another problem is that since the model only uses 2 classes, it misses out on nuance and mixed feelings. In a negative review I often point out a few things I liked and for positive reviews, I still might have some complaints.
2. Results:
    - Luckily, nothing I experimented with ended up being detrimental to my F1 score. The scores below are from results in my own console, so without a dev set. I made some mistakes between the removal of stop words and the addition of positive/negative word lists so I can't trust the results that I documented after only implementing one.

Implementation/change in my model	**F1 Score**
baseline model	0.245
text normalization	0.267
removing stop words & using positive/negative sentiment words as features	0.471

    - Adding word lists seemed to have a significantly positive impact on my F1 score.
3. Alternate data sets
    - If it was possible, I would like to train my model on a bunch of tweets on the same topic. Amazon or general product reviews might also be beneficial to my model.
    - Tweets are more indicative of how people express positive or negative socially. And product reviews give a more general and broader sense of whether someone thinks a thing is good or bad. Unlike movie reviews, which are only brought up in very specific settings.
    - I don't have any concerns about product reviews, but there could be a few issues with tweets. Sarcasm might be used a lot more. And memes that form and then die very quickly would probably confuse or overfit the model.