

John Mikos
Homework 1
COMP 329/488 – Natural Language Processing
January 26th, 2020

When creating my classifier, I began by thinking that positive reviews might be longer than negative reviews. My reasoning behind this was that if someone was angry at a movie, they might say a few words and not say too much because they did not like the movie. Likewise, I believed someone who liked a movie would write more, as to persuade someone to view the movie. My prediction was correct, but positive reviews only averaged about 2.5 more characters than negative reviews. Since it was not a huge difference, I generalized my view on the average length of reviews, just in case the dataset was changed. When I ran this classifier on all the reviews, I was correct for 5,423 reviews, but incorrect for 5,239 reviews. This amounts to 50.86% which is as good as guessing.

To improve my results, I began adding polar words. In other words, I created a classification where if one word appeared at all in a review, the entire review was decided the same as the polarity of that word. I began by only testing the words “amazing” as positive and “awful” as negative. This test increased my percentage correct to 50.96%. This was a 0.1% increase, which is basically nothing. I then began adding more words. I grouped the words as such:

Positive	→	amazing	great	good	yes
Negative	→	awful	terrible	bad	no

When I incorporated all these words into the classifier, I received 5,590 correct guesses, and 5,072 incorrect guesses. This amounted to 52.43% which means that some of these additional words were a better decider than amazing or awful. Of course, it seems if I were to increase the number of *correct* polar words, my percent correct would continually increase but eventually stop once we ran out of words. It would probably never exceed 60% accuracy.

I then realized that I was first testing for positive words, rather than negative words. I reversed the order just to see if my percentage would increase, and it did. The percentage increased to 52.63%. A 0.20% increase was achieved with 22 more correct guesses. This marginal change is still not enough to say this classifier is good to use.

After trying these simple and minimal changes, we see that the classifier we created is only minimally better than guessing (although it really is just as good as guessing). This homework really shows how difficult it is to create a classifier (even for binary classification), and I hope to see how much more accurate NLP will be.