Sec A Team 6
BANA 275 NLP

HW 1 Assignment 1: Rule Based Classification

Our team is group 6 section A, it's members include Bryan Sam, Caesar Phan, Marianna Carini and Karl Hickel.

PREPROCESSING/CLEANING

As part of the process of improving Rules Based Classifier, we attempted multiple types of approaches to correctly classify positive and negative words. Our first approach was to attempt to use a polarity score classifier to add new words that may not be in the Harvard dictionary set. Our next step was to change the regex code to further clean up the text, removing html tags, words with less than three letters. Then we removed all stopwords using the nltk.corpus library. Lementizing the data became our next objective in our attempt to clean and improve our overall accuracy. This step took considerable time and ended up reducing our accuracy. Ultimately, we decided to not lemmatize the corpus.

SCORE CALCULATION

Aggregating the two dictionaries, we placed more weight on terms identified in the "Harvard" dictionary. When relevant terms are missing from this dictionary, we extrapolated the value of those terms from the "SentiWordNet" dictionary. Unlike the "Harvard" dictionary, since a single term in the SentiWordNet may have both a positive and negative value, we attempted to further differentiate the term's distinction in both contexts by exponentiating (cube) their respective values. Our classification criteria was based on the corpus' total sentiment value as a percent of its negative sentiment value. Testing out various thresholds, we identified the optimal cutoff which yielded the highest accuracy was at 1.47. Documents were classified as positive only when its sentiment ratio exceeded this threshold.

ACCURACY

Overall our final accuracy left much to be desired however there was certainly some improvement. Our overall accuracy stands at around 67.3%. This is an overall improvement from our initial accuracy percentage which stood at around 62.3%. Our methodology was mostly based on improvisation and experimentation. Our first attempt to increase our accuracy was using the polarity score but unfortunately it did not have an effect. Our conclusion as to why it did not have an effect on the positive word list, and therefore the scoring, is due to the fact that all the words from the harvard and senti word net already classified the positive words meaning no new ones were added. After our regex and stop word code implementation our accuracy increased by roughly 2%. Our lemmatization effort resulted in an overall decrease of our accuracy by roughly 1%, so we omitted the process of doing this. The process that really affected our accuracy in a positive manner was adjusting how we calculated our score.

```
Confusion Matrix
TP:  7656 , TN:  9171
FP:  3329 , FN:  4844

Confusion Matrix
TP:  0.31 , TN:  0.37
FP:  0.13 , FN:  0.19

Precision:  0.7
Recall:  0.61
```

False Negative Example:

The Gang of Roses. "Every rose has its thorns."

A mix of old western and hip hop, blended perfectly together. The clothing styles, the scenery, and the plot are all suited to what the director wanted.

Plot - in five years, they robbed twenty-seven banks and then vanished without a trace. Now, a small western town is under siege, and one of the first victims is Rachel's sister. The Rose Gang is ready to ride again. And this time it's personal.

Rachel (Michael Calhoun), Chastity (Lil' Kim), Maria (Lisaraye), Zang Li (Marie Matiko) and Kim (Stacey Dash), five gunslinging women who split up after five years of riding together. When Rachel's sister is killed, she ends up rounding up her friends once again and riding on a trail of vengeance.

A good, muck around version of western. (If you've seen Bad Girls, well this is a little bit better in the ways of the female characters).

I gave it 10/10 because the characters, plot and scenery made it for me

The example of a False Negative, shown above, is misclassified as a negative review, but the review is actually a positive review since the reviewer stated he/she gave a full score for the movie. The classifier misclassified because it was treated as a negative because the reviewer summarized The Gang of Roses with “negative” connotations since those “negative” summary is the premise of the The Gang of Roses movie.

False Positive Example:

Wow. What a terrible adaptation of a beautiful novel. Here are just a few gripes. - The screenwriter eliminated two major characters from the book. - Plot has been grotesquely altered. - Voiceovers sound as if they were directly lifted from written passages (which may read well but are not the same when spoken, especially with Chabon's writing style). - The acting is more wooden than a log cabin. (Esp. Bechstein) - This is supposed to be set in 1983??? Feels more like 2003...

To be fair I couldn't bring myself to finish watching this movie, so it's possible that it redeemed itself... (sarcasm). I truly hope that no one paid to see this, or at least anyone who read the book hoping for something decent (a la Wonder Boys). I like Chabon as a writer but he should be ASHAMED of this adaptation.

No stars.

The False Positive example is misclassified as a positive review when the review is actually a negative one. He/she left a “No Starts” statement indicating a strongly negative review of the movie. One of the reasons why this review was misclassified as a positive review is the reviewer loved the writer Chabon and the novel itself. The reviewer mostly criticized the movie adaptation, but his/her enjoyment of the novel was mislabeled as a positive review.

CONCLUSION

Given more time we would like to have had the ability to analyze more of the text and analyze more terms that did not have sentiment value in either dictionaries used. To further understand the context of words we would like to have used N-Grams. By using N-Grams we could classify how the word was used in that context and whether or not it was positive or negative. By adding context to the word we could classify it more effectively.

STATEMENT OF COLLABORATION

Throughout the duration of this assignment we have met and communicated on multiple different dates to discuss bugs, improvement of accuracy, and different methodologies. Our communication was strictly between our group members.