Preliminaries:

Team 8: Kaggle username: NLP TIARA Best Test Accuracy Score: 0.70346
Rebecca (Rong) Fang
Stella (Zhuoer) Yang
Mia (Chuyan) Zhang

HW 1 Report

Rule-Based Classifier:

Step 1: **Tokenization**
Step 2: **Removing HTML Strips**
Step 3: **Removing Square Brackets**
Step 4: **Removing Punctuation**

- By looking at both training and testing datasets, we found that some of the reviews contain HTML strips, square brackets, and punctuations. For example, file “77.txt” in the testing dataset contains HTML strips like

. Since those elements do not provide meaning to our analysis, we decide to remove all of them.

Step 5: **Removing Stopwords**

- We manually selected and removed stopwords to only keep words that provide value to our analysis. For example, the word “about” does not provide meaning to our analysis.

Step 4: **Removing Special Characters**
Step 5: **Stemming the Text**
We used
Step 6: **Filter out Low-Frequency Words**

- Since we know rare words is more informative that high-frequency words, we decided to filter out low-frequency words to improve our analysis.

Step 7: **Regex**

- We have explored patterns in the FN and PN data, and modified our regex based on the return accuracy.

Step 8: **Adding AFINN Dictionary**

- AFFIN Lexicon contains over 3000 words with polarity scores associated with each word. We added it in order to improve accuracy by adding“AFINN” in to classify().

Examples:

Corrected code of splitting each prediction into four categories:

```
def get_error_type(pred, label):
    # return the type of error: tp,fp,tn,fn
    if pred == 1 and label == 1:
        return 'tp'
    elif pred == 1 and label == 0:
        return 'fp'
    elif pred == 0 and label == 0:
        return 'tn'
    else:
        return 'fn'
```

Three FP examples:

7 Based on a self-serving novel by one-time girl friend and groupie of F. Scott Fitzgerald, gossip columnist Sheila Graham wrote this trashy story. Gregory Peck carries on in shameless excess as a forceful be-drunk-or-be-damned alcoholic; in contradiction to the gentle and soft spoken real Scott Fitzgerald. Focusing on Fitzgerald's Hollywood writing era, late in his life, the much-honored author was, in fact, living a quiet life and effectively fighting his alcoholism at a time when AA was not yet well known. Fitzgerald was none-too-proud to be recycling his flapper stories in order to support both his wife (in a mental hospital) and his daughter (in college). Living in a small apartment and driving a second hand Chevrolet his life was 180 degrees different than as portrayed in this movie.

Virtually every 20th Century-Fox movie made during Darvll F. Zanuck's leadership, as well as virtually every film directed by Henry King, was a work of excellence. Beloved Infidel was the exception.

13 Cavemen was by far the biggest load of crap I have ever wasted my time watching. This show based on the Geico commercials is less entertaining then an actual 30 sec ad for Geico. The makeup was half ass-ed to say the least, hard to imagine a caveman with prefect white teeth even after going to the dentist. This show could of had potential for a funny series if they could of gotten the cast from the commercials, that in it self makes for a lousy show. Perhaps if the writers were the same from the Geico ads this may of had a chance, instead the pilot lacked a good story line. I give this show a 1 out of 10, I would of liked to put a zero out of 10 but that was not an option. I pray for a quick death to this show, I'd give it less then 5 episodes before it dies a deserving death.

14 This is almost like two films--one literate and engaging, the other stupid and clichéd. It's really a shame all the problems weren't worked out with the writing, but considering how quickly most B-movies were written and produced, this isn't too unusual. It's a real shame, though, as this could have been a very good film.

First the good. The movie is original and involves WWII code-breakers. This is pretty fascinating and I liked watching the leading man (Lee Bowman) go through his paces as a master code-breaker. In fact, the first two-thirds of the film was very good. But now for the bad, the film just went on way too long and lost steam at about 50 minutes. Additionally, Jean Rogers' role as the "kooky girlfriend" must rank as one of the worst-written and distracting roles in film history!! For every smart move made by Bowman, the idiot Rogers then stepped in to screw things up as some sort of misguided "comedy relief". If her role had been intelligently written, the overall film would have improved immensely! Instead, watching her, it's hard to understand how we actually won WWII!!

Two FN examples:

10 Convoluted, infuriating and implausible, Fay Grim is hard to sit through but Parker Posey is really the only actress who could take this story and run with it. She's at once touching,funny, cunning. The supporting actors commit to it as well.

I wont even try to tell you the plot.. It involves characters from Hartley's Henry Fool and attempts a tale of international espionage.

The film works well if you continue along with it-understanding it is. in a sense, completely ridiculous. It becomes more and more ridiculous as you plod along. (I resisted the temptation to turn off the DVD twice).

Fay Grim requires an adventurous film-goer willing to tackle something that isn't cookie-cutter. In the end, it offers something that defies description.

26 Ossessione

Luchino Visconti's debut film, this Italian noir is generally credited with launching the Neorealist movement--well, it says so right on the back of the box--and is a sometimes penetrating, sometimes lugubrious portrait of lonesome individuals in moral flux. Set in Fascist Italy, an assortment of supporting characters--including an ingenuous drifter who espouses Communist virtues--embody the remote desperations of a country searching for its identity from without, drifting phantasms longing for a soul. Although Visconti's compassion for the disenfranchised and his ability to express their lamentable conditions was already well-developed, the spider web of deceit is tenuous--although a staple of noir is to posit a protagonist manipulated by fate and the femme fatale, Gino here is so unhinged to begin with that you fear he might deserve it--the cosmic irony too didactic, the illicit relationship strained with bathos. All the same, it's incisive and essential, although its actual impact on film history is certainly debatable.

Analysis:

We have filtered out low-frequency words because rare words are more informative. However, we were not able to adopt Ngram, and lexicons only allow us to analyze word by word instead of phases.

The occurrence of False Negative and False Positive

Many of the positive movie reviews were classified as negative even when the score rating was mentioned in the review. For example, review 1 and review 2 gave a score of 10/10 and 8/10. However, with the removal of punctuation and special characters, we were not able to classify the review based on scores. Moreover, those reviews contain negative words such as “bad”, “killed”, ignoring” and made the phase classified as negative. Thus, without a combination of words together, analyzing the single words would not allow the machine to understand the positive meaning of the sentence when negative words occur.

Statement of Collaboration:

Initial data exploration: All Members

Tuning Parameters & Variables: All Members, Shujiang Tan, Lucas Law

Result Analysis: All Members