CS598: Practical Statistical Learning
Project 3 Report: Movie sentiment analysis

Javier Huamani (huamani2@illinois.edu)
Sudha Natarajan (Sudha2@illinois.edu)

1. Contributions

Both teammates were equally involved in the exploration, data analysis, modeling, predicting and report creation of the movie sentiment analysis data; we had check-in meetings to share findings and updates on the progress of the model creation and the report creation steps.

2. Objective

We were provided with a dataset consisting of 50,000 IMDB movie reviews, where each review is labelled as positive or negative. In this project we build a binary classification model to predict the sentiment of a movie review and label it positive or negative. The final goal is to predict the sentiment of a review with vocabulary size less than equal to 1000. Using AUC as the evaluation metric, our performance target is to produce AUC equal to or bigger than 0.96 over all the five splits of test data.

3. Data Assessment and analysis

The initial dataset provided to us was alldata.tsv, which has 50,000 rows and 4 columns, and each row contains a movie review. The following are the 4 columns:

- Col 1: "id" the identification number.
- Col 2: "sentiment", 0 = negative and 1 = positive.
- Col 3: "score", the 10-point score assigned by the reviewer.
    - Scores 1-4 correspond to negative sentiment.
    - Scores 7-10 correspond to positive sentiment.
- Col 4: "review"

The following were the primary cleansing exercises that were undertaken for all Train/Test splits:

- Removal of html tags from "review"
- Lowercase "review"
- Remove "score"

An extensive process using t-tests and lasso regularization was utilized to reduce the overall vocabulary size, and therefore the DTM sizes. An html markdown file MyVocabGenerator.html was submitted to highlight the overall process.

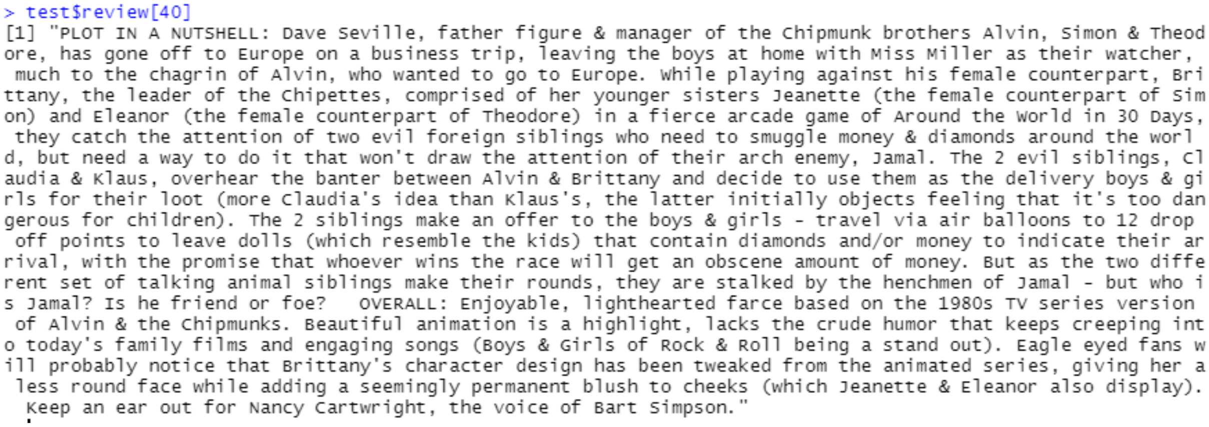
4. Approach

We followed the various approaches suggested by Dr. Liang. Our inputs included the comprehensive vocabulary of 981 terms (1-grams to 4-grams) and the training and test data per split. We created Document Term Matrices (DTMs) for each train and test split using the comprehensive vocabulary. The training DTM was used to train a binary logistic regression model with ridge regression (alpha of 0). We decided to use logistic regression because its effective with binary classification problems, its simplicity leads to improved performance and the size of our dataset (more entries than features) work best with it. We used cross validation to estimate an optimal lambda value which maximized the AUC. However, initially we were unable to meet the threshold of 0.96 AUC for some splits. We decided to use elastic net with an alpha term of 0.1 which preferences an L2 over L1 penalty. Using cross validation again, we were able to output lambdas which surpassed the threshold 0.96 AUC for all 5 splits.

5. Validation

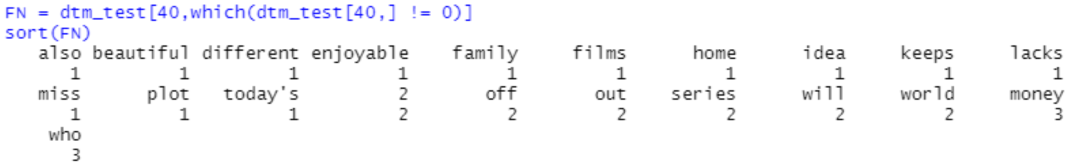
Despite reaching the threshold AUC for all 5 splits, it is important to study a subset of incorrect predictions in order to work through possible improvements for the task at hand. Two incorrect predictions, a False Negative Sentiment and False Positive Sentiment, were gathered for discussion.

5.1. False Negative Sentiment



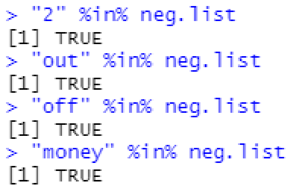
```
> test$review[40]
[1] "PLOT IN A NUTSHELL: Dave Seville, father figure & manager of the Chipmunk brothers Alvin, Simon & Theod
ore, has gone off to Europe on a business trip, leaving the boys at home with Miss Miller as their watcher,
 much to the chagrin of Alvin, who wanted to go to Europe. While playing against his female counterpart, Bri
ttany, the leader of the Chipettes, comprised of her younger sisters Jeanette (the female counterpart of Sim
on) and Eleanor (the female counterpart of Theodore) in a fierce arcade game of Around the World in 30 Days,
 they catch the attention of two evil foreign siblings who need to smuggle money & diamonds around the worl
d, but need a way to do it that won't draw the attention of their arch enemy, Jamal. The 2 evil siblings, Cl
audia & Klaus, overhear the banter between Alvin & Brittany and decide to use them as the delivery boys & gi
rls for their loot (more Claudia's idea than Klaus's, the latter initially objects feeling that it's too dan
gerous for children). The 2 siblings make an offer to the boys & girls - travel via air balloons to 12 drop
 off points to leave dolls (which resemble the kids) that contain diamonds and/or money to indicate their ar
rival, with the promise that whoever wins the race will get an obscene amount of money. But as the two diffe
rent set of talking animal siblings make their rounds, they are stalked by the henchmen of Jamal - but who i
s Jamal? Is he friend or foe?   OVERALL: Enjoyable, lighthearted farce based on the 1980s TV series version
 of Alvin & the Chipmunks. Beautiful animation is a highlight, lacks the crude humor that keeps creeping int
o today's family films and engaging songs (Boys & Girls of Rock & Roll being a stand out). Eagle eyed fans w
ill probably notice that Brittany's character design has been tweaked from the animated series, giving her a
 less round face while adding a seemingly permanent blush to cheeks (which Jeanette & Eleanor also display).
  Keep an ear out for Nancy Cartwright, the voice of Bart Simpson."
```

The review above shows one example of a False Negative Sentiment review for the 1st split. The reviewer in this case used ~75% of their review stating a summary of the movie plot prior to stating their opinion of the actual movie. The following is a subset of the test DTM which shows the words with non-zero counts for the False Negative Sentiment review:



```
FN = dtm_test[40,which(dtm_test[40,] != 0)]
sort(FN)
     also beautiful different enjoyable    family     films      home      idea     keeps     lacks 
        1         1         1         1         1         1         1         1         1         1 
     miss      plot   today's         2       off       out    series      will     world     money 
        1         1         1         2         2         2         2         2         2         3 
      who 
        3 
```

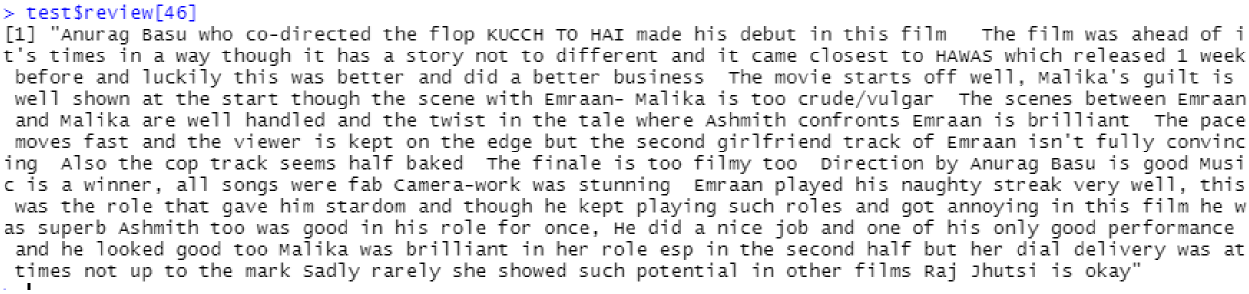
Words such as beautiful and enjoyable are positive words which have the lowest counts, possibly due to the fact that they were in the shorter opinion section of the review. However, the following words in the DTM were all found within the negative terms list:



```
> "2" %in% neg.list
[1] TRUE
> "out" %in% neg.list
[1] TRUE
> "off" %in% neg.list
[1] TRUE
> "money" %in% neg.list
[1] TRUE
```

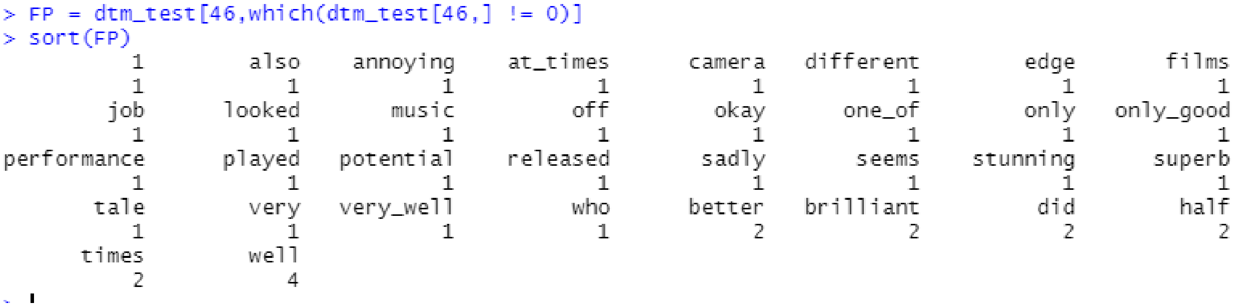
The words “2”, “off” and “money” were all found within the plot summary. Numbers have been found within both the negative and positive words list as a means for reviewers to indicate their own numeric rating of a movie. However, in this case the “2” was an arbitrary number used within the summary. It is possible that without the initial irrelevant summary section, the binary classifier would’ve classified the review as positive instead of negative

5.2. False Positive Sentiment



```
> test$review[46]
[1] "Anurag Basu who co-directed the flop KUCCH TO HAI made his debut in this film   The film was ahead of i
t's times in a way though it has a story not to different and it came closest to HAWAS which released 1 week
 before and luckily this was better and did a better business  The movie starts off well, Malika's guilt is 
 well shown at the start though the scene with Emraan- Malika is too crude/vulgar  The scenes between Emraan
 and Malika are well handled and the twist in the tale where Ashmith confronts Emraan is brilliant  The pace
 moves fast and the viewer is kept on the edge but the second girlfriend track of Emraan isn't fully convinc
ing  Also the cop track seems half baked  The finale is too filmy too  Direction by Anurag Basu is good Musi
c is a winner, all songs were fab Camera-work was stunning  Emraan played his naughty streak very well, this
 was the role that gave him stardom and though he kept playing such roles and got annoying in this film he w
as superb Ashmith too was good in his role for once, He did a nice job and one of his only good performance 
 and he looked good too Malika was brilliant in her role esp in the second half but her dial delivery was at
 times not up to the mark Sadly rarely she showed such potential in other films Raj Jhutsi is okay"
```

This particular review was confusing because it gave praise to the actors, musical score and other minor aspects of the movie. However, it also criticized the actual score and the ground truth states that overall this is a negative sentiment review. The following is a subset of the test DTM which shows the words with non-zero counts for the False Positive Sentiment review:



```
> FP = dtm_test[46,which(dtm_test[46,] != 0)]
> sort(FP)
          1        also    annoying    at_times      camera   different        edge       films 
          1           1           1           1           1           1           1           1 
        job      looked       music         off        okay      one_of        only   only_good 
          1           1           1           1           1           1           1           1 
performance      played   potential    released       sadly       seems    stunning      superb 
          1           1           1           1           1           1           1           1 
       tale        very   very_well         who      better   brilliant         did        half 
          1           1           1           1           2           2           2           2 
      times        well 
          2           4 
```

Positive words like “well”, “better”, “brilliant”, “stunning”, etc. are shown to have the highest counts in the DTM. The reviewer used more impactful words more often when discussing the praise in the review. It is understandable that the classifier would’ve incorrectly assigned the review with a Positive sentiment due to the ambiguity of the review.

6. Results

The following table shows the AUC we obtained for each of the 5 splits:

Split	AUC
Split 1	0.9629556
Split 2	0.9623818
Split 3	0.9617547
Split 4	0.9624445
Split 5	0.9617127

The overall average AUC for all 5 splits is: **0.96224986**

The total running time for all 5 splits on Windows 10 PC, 16 GB Ram, Intel Core i5-8350U CPU system is: **7.15 minutes.** The average running time for each split is: **1.43 minutes.**

7. Conclusion

We got to work in a real-life use case with a practical movie reviews dataset to predict the sentiment of movie reviews. Using Professor's code with the various approaches helped us walk through and understand the different steps, trial and error process and tuning to the desired outcome. We can envision various real life use cases that this process might apply to and are confident about the approach and the process to follow.

Validating our results by taking a deep dive at the data helped us understand some possible reasons why our classifier made some mistakes despite surpassing the threshold AUC. The deep diver made it clear that ambiguity and irrelevant information can have a detrimental effect on the performance of the classifier.

8. Acknowledgement

- Professor Dr. Liang's code and approaches from various posts on Campuswire
- Sample project reports posted on Campuswire by Dr. Liang.