Project 3: Movie Review Sentiment Analysis

Team MAK: Mriganka Sarma (ms76), Ajay Menon (kamenon2), Kai Pak (kaipak2)
CS598/STAT542 – Practical Statistical Learning (PSL), Spring 2021

1. Introduction and Goal

In this project, we are presented with an IMBD dataset containing 50,000 labeled movie reviews that includes the written review, the reviewers' 1-10 score, and the general sentiment (positive/negative). The goal is to build a model that can predict the sentiment of reviews based on only the text itself using natural language processing (NLP) and statistical modeling techniques. The highwater goal of this project is to build a binary classifier with a vocabulary size of less than or equal to 1000 words achieving an accuracy measure of **≥ 0.96 AUC (Area Under Curve)** on all five test datasets. Utilizing linear-based learners we are able to successfully beat this benchmark.

2. Dataset Description

The dataset consists of 50,000 labeled movie reviews with the following columns:

ID	Name	Description
1	id	Identification number of the record.
2	sentiment	0/1 – The binary target prediction.
3	score	1-4 (neg), 7-10 (pos). Dataset does not contain 5-6.
4	review	The actual text reviews.

Note, that the reviews scores are in a generally bimodal distribution: they score either very negatively or positively. This likely aids in predictive tasks such as this since the reviews will tend to be quite opinionated. Such sentiment analysis tasks can be modeled with a relatively limited set of vocabulary such as with the use of sentiment lexicons (Wilson 2004).

Using the project provided script, we split the data in the following way:

- First read the complete dataset "alldata.tsv".
- Now using the given train/test split row numbers in file "splits_S21.csv", split the dataset into five sets each containing train and test splits as below:
  - "train.tsv" – doesn't contain the "score" column
  - "test.tsv" – contains the "id" and "review" columns
  - "test_y.tsv" – contains the "id", "sentiment", and "score" columns
- Each of these sets are stored in a subfolder specific to that split.

The data is split so that 25,000 observations are train leaving 25,000 for test.

3. Data Cleaning

General NLP tools such as word2vec, and regular expressions have been used to clean and prepare the data for machine learning algorithms. Since these reviews come from HTML pages, an initial step is to remove any tags, e.g.
 etc. using regular expressions. Then, with word2vec, we further process the text by lower-casing and removing special characters before vectorizing the processed tokens to form the preprocessed vocabulary input for the models (more details to follow). It also generates the sparse matrix used to represent the test and train datasets. Each row represents a review and each column the count of the word/term in the review through the vocabulary input.

4. Vocabulary Construction and Model Selection

We tried different approaches to build a model for this task by combining different approaches for vocabulary construction with different model selection for the classification task. All those approaches are shown in the below table:

Table 1: Approaches

Approach #	Train Data	Model Selected	Vocab Size	Filtered Vocab
1	Split_1	Ridge Regression	961	Y
2	All Data	Ridge Regression	992	Y
3	All Data	Ridge Regression	976	N
4	All Data	XGBoost	992	Y
5	All Data	XGBoost	976	N

The vocabulary is initially generated by producing n-grams ranging from one to four tokens. This results in a preprocessed vocabulary containing over 30,000 unique tokens. Using a bag-of-words model, this means there would be potentially greater than 30,000 predictors for any model we could train which would make interpretation quite difficult.

In some of our approaches (1, 2, 4), we have filtered this list of 30,000+ words into the top 2000 positive and negative words by using the magnitude of their t-statistics derived from the "two-sample t-test". Following equation has been used for the "two-sample t-test" for testing population of positive and negative sentiments.

$$\frac{(\bar{X}-\bar{Y})}{\left(\sqrt{\frac{s_X^2}{m}+\frac{s_Y^2}{n}}\right)}$$

where s^2_* denotes the sample variance of X and Y, and m and n the size of their individual population. Based on the 2000 words generated by following t-statistics value, we have observed that negative words have a higher distribution (~1257 words) compared to positive words (~743) while considering split_1 train dataset. Similar distribution is observed for alldata.tsv as well.

As it can be seen from the table 1, in our first approach, we have used the data from the first split only as the training data, whereas for all the remaining approaches we have used the full dataset as the training data. Then we have used different permutations of training data, filtered words, and classification model to evaluate our prediction on the test data.

5. Model Validation

5.1. Model Performance

With each of the approaches listed in table 1, we were able to beat the benchmark of AUC ≥ 0.96. The below table lists the performance of each of these approaches.

Approach #	Split_1	Split_2	Split_3	Split_4	Split_5
1	0.9601	0.963	0.963	0.9635	0.9626
2	0.9655	0.9638	0.9648	0.9646	0.9638
3	**0.9693**	**0.9685**	**0.9688**	**0.9694**	**0.9682**
4	0.9627	0.9626	0.9626	0.9629	0.9619
5	0.9671	0.9673	0.9669	0.9679	0.9664

Approaches 1, 2, and 3 used logistic regression model for the classification, whereas approaches 4 and 5 used XGBoost for classification. We were able to get high AUC scores over all the test splits using the simple logistic regression model for classification as listed in approach #3 in table 1. Since the approach #3 produced the best results, so we selected this approach as our final model for the task.

5.2. Vocabulary Size

Approach #3 helped us reduce our vocabulary size to **976**. Although performing the "two-sample t-test" helped in getting words that are more relevant with sentiment for approach 1, we have observed that we were able to get interpretable sentiment words by using the full dataset as the training data even without filtering the words based on t-statistics values. Moreover, when the full data was used as training data, this filtering actually reduced the AUC scores over all the test splits. So our approach #3 uses all data as training data without any t-statistics based filtering and by using Lasso regression reduces the final vocabulary size down to 976.

5.3. Model Limitation

The model performance gets limited due to wrong sentiment and scores assigned to a review. For example, ID's 15207 and 49333 have negative reviews, but were labelled as sentiment 1 and score 10 indicating them to be positive reviews. Thus, model prediction could worsen because of such invalid records in dataset.

There are also reviews in the input data which lack any sentiment relevant words. For example, the below review text has a positive sentiment and rating score 8. However, due to the lack of any indicative sentiment words, our model has classified this review as a negative sentiment with probability score "**0.431642882314865**".

> I know what you're saying, \Oh man, Pinochio is not scary!\" but this movie goes beyond alot more than a maniacal pinochio. Behind it tells the story of a mother and her daughter who is oddly attached to her doll Pinnochio who seems to talk to her. The only weird thing is that noone else can hear the doll except her. In the end is shocking revelation that, as did I, will shock you. Watch it. Give it a try."

5.4. Error Analysis

The following review text recommends a movie in a sarcastic way, although has given score 1. But due to the presence of positive words like "highly recommend", the model misclassified this review as a positive review with moderately high probability score **"0.569755768079311"**. So, apparently the model is not able to handle sarcasm very well.

> I would **highly recommend** seeing this movie. After viewing it, you will be able to walk out of every other bad movie EVER saying \at least it wasn't The Omega Code.\"

Forget my money, I want my TIME back!"

6. Tech Spec

The model was run on the following machine:
MacBook Pro (15-inch, 2017) with 2.8 GHz, 4-core, Intel Core i7 CPU, 16 GB Memory, SSD.

The following table lists the overall computation time and the computation times for each of the splits in seconds using the approach #3.

Split_1	Split_2	Split_3	Split_4	Split_5	Overall
52.279	51.743	51.255	51.126	51.781	258.184

7. Interpretability of the Algorithm

7.1. Positive Review

Here we are presenting a list of the top 25 positive words/terms with their beta values:

Table 2: Positive terms with their beta values

	Term	Beta
1	7_10	1.73073653923649
2	well_worth	0.988981697097815
3	7_out	0.943624061123603
4	8_10	0.933834416438047
5	definitely_worth	0.889425640300297
6	must_see	0.820110871011616
7	refreshing	0.778163023013647
8	10_10	0.772708500731569
9	excellent	0.673109848093651
10	9_10	0.671159766473525
11	pleasantly_surprised	0.58864625566627
12	if_don't_like	0.586983541415454
13	superb	0.585539199031047
14	wonderfully	0.57147099155154
15	highly_recommend	0.562706393182768
16	gem	0.529333852723964
17	wonderful	0.507730239765961
18	amazing	0.499302368624268
19	loved_this	0.497267440601616
20	beautifully	0.48873409327531
21	favorite	0.473183551602298
22	perfect	0.469695868152225
23	funniest	0.468769380948661
24	7_out_of	0.460499624504575
25	8	0.452924090293459

Since we have selected the logistic regression model for our classification task, our final prediction is a linear combination of these terms. For example, we picked the top rated positive review from our classification results for evaluation of the interpretability of our chosen model. The top positive review has the highest probability score **"0.999999999983402"** among all the positive reviews. The review text is shown below with the positive words highlighted that have contributed towards the high probability of this review.

> The film has the **best** cinematography of the bunch, mainly because it is in a **stunning** black and white. The segment is dreamlike and **beautiful**. **7/10**.

Jean-Luc Godard - \"Armide\" - I chose to brave this much-maligned film for the Godard and Altman segments. With Godard, I was much more **impressed** than I thought I would be. I can't claim to have seen all that many of his films since he made so many that almost no one has seen, but, judging from what I have seen, this may be his **best** work since the 60s. It is the **funniest** segment in this film, and the most artistically accomplished. Bravo, Jean-Luc! **9/10**.

Julien Temple - \"Rigoletto\" - a **very funny** segment, it is also quite predictable. Still, this story about a husband and wife who are cheating on each other at the same resort is **wonderfully** filmed with long, complex tracking shots that depend on precisely timed choreography from the actors. It also has a great self-referencing joke about omnibus films themselves. The final scene is very weak. **7/10**.

Bruce Beresford - \"Die tote Stadt\" - this short segment involves too lovers in (I think) Venice. It is pretty, with some **nice** shots of doves flying about the city. It is slight, but **nice**. **7/10**.

Robert Altman - \"Les Boréades\" - not one of the better segments, unfortunately, this is more of a music video than a concept short film. It involves the occupants of an insane asylum attending a theatrical performance. The music and images work **well** together, so at least I can give it credit for being a **good** music video. **7/10**

Franc Roddam - \"Liebestod\" - somewhat unfortunate for Beresford's segment, this segment is very similar to it. As you might assume from my phrasing, this one struck me much more. It is about a young man and his girl going to Las Vegas on a fatalistic voyage. **8/10**.

Ken Russell - \"Nessun dorma\" - maybe the most visually striking segment, it plays in a fantasy world more than in reality. It is a **beautiful** tale of a fallen angel. **8/10**.

Derek Jarman - \"Depuis le jour\" - I have heard a lot about Jarman, and this is the first piece of filmmaking I have seen from him. Hopefully, I'll see more in the future. This one is also music-videoish, but it is better than Altman's segement. It mainly concerns an old woman remembering her younger days. The editing and the use of different film stocks to represent both time and emotion are very **beautiful**. **8/10**.

Bill Bryden - \"I pagliacci\" - the sad clown, possibly one of the most famous arias (particularly memorable from an episode of Seinfeld), this serves as the material separating each segement and the finale. It is **simple** and **effective**. **7/10**.

Overall, I give it a solid **7/10**. It isn't anywhere near as bad as you've heard."

As we can see from the list of positive terms in table 2, the terms "7/10", "8/10", "9/10" have very high weights. This review text has many appearances of these terms which have contributed heavily towards assigning a very high probability to this review. Apart from these top 25 positive words/terms, there are many other positive words in the review which contribute towards correctly classifying this as a positive review.

7.2. Negative Review

Here are the top 25 negative terms from our model.

Table 3: Negative terms with their beta values

	Term	Beta
1	4_10	-1.66975011793272
2	4_out_of_10	-1.30493830809791
3	3_10	-1.23806517048512
4	waste	-1.20741646434771
5	3_out_of_10	-1.20649611094279
6	1_2_from	-1.16104537144591
7	worst	-1.0423891229862
8	not_recommend	-0.996541688086449
9	grade_d	-0.952269113314269
10	2_10	-0.901485054584791

11	awful	-0.887109659356933
12	wanted_to_like	-0.82102390775355
13	not_worth	-0.820400358504664
14	poorly	-0.789723888614148
15	disappointment	-0.770731477917552
16	skip_this	-0.698531410984345
17	laughable	-0.672888300637111
18	miscast	-0.653891149487988
19	redeeming	-0.648777382272104
20	had_high	-0.642978696296286
21	not_funny	-0.638554318494171
22	forgettable	-0.63051048866905
23	fails	-0.628640572071134
24	2_out_of_10	-0.625314721175724
25	dull	-0.617924732485417

The script by Fragasso is an absolute **mess**, none of it is well thought out & is just as stupid as it gets. The scenes of zombie birds attacking people are not only technically inept but the whole idea is just **absurd**. The zombies themselves have no consistency whatsoever, look at the scene where Patricia is on the bridge & the zombies are slow as they shuffle along but then look at the scene earlier on where she was attacked by the zombie with the machete because that one runs around like it's on steroids, then for no reasonable explanation about 10 minutes before the film finishes the zombies suddenly develop the ability to speak which also looks daft. There are so many things **wrong** with Zombi 3, scene after scene of **terribly** thought out & **ineptly** directed action, **awful** character's & really **dull** broken English dialogue which doesn't make sense half the time. Then there's the **embarrassing** scene where the zombie head inside the fridge suddenly develops the ability to fly through the air & bite someones neck, the scene when the guy's in white contamination suits at the end are about to kill Kenny & Roger but instead of using their automatic rifles they decide to try & kill them by hand, even when Kenny picks up a gun himself they still refuse to use their rifles & when Kenny starts to shoot them all they still refuse to use their rifles & it's one of the most **ineptly** handled scenes ever put to film & then there's the end where Kenny takes off in the helicopter but can't rest it down on the ground for literally a few seconds to pick his buddy up & then a load of zombies suddenly spring up from under some piles of grass, what? Since when did zombies hide themselves yet alone under piles of grass? This all may sound 'fun' but believe me it's not, it's a really **bad** film that is just **boring**, repetitive & simply doesn't work on any level as a piece of entertainment except for a few unintentional laughs.

It's hard to know who was responsible for what exactly but none of the footage is particularly well shot. It has a bland lifeless feel about it & for some reason the makers have tried to bath every scene in mist, the problem is they clearly only had one fog machine & you can see that at one corner of the screen the mist is noticeably thicker as it is coming straight out of the machine & thinning out as it disperses across the scene. Since a lot of it is set during the day it doesn't add any sort of atmosphere whatsoever & when they do get it right & the mist is evenly spread across the screen it just looks like they shot the scene on a foggy day! The direction is **poor** with no consistency & it just looks & feels bottom of the barrel stuff. Even the blood & gore isn't up to much, there's a gory hand severing at the start, a scene when something rips out of a pregnant woman's stomach, a legless woman (what actually took her legs off in the pool by the way & why didn't it take the legs off the guy who jumped in to save her?) & a few OK looking zombies is as gory as it gets. For anyone hoping to see a gore fest the likes of which Fulci regularly served up during the late 70's & early 80's will be **very disappointed**, there aren't any decent feeding scenes, no intestines, no stand out 'head shots' & very little gore at all.

Technically the film is **poor**, the special effects are **cheap** looking, the cinematography is **dull**, the music is **terrible**, the locations are **bland** & it has rock bottom production values. This was actually shot in the Philippines to keep the cost down to a minimum. The entire film is obviously dubbed, the acting still looks **awful** though & the English version seems to have been written by someone who doesn't understand the language that well.

Zombi 3 is not a sequel to Fulci's classic zombie gore fest Zombi 2 (1979), it has nothing to do with it at all apart from the cash-in title. I'm sorry but Zombi 3 is an amateurish **mess** of a film, it's **boring**, it makes **no sense**, it's **not funny** enough to be entertaining & it lacks any decent gore. One to **avoid**.

To explain the negative reviews predicted by our model, we picked the second worst review with a probability score **"1.57530866467623E-09"**. As we can see in the above review text, there are some of the top 25 negative words, e.g. "awful", "dull", present in this review. The review also contains many other negative words which are not in the top 25 but have significant negative weights, such as "mess", "absurd", "poor", "terrible", "wrong", "inept", "bad", "boring", "very disappointed", "no sense", "not funny", "avoid" etc. So linear combination of all of these negative weighted terms have contributed towards rating this review as the 2nd most negative review classified by our model. Intuitively also, we can see the direct correlation between the negative words/terms and the negative classification of this review.

8. Conclusion and Future Direction

We made a few interesting observations during the development of our model.

We expected to see improvements in the AUC scores by increasing the stop words and by stemming the vocabulary. We tried to increase the stop_word count by using stopword ("en") library and enable stemming using the SnowBallC package. However, we saw a negative effect on the final AUC scores. Hence this approach was not included.

Reviews had high counts of
 html tokens and were removed. Some other special characters like \" were also present but removing them using gsub('[[:punct]]') again did not seem to improve the AUC scores.

If a word like "surprising" or "truly" exists and a positive sentiment (1) is assigned to it during training, then it's possible that the model predicts inaccurately during test since the contextual-words are not really taken into account. Basically, since our model training lacks "part-of-speech" tagging, misclassification can happen. A possible remedy to this could be using additional parameters like "skip_grams_window_context" and "skip_grams_window" which can possibly help reduce misclassification. A future direction would be to explore these parameters and the effect of their inclusion on the performance of the model.

9. References

F. Liang, R code and posts on Piazza, CS598 PSL, Spring 2021 UIUC.

T. Wilson, J. Wiebe, P. Hoffman. Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. ACM. 2005.

https://cran.r-project.org/web/packages/text2vec/vignettes/text-vectorization.html