DS-GA 1011 Fall 2018
Bag of N-Gram Document Classification

Shizhan Gong, sg5722

October 7, 2018

The goal of this assignment is to build a bag of n-grams model for predicting the sentiment of the movie reviewers given the textual review for the movie. I first built a baseline model which is the modification of the model from the lab session on Bag of Words model. Then I did ablation study to test the effect of some hyper parameters and compare it with baseline models. Finally, I extended the model to predict the rating of the review.

1 Baseline Model

In the baseline model, a simple network which only contains an embedding layer and a linear layer are used. 1-gram model is used and the embedding size set to be 200. Vocabulary size is 10,000 and each document is padded to contain 300 words. Adam methods with learning rate equals to 0.001 are used to train the model. We randomly split the train data set into 20,000 train examples and 5,000 validation examples. I used train set to train the model and use validation set to determine the stopping point and compare models. The batch size is set to be 50.

2 Tokenization Schemes

I tried the following three tokenization schemes. First I tried to tokenize without any other pose-processing expect for lowercasing. Then I tried to remove punctuation after tokenization. Finally, I tried to remove both punctuation and stop words. I used *spacy.lang.en.stop_words.STOP_WORDS* as the list of stop words. The results are given in the Table 1.

Table 1: Model performance of different tokenization schemes

tokenization schemes	train loss	val loss	val acc
remove nothing	0.2492	0.2932	0.8722
remove punctuation	0.2606	0.2949	0.8778
remove punctuation and stopwords	0.2127	0.2854	0.8822

From the result above we can see that when we remove both stopwords and punctuations, the performance of the model is the best. Therefore, in the following analysis, we choose to remove both stopwords and punctuations.

3 Model Hyperparameters

Then I try to explore the effect of model hyperparameters on the performance of the data. I first try to use different n for n-gram. When building vocabularies, I first build n-grams without removing punctuation and stopwords, and then remove them from the single gram. Since now each documents contains both single words and n-grams, the maximum length of a document will become larger accordingly. Other hyperparameters remaining the same, we obtain the results in Table 2. From the result we can see that 2-gram performs the best. Actually, since we have limited the size of the vocabulary, only very few 3-gram and 4-gram are incorporated into the vocabulary. This is reasonable since only a few meaningful phrase are larger than three words. Therefore we choose to apply 2-grams.

Table 2: Model performance of different n-grams

n-gram	sentence length	train loss	val loss	val acc
1-gram	300	0.2112	0.2877	0.8842
2-gram	600	0.2019	0.2666	0.8920
3-gram	900	0.2133	0.2672	0.8906
4-gram	1200	0.2242	0.2744	0.8876

Then I changed the vocabulary size from 10,000 to 25,000 to see how the results respond to the changes. The results are shown in Table 3. The results also show that within the given range, with lager vocabulary size, the performance of the model becomes slightly better. Therefore we set the vocabulary size to be 25,000.

Table 3: Model performance of different vocabulary size

vocabulary size	train loss	val loss	val acc
10,000	0.2019	0.2666	0.8920
15,000	0.2024	0.2631	0.8950
20,000	0.1807	0.2591	0.8968
25,000	0.1721	0.2504	0.9014

Finally, the effects of embedding size to the model performance is also tested. The results are show in Table 4. We can see that the performance of the model with embedding size equaling to 400 is better than that of the model with embedding size equaling to 200. However, if we continue increasing the embedding size, we can no longer see a significant lift of the model performance in the validation set. Therefore, we decide the embedding size is 400.

Table 4: Model performance of different embedding size

embedding size	train loss	val loss	val acc
200	0.1721	0.2504	0.9014
400	0.1747	0.2505	0.9040
600	0.1615	0.2620	0.9041
800	0.1490	0.2453	0.9042

4 Optimization Hyperparameters

Afterwards, I try to use different optimization algorithm to train the model. Specifically, I compared SGD with Adam. The results show that Adam performs much better than SGD besides, the convergence speed of Adam is also much faster than SGD.

Table 5: Model performance of different optimization algorithm

algorithm	train loss	val loss	val acc
Adam	0.1721	0.2504	0.9014
SGD	0.1886	0.2994	0.8840

Then I tried different learning rate for the model. The result is shown in Table 6. When I set a relatively large learning rate such as 0.01, the model suffers from serious overfitting after only one epoch of training. When I set a small learning rate 0.0001, however, it takes much longer time for the model to reach its optimal point. Therefore, I incorporated linear annealing rate. Nevertheless, there is no significant improvement in terns of the model performance in the validation set.

Table 6: Model performance of different learning rate

learning rate	train loss	val loss	val acc
0.01	0.3468	0.2532	0.8950
0.001	0.1747	0.2505	0.9040
0.0001	0.1335	0.2483	0.9028
0.001-0.00005*epoch	0.1225	0.2457	0.9024
0.001-0.0001*epoch	0.1234	0.2455	0.9032

5 Network Structure

One of the widely used approaches to lift the performance of neural network is to increase the number of layers of the network. In this case, a add a hidden layer to the classification layer, and try to use different activation functions for this layer. The results are shown in the Table 8.

We did not see any improvement of the results however. Meanwhile, we observed serious overfitting at the early stage of the training since now the model is more complex. One way to mitigate overfitting is to incorporate dropout mechanism into the network. Therefore I

Table 7: Model performance of different activation function

activation function	train loss	val loss	val acc
Softmax	0.1325	0.2615	0.9016
ReLU	0.1597	0.2650	0.8972
TanH	0.1509	0.2872	0.8900

add an dropout mechanism at the output of the hidden layer and tested the effect of different dropout probability. The results are given in Table 8

Table 8: Model performance of different dropout rate

dropout probability	train loss	val loss	val acc
0.5	0.1996	0.2727	0.9016
0.4	0.2600	0.2643	0.9022
0.3	0.3283	0.2872	0.9000
0.2	0.2354	0.2644	0.9004

Still, we did not see significant improvement. Therefore in this particular case, increasing the model complexity does not contribute to more accurate results.

6 Result

Finally, we achieve the best model whose accuracy on the validation set is 0.9040. The detailed hyperparameters are given in Table 9. We then tested the model on the test set and obtain an accuracy of 0.8951.

Table 9: Model performance of different dropout rate

n-grams	2
vocabulary size	25,000
embedding size	400
optimization algorithm	Adam
learning rate	0.001
hidden layer	none

In addition, we also extend our model to predicting the rating of the review. With the same hyperparameter, we achieve the accuracy of 0.4420 in the test set. The result is not very satisfying. We also consider the top-3 accuracy, which is 0.6912. Therefore, it is not a effective model to use comments data to predict the score. This is reasonable since different people have different language style. Some people may tend to use some exaggerated words to outspeak their emotion while other people prefer to comment in a more rational way. Also sometime the person himself may have difficulties telling between a score of 7 or 8. It is easy to distinguish between positive and negative, but it is hard to use words to quantify how each person feel about the movie.

A Appendix: github repo

https://github.com/670973787/Bag-of-N-Gram-Document-Classification

B Appendix: correct & incorrect cases in the validation set

correct

- (positive) This movie is brilliant. The comments made before is from someone who obviously doesn't get it. The movie is campy- yes! But it is uplifting and fun. This movie is an underground hit and brings comparisons to Absolutely Fabulous. It is a must see!.
- (negative) If there was a God, he would have made sure this movie stayed in the toilet were it was crapped up. This is BY FAR the worst vampire movie I have ever seen. I may never watch a vampire film again because of this movie. It makes Zombie Lake look like The Sound of Music.
- (negative) Some 25 year olds behave like teenagers, coping with the death of a high-school mate, trying to find their purpose in live and love. The script is so lame that I had to force myself to even finish this movie. Stay away from it. 1/10.

incorrect

- (negative) It is enjoyable and fast-paced. <br / ><br / >There is no way on Earth that the actor playing Mat could be eighteen. However, the main thing is that he does act eighteen very convincingly. It must be a credit to his audition that he convinced them to cast him. I quite soon accepted him as being a naive young country boy.<br / ><br / >While his was the best performance, most of the others were also very engaging. In particular, the interplay between the policemen was natural and well-balanced, and worked very well.<br / ><br / >It is only about 45 minutes long, so the plot is not complex. More key is the style of the whole thing. It is very slick and vibrant, and the backdrops are atmospheric, especially from the fact that all the colours are extremely rich. The gangland is identifiable to foreign audiences, but still manages to be distinctly Australian.
- (positive) *****Spoilers herein*****<br / ><br / >What really scares you? Killer sharks, or maybe ghosts trying to bring back a message? Maybe a chainsaw wielding psychopath?<br / ><br / >Maybe. But those fears dont́ even compare to a horror which people dare not even speak of or consider–and that is the death of oneś own child. ”Pet Sematary” taps this base, primal adult fear, and then takes it to places that most could not bear to explore.<br / ><br / >Iv́e read comments about this film that include poor acting, characters making stupid decisions, etc. I disagree. The acting is actually first rate for a film like this. Maybe it is impossible for many to imagine the desperation resulting from such a scenario. But the filmś events are not

only logical, they may be absolutely inevitable if such a scenario were possible. This is the true horror of "Pet Sematary": It isnt́ that pets and people come back from the dead as evil killers who hunt with knives and scalpels, it is that anyone who has lost a child could become so desperate as to commit the crimes that Louis Creed does. Despite warning, or even past history.<br / ><br / >The movie takes those willing to go with it to the depths of a desperate human heart. The heart of a protector trying to make up for not being able to protect. And the results are horrifying. In fact, when the film dives into slasher territory near the end, itś almost a letdown, although I believe itś perfectly logical how it got there.<br / ><br / >I am a true horror fan, and I contend that this is one of the scariest horror films ever made. If you dont́ think so, see it again after you have children.

- (negative) I don't understand people. Why is it that this movie is getting an 8.3!!!!!!???? I had high hopes for this movie, but once i was about a half hour into it I just wanted to leave the theater. In the vast majority of the reviews on this site people are saying that this is one of the best action movies they've seen (or of the summer, year, etc.) They say it's an excellent conclusion. WTF!!!!!!!!!?????? What has been concluded (besides the fact that Bourne can ride motorcycles, shoot, and fight better than anyone else he comes across)? What do you learn about Bourne's character in this movie?????????Absolutely f****** nothing!!!!!!! Okay, there's a lot of action, but what's so great about the action in this movie?? I don't like the cinematography and film editing. The shaky camera effect and fast changing shots were used TOO much and they get old fast (I didn't mind them in Supremacy because it was still easy to follow and was not used in excess) and made me quite dizzy. I was quickly wishing I had saved my $$$ for something else.<br / ><br / >This movie has no plot. All this movie is is a 115 minute chase seen. Bourne, who you learn absolutely nothing about in the entire 115 minutes of the movie, is a perfectionist at everything he attempts. There is absolutely no character development in this movie, you know nothing about anyone, and there is a wide array of new characters that are introduced in this installment. Some people said that this movie has incredible writing and suspense. ???????????!!!!!!!! What writing???? What suspense??? There's no suspense. Bourne is so perfect at doing everything he does, I don't think he has anything to worry about. If this is the best movie of the year 2007 I may just quit watching movies entirely!!!! <br / ><br / >Many people have also said that Matt Damon's performance in this movie is one of the best (if not the best) of his career. What performance?? How many lines did he have in this movie??? I have some respect for Damon because he has been in movies that I liked and has played different kinds of characters, but a good actor is someone that you can barely recognize from one movie to the next, someone who chooses different types of roles. Not someone who plays the same roles over and over again (which Damon doesn't do, but an example of someone who does is Vin Diesel).<br / ><br / >Anyways, this movie was a BIG disappointment to me. I do not recommend this movie but I do recommend the first two (Bourne Identity and Bourne Supremacy) and I most definitely recommend reading the three books (which are much different then the movies).

C Appendix: Training Curves

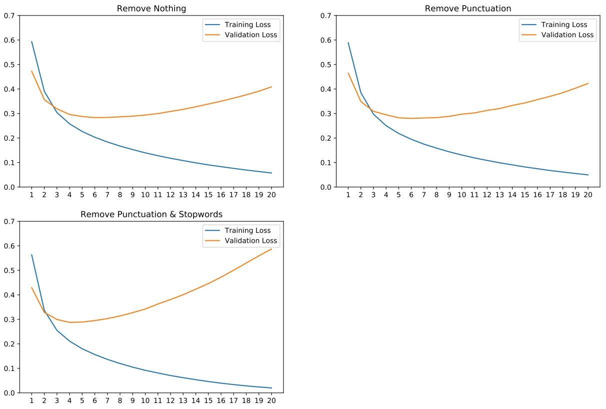

Figure 1: Tokenization Schemes.

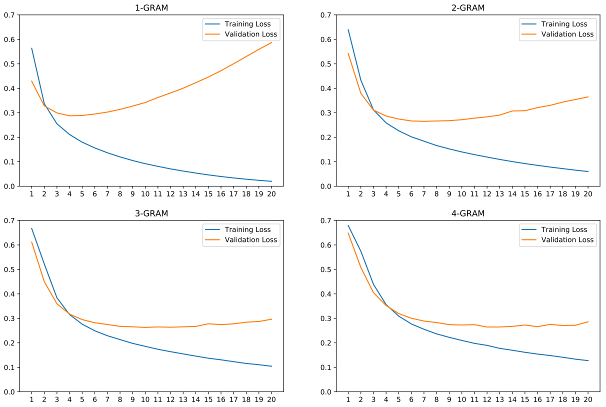

Figure 2: N-grams.

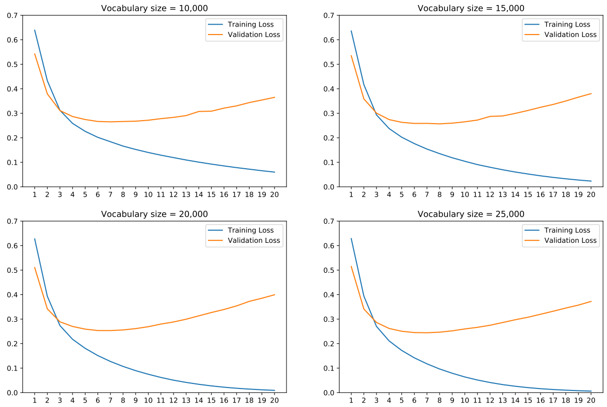

Figure 3: Vocabulary Size.

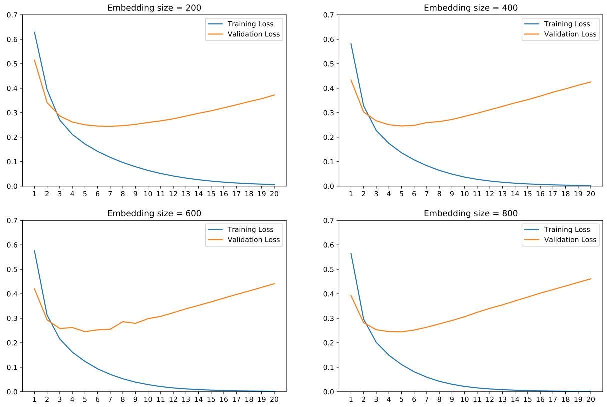

Figure 4: Embedding Size.

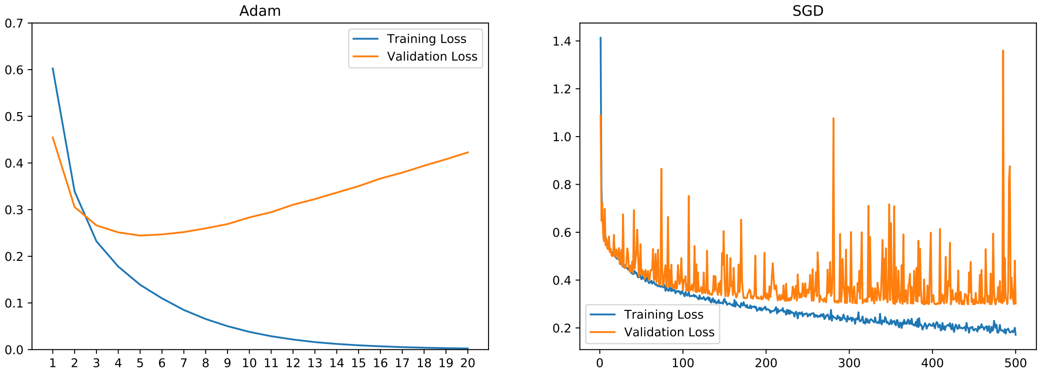

Figure 5: Optimization Algorithm.

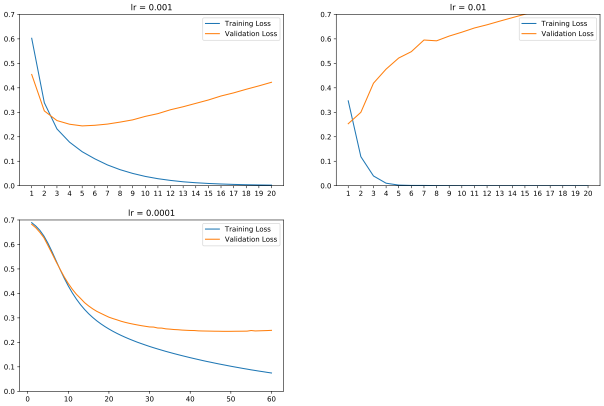

Figure 6: Learning Rate.

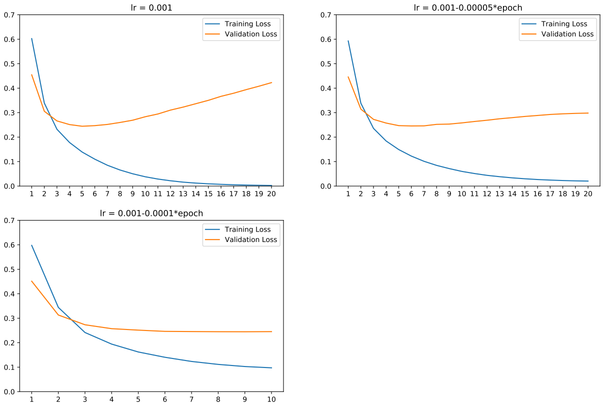

Figure 7: Linear Annealing Rate.

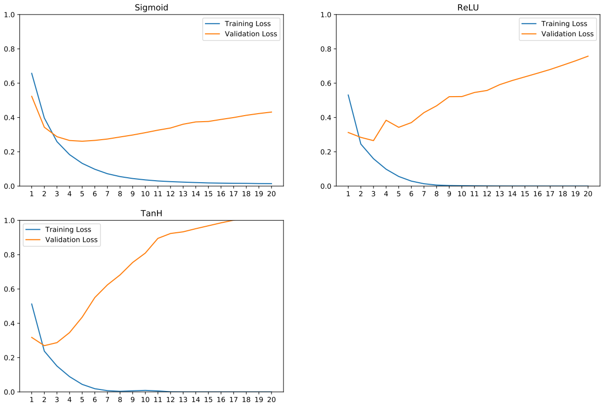

Figure 8: Activation Function.

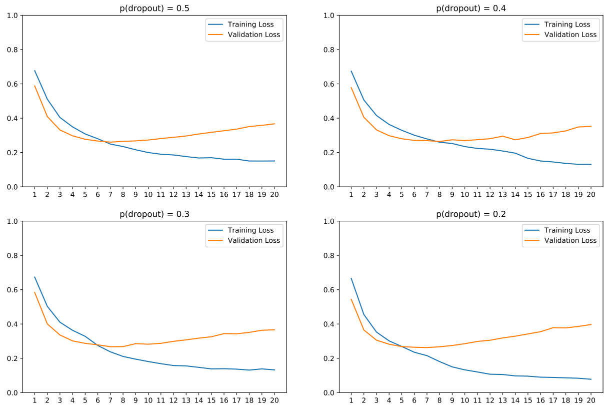

Figure 9: Dropout Probability.

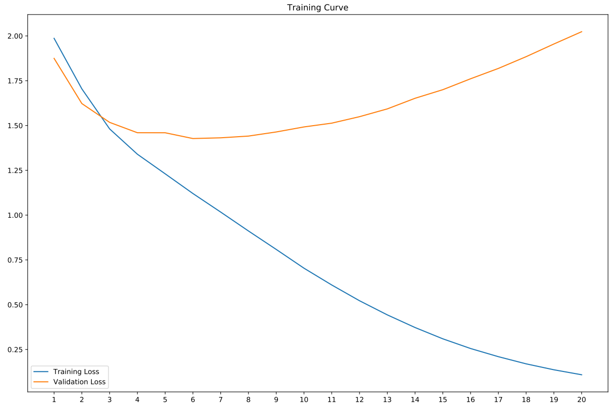

Figure 10: Predicting Score.