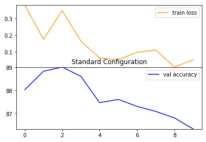
IMDB Reviews Sentiment Analysis Ablation Study

Manwen Li

https://github.com/manwenli/DS-GA-1011

0.1 Default Configuration



ID	N-Gram	Max Vocab size	Remove Punctu-ations	Remove Stop Words	Remove HTML Tag	Optimizer	Learning Rate	LR Linear Decay	Size of Embedding	Best Epochs	Best Validation Accuracy
0	4	10k	Yes	Yes	Yes	Adam	0.005	No	200	2	88.98%

Figure 1: Default configuration

The best validation accuracy from the bag of n-gram classifier with default configuration of the parameters is 88.98 percent, which happens after training for two epochs(epoch 0 and epoch 1). At the third epoch (epoch = 2), validation accuracy starts to decline. After being trained for epoch 10, the validation accuracy is even worse than the accuracy after training for only one epoch. The decline of validation accuracy is likely caused by overfitting; thus, early stopping approach during the course of training should be adopted after epoch 0 and epoch 1.

0.2 N-Gram

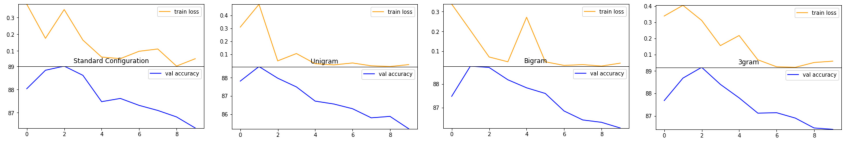


Figure 2: From left to right: n = 1,2,3,4

ID	N-Gram	Max Vocab size	Remove Punctu-ations	Remove Stop Words	Remove HTML Tag	Optimizer	Learning Rate	LR Linear Decay	Size of Embedding	Best Epochs	Best Validation Accuracy
0	4	10k	Yes	Yes	Yes	Adam	0.005	No	200	2	88.98%
1	1	10k	Yes	Yes	Yes	Adam	0.005	No	200	2	88.62%
2	2	10k	Yes	Yes	Yes	Adam	0.005	No	200	2	88.76%
3	3	10k	Yes	Yes	Yes	Adam	0.005	No	200	3	89.16%

Figure 3: I only changed the values of n in n-gram, while keeping other parameters as default. The result is summarized in this table.

According 3, while keeping other parameter values constant, 3-gram bag of word gives the highest validation accuracy 89.16 percent, which happens after training three epochs (epoch = 2). In addition, the best validation accuracy keeps increasing from n = 1 to n = 3 in n-gram, because bi-gram and 3-gram can better capture the context in users' reviews. For example, in the phrase "not so good", bi-gram will extract the "so good" part, while 3-gram can precisely capture the entire phrase. However, with 4-gram, some unnecessary information in the token will be captured, bringing in some noise to the context.

0.3 Tokenization Scheme

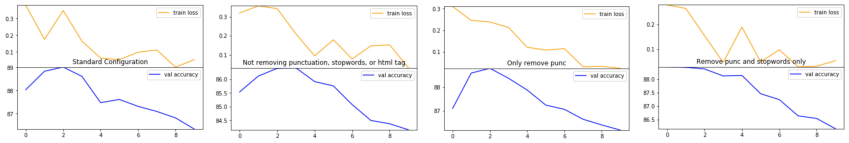


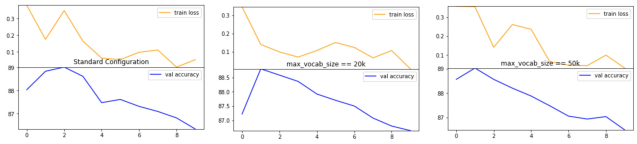
Figure 4: From left to right: remove punctuation,stop words, and html tag; does not remove anything; remove punctuation only; remove punctuation and stop words.

ID	N-Gram	Max Vocab size	Remove Punctuations	Remove Stop Words	Remove HTML Tag	Optimizer	Learning Rate	LR Linear Decay	Size of Embedding	Best Epochs	Best Validation Accuracy
0	4	10k	Yes	Yes	Yes	Adam	0.005	No	200	2	88.98%
4	4	10k	No	No	No	Adam	0.005	No	200	4	86.42%
5	4	10k	Yes	No	No	Adam	0.005	No	200	3	88.8%
6	4	10k	Yes	Yes	No	Adam	0.005	No	200	1	88.46%

Figure 5: Only tokenization schemes are changed, while keeping other parameters as default.

Before punctuation, stop words, and html tags are removed, the best validation accuracy is as low as 8642 percent. As soon as the punctuation is separated from words, the validation accuracy jumps to 88.8 percent, because the binary sentiment (positive, negative) from a review does not depend on the punctuation. Tokenizing punctuation greatly increases the vocabulary size. When we only take the top k tokens into considerations building the model, tokens of punctuation will bring significant noise because they are the most common elements in the text. Removing the html tag also increases validation accuracy for the same reason.

0.4 Maximum vocabulary size



ID	N-Gram	Max Vocab size	Remove Punctuations	Remove Stop Words	Remove HTML Tag	Optimizer	Learning Rate	LR Linear Decay	Size of Embedding	Best Epochs	Best Validation Accuracy
0	4	10k	Yes	Yes	Yes	Adam	0.005	No	200	2	88.98%
7	4	20k	Yes	Yes	Yes	Adam	0.005	No	200	2	88.8%
8	4	50k	Yes	Yes	Yes	Adam	0.005	No	200	2	89.02%

Figure 7: Only maximum vocabulary sizes are changed, while keeping other parameters as default.

Only top k tokens are taken into consideration for the classifier. If the value for k is too small, not enough information is captured; thus, the validation accuracy is the highest when the maximum vocabulary size = 50k. However, maximum vocabulary size of 10k gives better result than maximum vocabulary size of 20k, which worths further investigation.

0.5 Optimizer

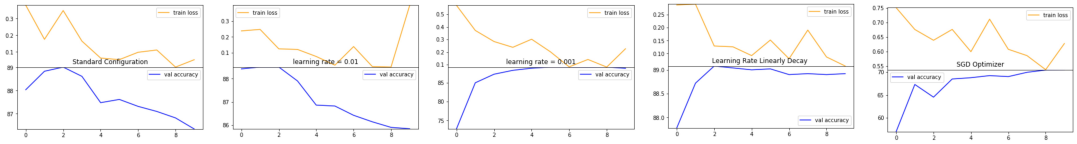


Figure 8: From left to right: Adam with constant lr = 0.005, Adam with constant lr = 0.01, Adam with constant lr = 0.001, Adam with linearly annealing lr where gamma = 0.3, SGD with constant lr = 0.005,

ID	N-Gram	Max Vocab size	Remove Punctu-ations	Remove Stop Words	Remove HTML Tag	Optimizer	Learning Rate	LR Linear Decay	Size of Embedding	Best Epochs	Best Validation Accuracy
0	4	10k	Yes	Yes	Yes	Adam	0.005	No	200	2	88.98%
9	4	10k	Yes	Yes	Yes	Adam	0.01	No	200	2	88.5%
10	4	10k	Yes	Yes	Yes	**Adam**	**0.001**	**No**	200	6	**89.1%**
11	4	10k	Yes	Yes	Yes	Adam	0.005	Yes	200	3	89.08%
12	4	10k	Yes	Yes	Yes	SGD	0.005	No	200	10	70.54%

Figure 9: Only optimizer type and learning rates are changed, while keeping other parameters as default.

According to the table and training curve, Adam optimization tends to converge faster, which makes it works well for small values of epochs. During the course of 10 epochs of training, the training curve from SGD does not converge and the validation accuracy keeps increasing; it requires larger number of training epochs. With constant learning rates of 0.005 and 0.01, the best validation accuracy occurs after training for 3 epochs. When the learning rate starts at 0.005 but decreases by 60 percent after each epoch, the best validation accuracy occurs after training for 4 epochs, and this configuration increases the best validation accuracy to 89.08 percent. Finally, when the constant learning rate equals 0.001, the best validation accuracy happens after training for 7 epochs.

When the learning rate is large, although less epochs of training are needed, the model will miss the ”optimal point” that gives the highest validation accuracy possible. With much smaller learning rate, it takes longer to optimize but it can get closer to the optimal point. The better configuration might be starting with a medium learning rate and decrease the learning rate every epoch.

0.6 Embedding Size

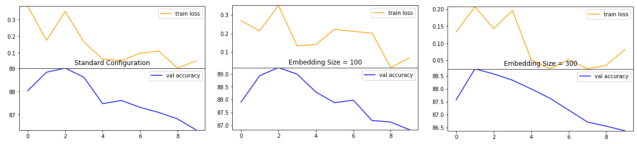


Figure 10: From left to right: Embedding size = 200, 100, 300

ID	N-Gram	Max Vocab size	Remove Punctu-ations	Remove Stop Words	Remove HTML Tag	Optimizer	Learning Rate	LR Linear Decay	Size of Embedding	Best Epochs	Best Validation Accuracy
0	4	10k	Yes	Yes	Yes	Adam	0.005	No	200	2	88.98%
13	4	10k	Yes	Yes	Yes	Adam	0.005	No	**100**	3	**89.26%**
14	4	10k	Yes	Yes	Yes	Adam	0.005	No	300	2	88.78%

Figure 11: Only embedding sizes are changed, while keeping other parameters as default.

The lowest embedding size (=100) gives the highest validation accuracy. As the embedding size increases, the validation accuracy decreases. A possible reason is that the model tends to overfit when the embedding sizes are high, providing some noisy information from the high dimension space.

0.7 Best Configuration and Test Accuracy

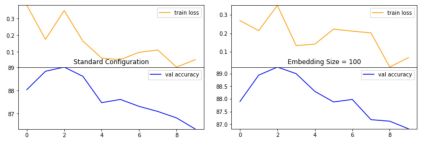


Figure 12: From left to right: Default configuration and the best configuration

ID	N-Gram	Max Vocab size	Remove Punctuations	Remove Stop Words	Remove HTML Tag	Optimizer	Learning Rate	LR Linear Decay	Size of Embedding	Best Epochs	Best Validation Accuracy
15	3	50k	Yes	Yes	Yes	Adam	0.005	Yes	100	4	89.56%

Figure 13: The configuration that gives the highest accuracy on validation set.

Using the configuration that gives the highest validation accuracy from each section above, the model gives test accuracy of 87.568 percent.

0.8 Predictions: full text refer to notebook

Correct prediction 1: Text: worst movies 've seen 'm sure satire movies uses stupid ¡unk¿ cliché think ... maybe worth additionally 're forced date n't like suggest watching movie 'll probably leave 's fair price pay guess

Label: -1

Correct prediction 2 : romance air love bloom victorian era england light hearted story set society time manners ¡unk¿ ladies charming elegant gentlemen dashing emma based novel ... settings proceedings offer sense calm allows ¡unk¿ simply ¡unk¿ ¡unk¿ paltrow won oscar best actress shakespeare love years making

Label: 1

Correct prediction 3: loved curse frankenstein rushed frankenstein destroyed cushing ... ¡unk¿ lee time great disappointment movie lee frankenstein 's monster altogether ... on weak script thought directing flat couple nice shots excitement atmosphere suspense generated director curse

Label: -1

Incorrect prediction 1: youth sexuality french countryside – unique films 're going mean feat considering hard find copies combination ¡unk¿ censorship 's erotic disgusting occasionally funny ¡unk¿ boring middle n't ¡unk¿ bizarre film 've seen

True label: 1

Incorrect 2: like horror fans force fed banal big budget hollywood remakes mtv high school slasher tripe 20 years original horror genre movie ... despite 's low budget ¡unk¿ use 35 mm stock adding quality cg effects mix director james ¡unk¿ created feels ¡unk¿ a old school horror fans

True label: 1

Incorrect 3: movie distinct albeit ¡unk¿ rough humanity borderline ¡unk¿ ¡unk¿ lyrical score points comic ¡unk¿ 're ¡unk¿ like ¡unk¿ woman train semi pitiful vulnerability ¡unk¿ far away ¡unk¿ sucks breasts like baby ...making mockery notion sensitivity honesty hitting numerous points possible

True Label: 1