Assignment1 Report

Adil Kaan Akan

2171155

1 Experiment results

In the following results, the table shows the loss, categorical accuracies, top 2 accuracies and top 3 accuracies. However, the graphs are from training. I save the history of the training and plot the values. The graphs are Loss vs Epochs, Accuracies vs Epoch and Confusion matrix, respectively.

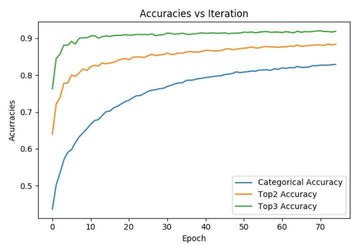
1.1 Min_df and Method

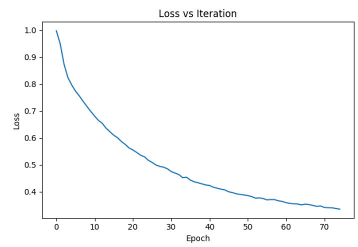
The first section is for choosing the best minumum document frequency and best method to encode features. At the end, we will choose the best minumum document frequency and best method according to top 3 accuracy metric.

1.1.1 Min_df=0.5, Method=Existance, others are default

The result for that experiment is on the table below.

Method	Min df	Loss	Categorical Accuracy	Top 2 Accuracy	Top 3 Accuracy
Existance	0.5	1.345	0.33	0.63	0.83





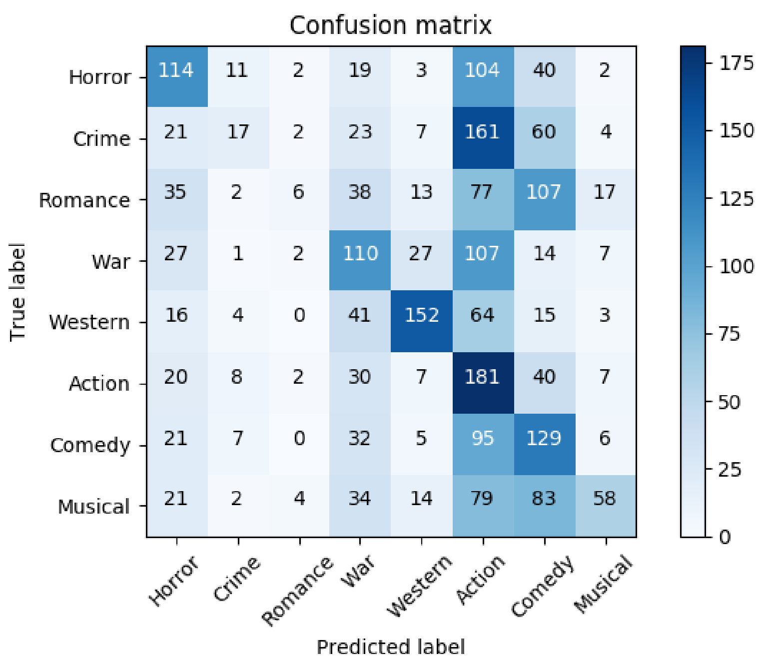
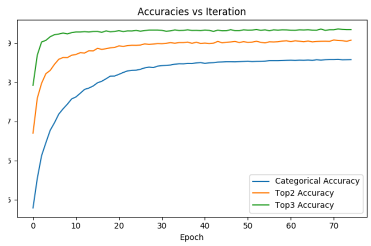


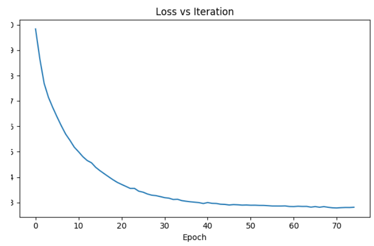
Figure 1: Confusion matrix.

1.1.2 Min_df=0.25, Method=Existance, others are default

The result for that experiment is on the table below.

Method	Min df	Loss	Categorical Accuracy	Top 2 Accuracy	Top 3 Accuracy
Existance	0.25	1.135	0.43	0.82	0.955





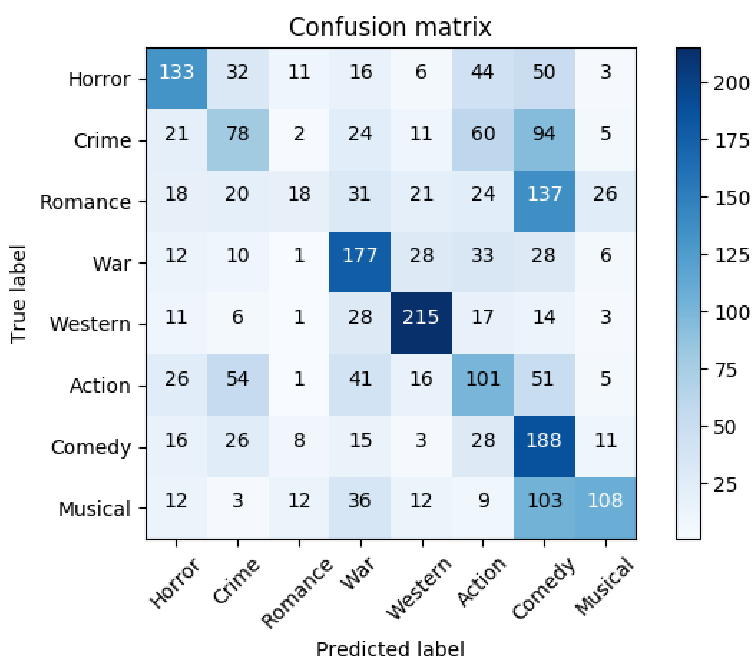
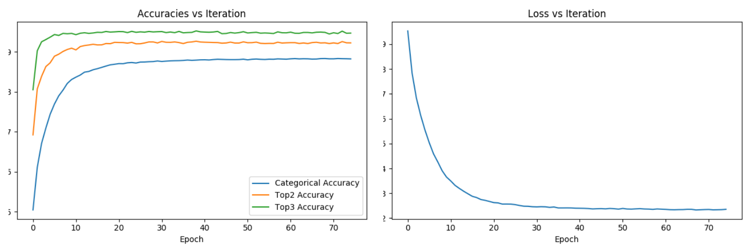


Figure 2: Confusion matrix.

1.1.3 Min_df=0.1, Method=Existance, others are default

The result for that experiment is on the table below.

Method	Min df	Loss	Categorical Accuracy	Top 2 Accuracy	Top 3 Accuracy
Existance	0.1	1.071	0.46	0.88	0.969



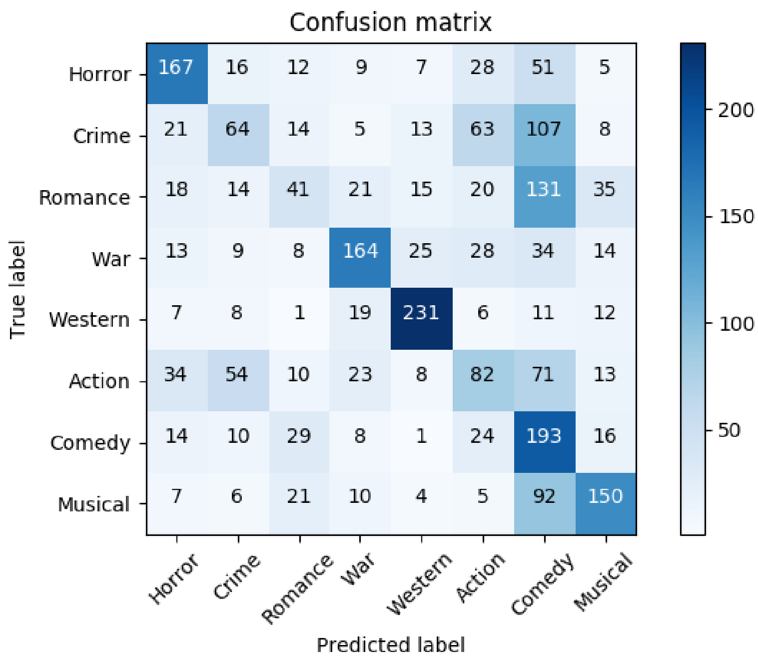
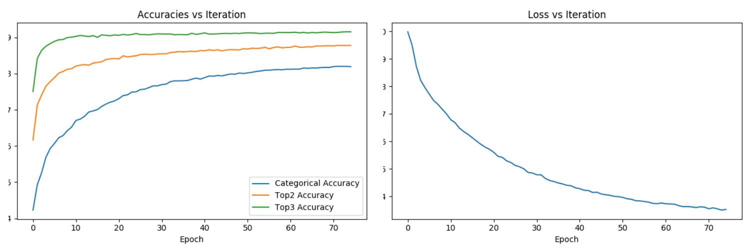


Figure 3: Confusion matrix.

1.1.4 Min_df=0.5, Method=Count, others are default

The result for that experiment is on the table below.

Method	Min df	Loss	Categorical Accuracy	Top 2 Accuracy	Top 3 Accuracy
Count	0.5	1.040	0.479	0.955	0.993



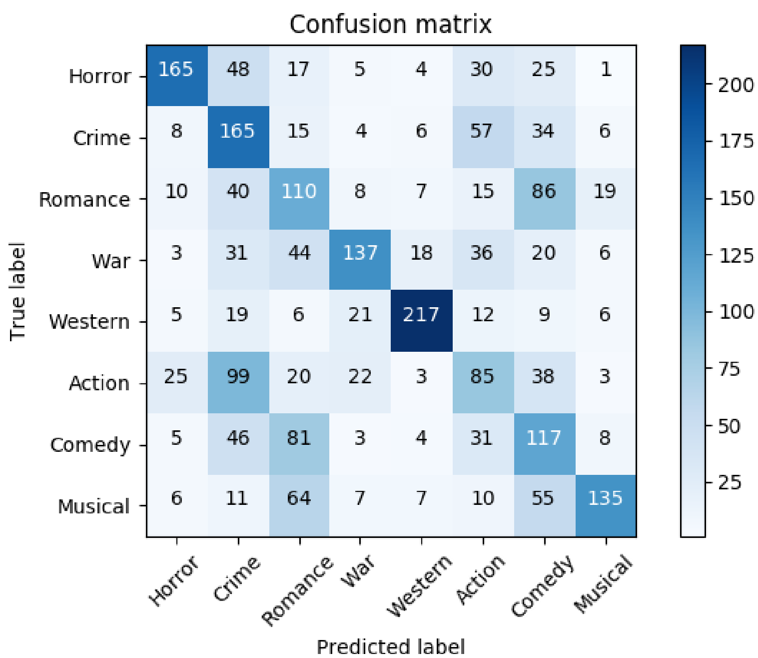
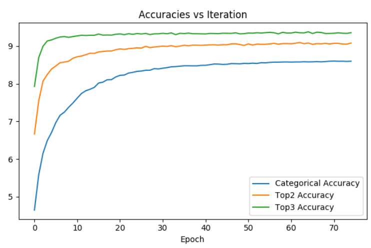


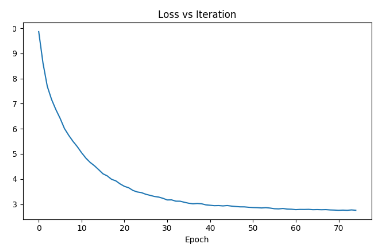
Figure 4: Confusion matrix.

1.1.5 Min_df=0.25, Method=Count, others are default

The result for that experiment is on the table below.

Method	Min df	Loss	Categorical Accuracy	Top 2 Accuracy	Top 3 Accuracy
Count	0.25	1.012	0.493	0.976	0.996





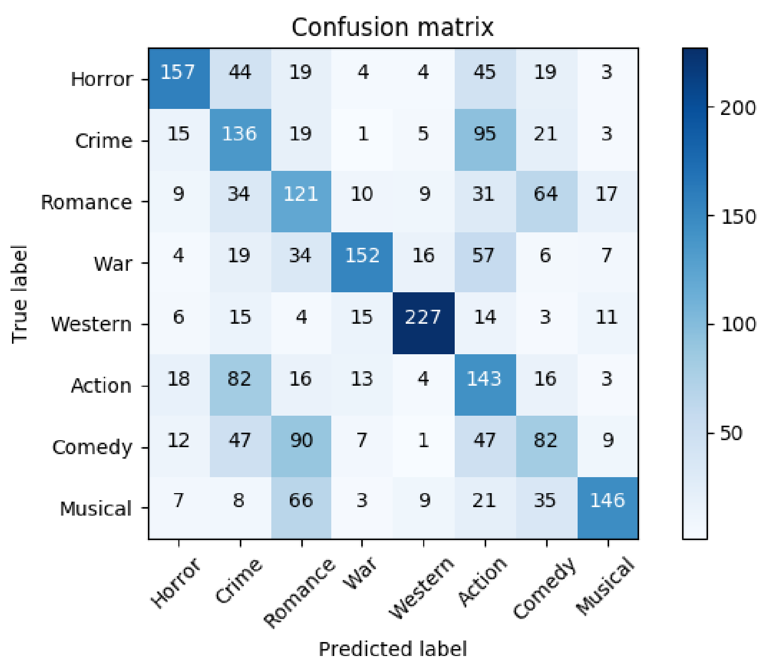
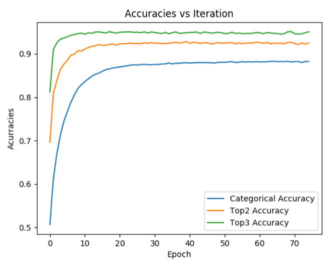


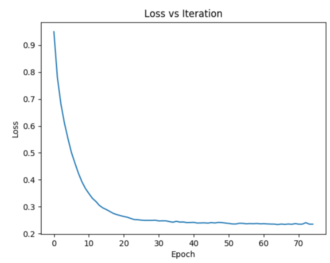
Figure 5: Confusion matrix.

1.1.6 Min_df=0.1, Method=Count, others are default

The result for that experiment is on the table below.

Method	Min df	Loss	Categorical Accuracy	Top 2 Accuracy	Top 3 Accuracy
Count	0.1	0.994	0.502	0.974	0.996





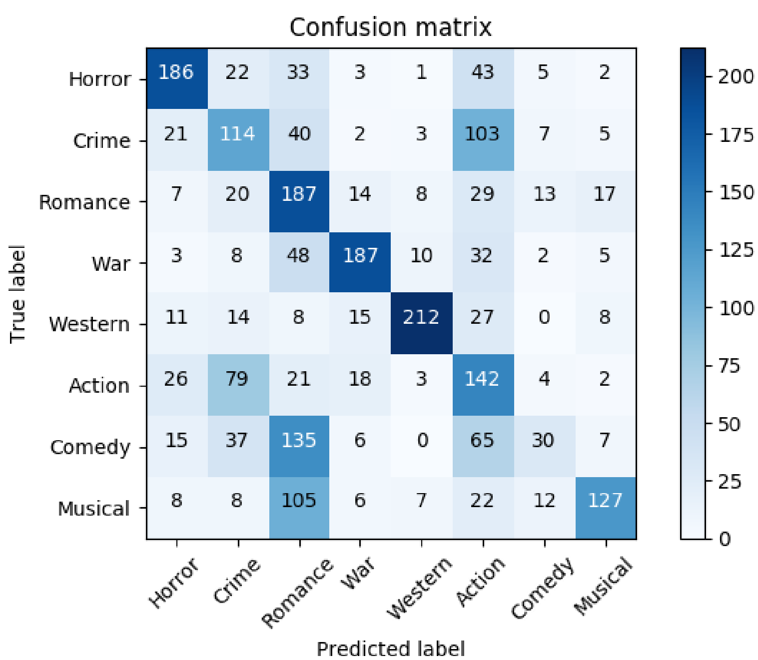
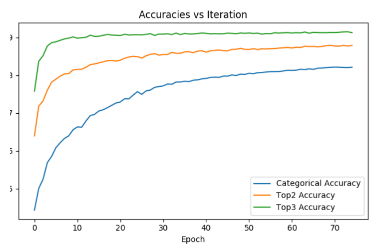


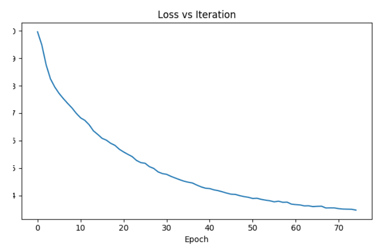
Figure 6: Confusion matrix.

1.1.7 Min_df=0.5, Method=Tf-Idf, others are default

The result for that experiment is on the table below.

Method	Min df	Loss	Categorical Accuracy	Top 2 Accuracy	Top 3 Accuracy
Tf Idf	0.5	0.951	0.510	0.740	0.851





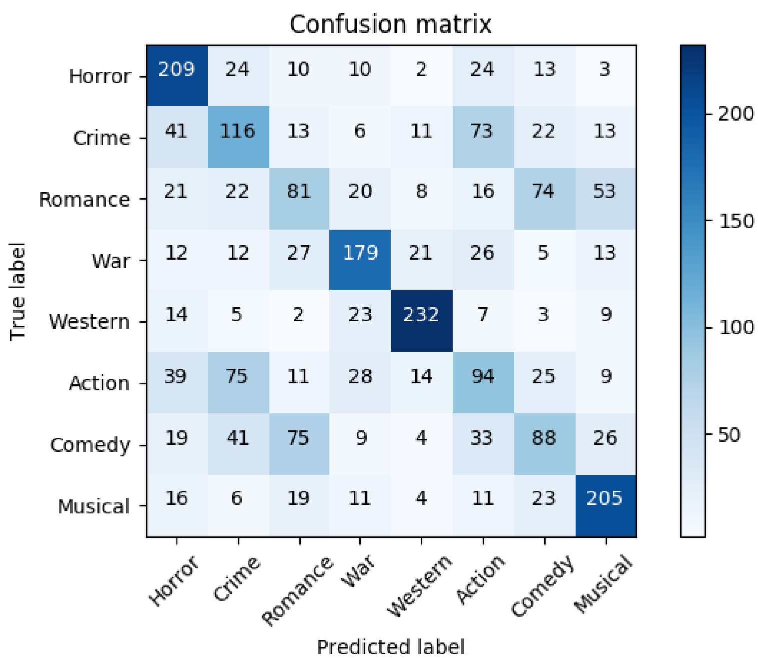
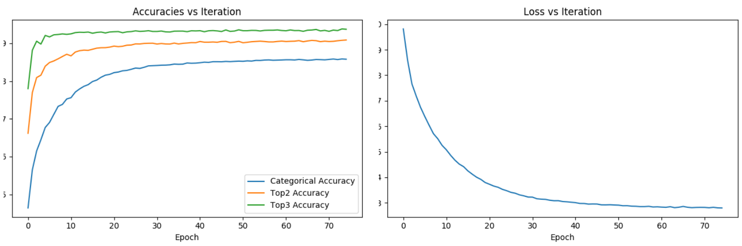


Figure 7: Confusion matrix.

1.1.8 Min_df=0.25, Method=Tf-Idf, others are default

The result for that experiment is on the table below.

Method	Min df	Loss	Categorical Accuracy	Top 2 Accuracy	Top 3 Accuracy
Tf Idf	0.25	0.932	0.524	0.757	0.866



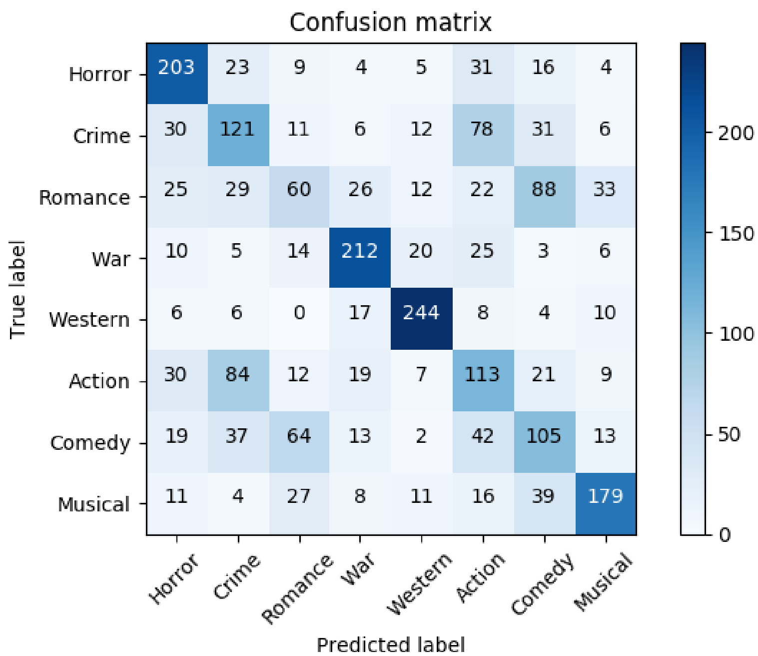
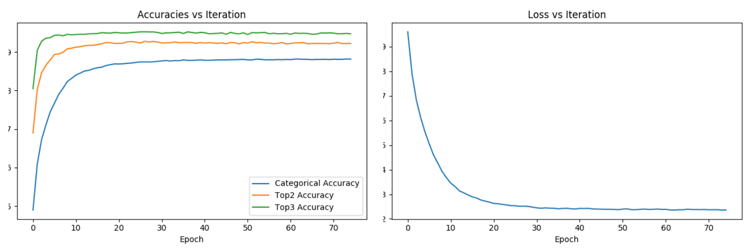


Figure 8: Confusion matrix.

1.1.9 Min_df=0.1, Method=Tf-Idf, others are default

The result for that experiment is on the table below.

Method	Min df	Loss	Categorical Accuracy	Top 2 Accuracy	Top 3 Accuracy
Tf Idf	0.1	0.933	0.53	0.774	0.880



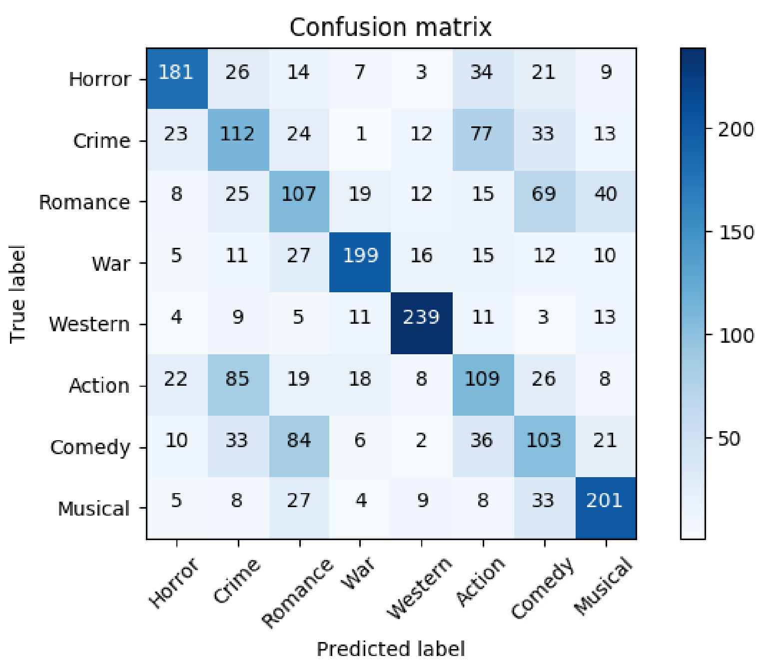


Figure 9: Confusion matrix.

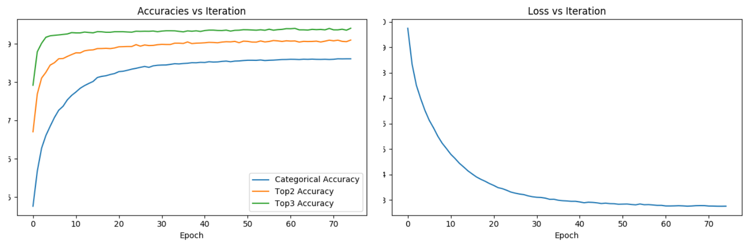
According to experiments, best min_df is **0.1** and best method is **count**.

1.2 Layer Configurations

1.2.1 Layer configuration=(512), Min_df=0.25, Method=Count, others are default

The result for that experiment is on the table below.

Layer Configuration	Loss	Categorical Accuracy	Top 2 Accuracy	Top 3 Accuracy
(512)	0.951	0.497	0.963	0.995



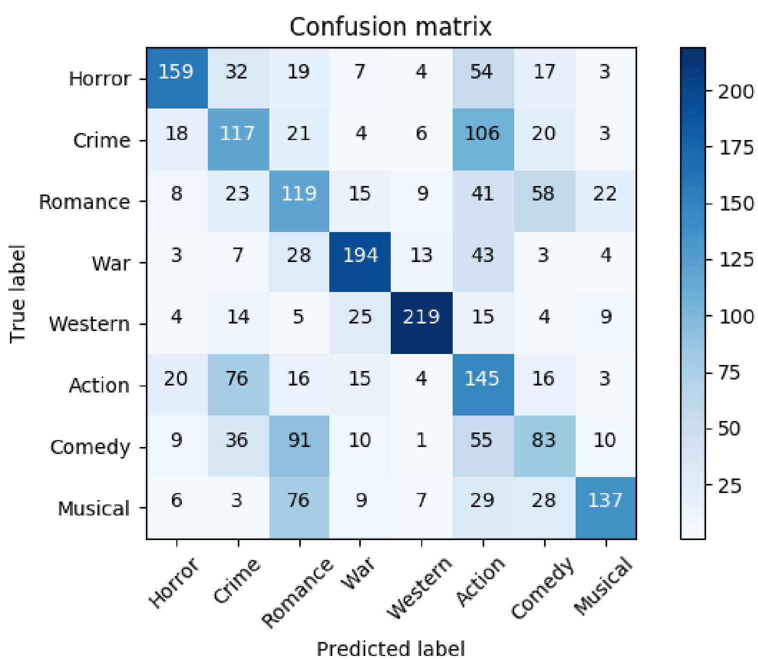
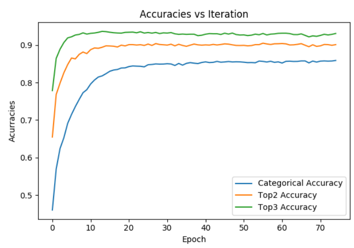


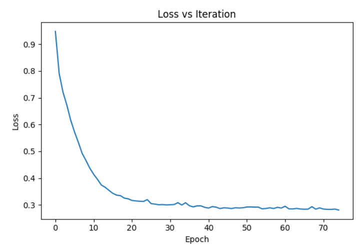
Figure 10: Confusion matrix.

1.2.2 Layer configuration=(512, 256), Min_df=0.25, Method=Count, others are default

The result for that experiment is on the table below.

Layer Configuration	Loss	Categorical Accuracy	Top 2 Accuracy	Top 3 Accuracy
(512, 256)	0.951	0.524	0.978	0.998





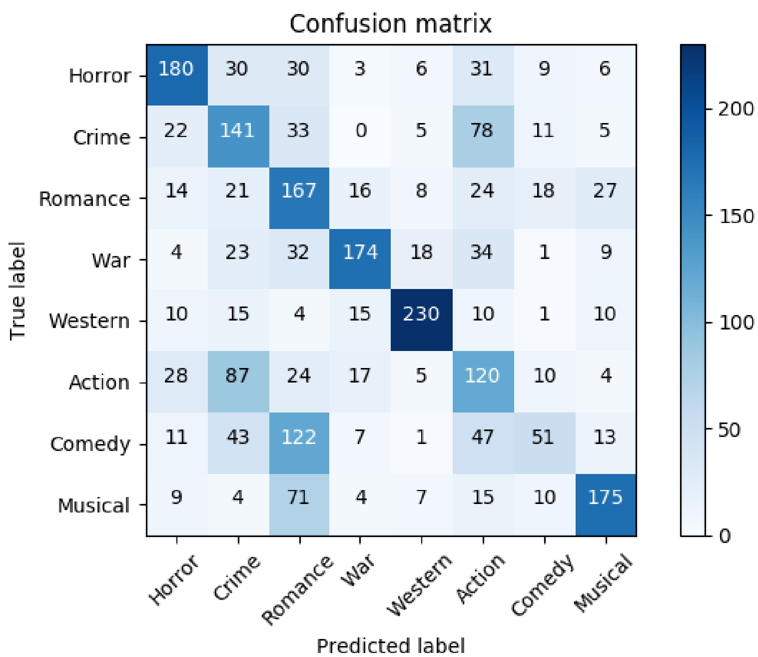
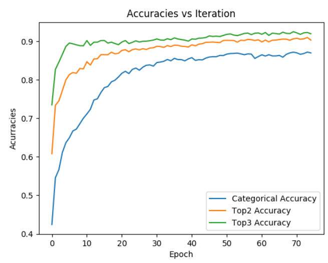


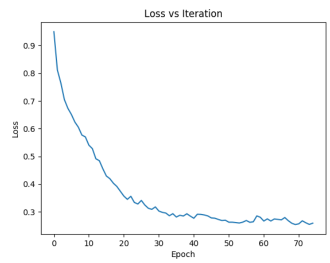
Figure 11: Confusion matrix.

1.2.3 Layer configuration=(512, 256, 128), Min_df=0.25, Method=Count, others are default

The result for that experiment is on the table below.

Layer Configuration	Loss	Categorical Accuracy	Top 2 Accuracy	Top 3 Accuracy
(512, 256, 128)	1.002	0.498	0.986	0.997





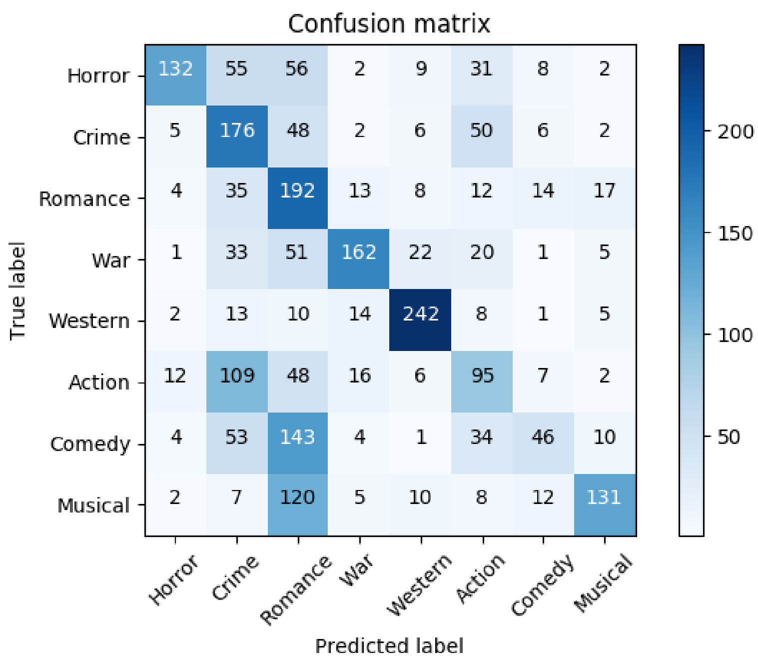


Figure 12: Confusion matrix.

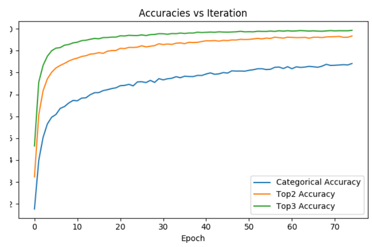
According to experiments, best layer configuration is **(512,256)**.

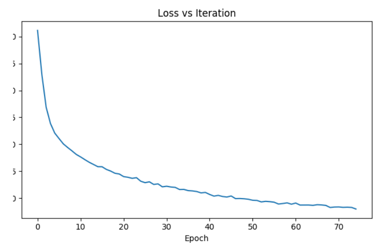
1.3 Activations and Loss Functions

1.3.1 Activation=Sigmoid, Loss=Categorical crossentropy, others are choosen bests

The result for that experiment is on the table below.

Activation Function	Loss Function	Loss	Categorical Accuracy	Top 2 Accuracy	Top 3 Accuracy
Sigmoid	C.Crossentropy	4.53	0.168	0.314	0.441





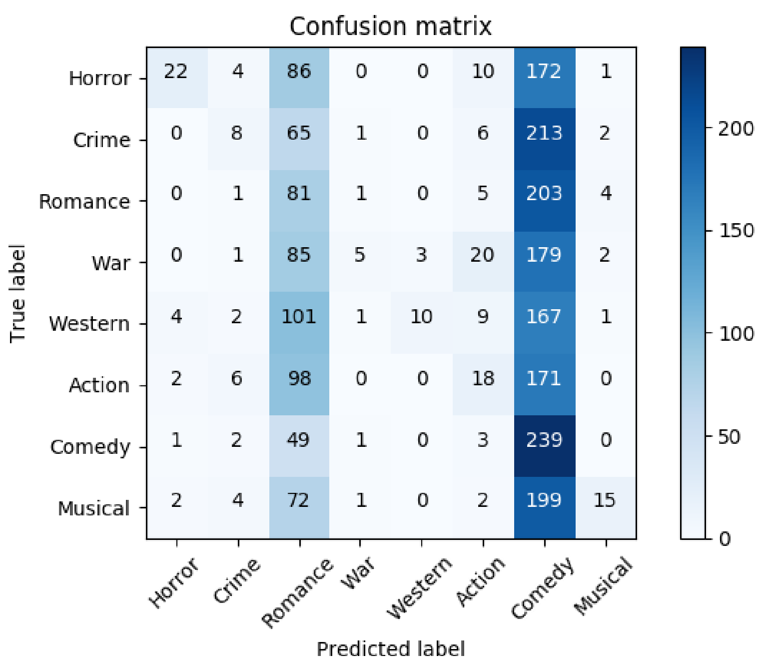
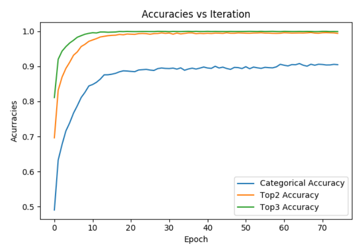


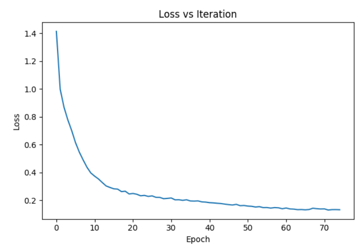
Figure 13: Confusion matrix.

1.3.2 Activation=Relu,Loss=Categorical crossentropy, others are choosen bests

The result for that experiment is on the table below.

Activation Function	Loss Function	Loss	Categorical Accuracy	Top 2 Accuracy	Top 3 Accuracy
Relu	C.Crossentropy	8.205	0.486	0.945	0.993





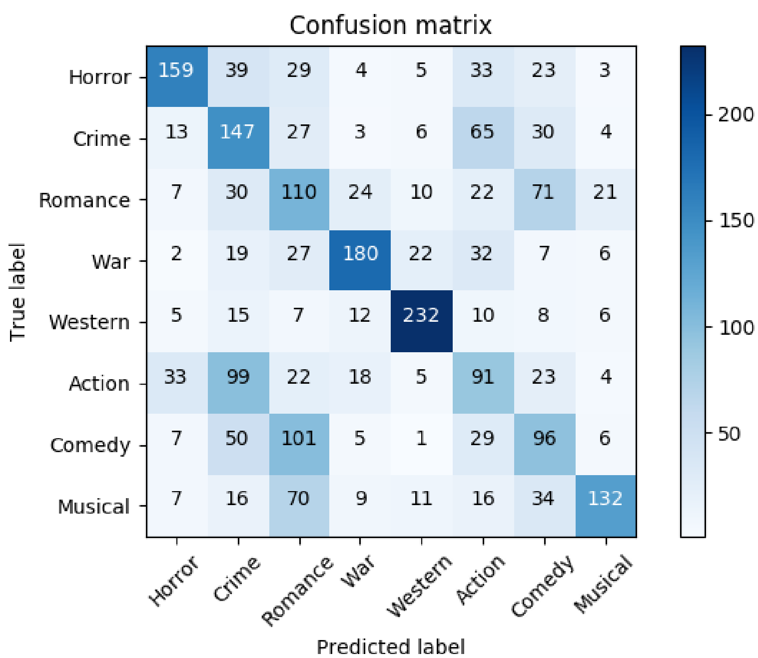
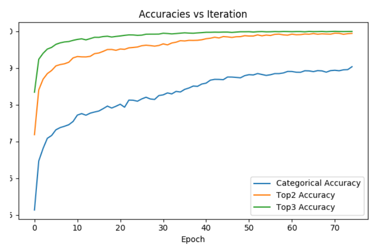


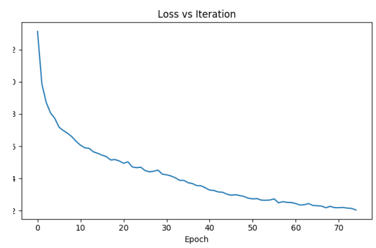
Figure 14: Confusion matrix.

1.3.3 Activation=TanH, Loss=Categorical crossentropy, others are choosen bests

The result for that experiment is on the table below.

Activation Function	Loss Function	Loss	Categorical Accuracy	Top 2 Accuracy	Top 3 Accuracy
TanH	C.Crossentropy	3.709	0.453	0.687	0.822





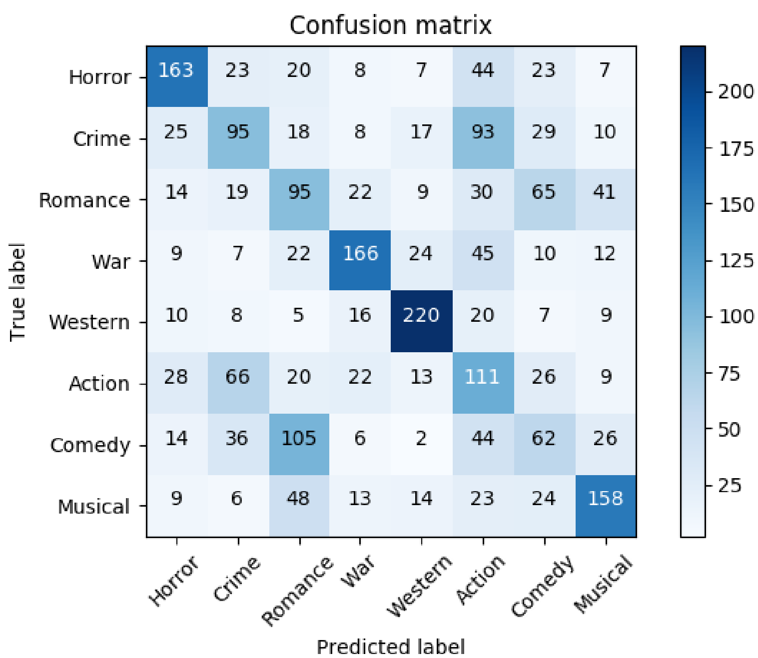
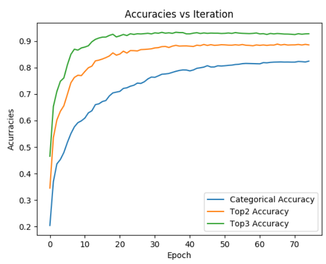


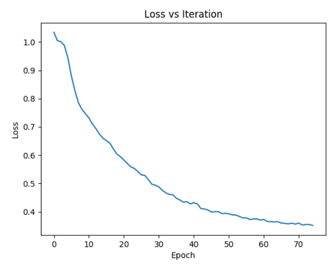
Figure 15: Confusion matrix.

1.3.4 Activation=Sigmoid, Loss=Categorical Hinge, others are choosen bests

The result for that experiment is on the table below.

Activation Function	Loss Function	Loss	Categorical Accuracy	Top 2 Accuracy	Top 3 Accuracy
Sigmoid	Categorical Hinge	1.120	0.412	0.686	0.812





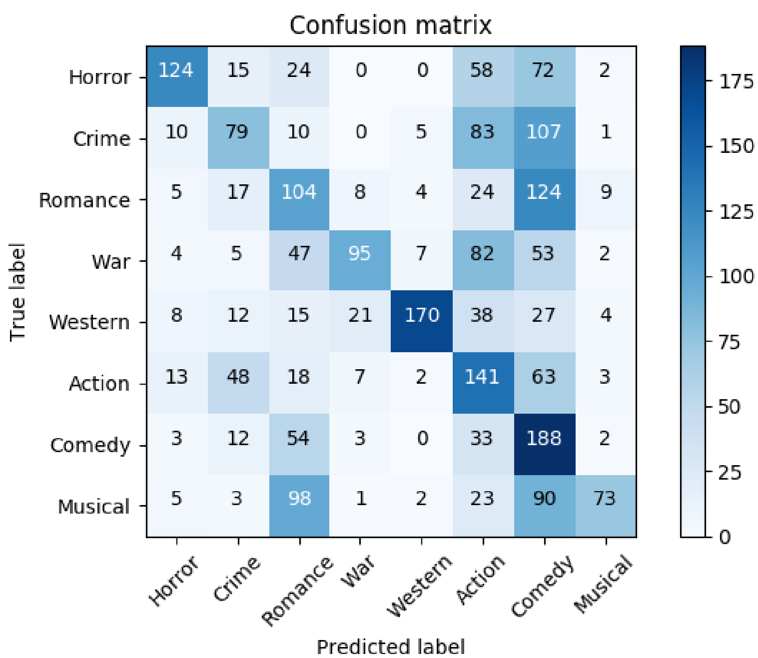
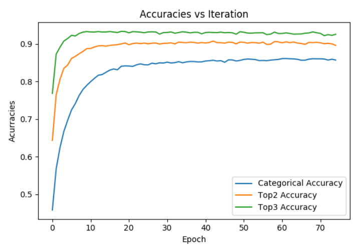


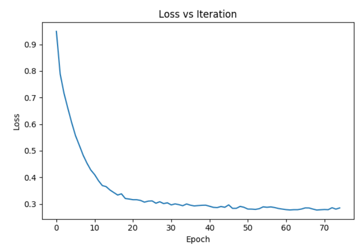
Figure 16: Confusion matrix.

1.3.5 Activation=Relu,Loss=Categorical Hinge, others are choosen bests

The result for that experiment is on the table below.

Activation Function	Loss Function	Loss	Categorical Accuracy	Top 2 Accuracy	Top 3 Accuracy
Relu	Categorical Hinge	0.943	0.527	0.986	0.998





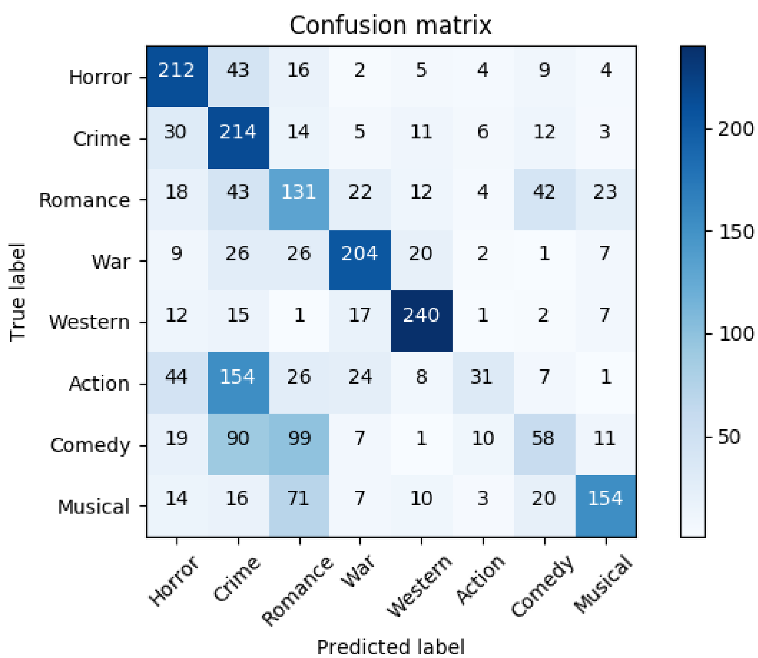
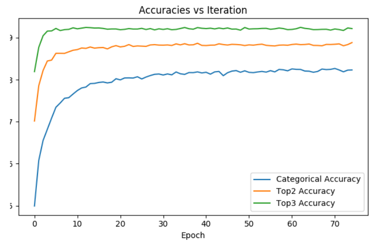


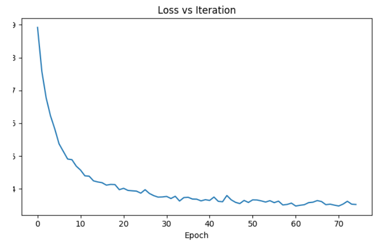
Figure 17: Confusion matrix.

1.3.6 Activation=TanH, Loss=Categorical Hinge, others are choosen bests

The result for that experiment is on the table below.

Activation Function	Loss Function	Loss	Categorical Accuracy	Top 2 Accuracy	Top 3 Accuracy
TanH	Categorical Hinge	0.964	0.513	0.736	0.848





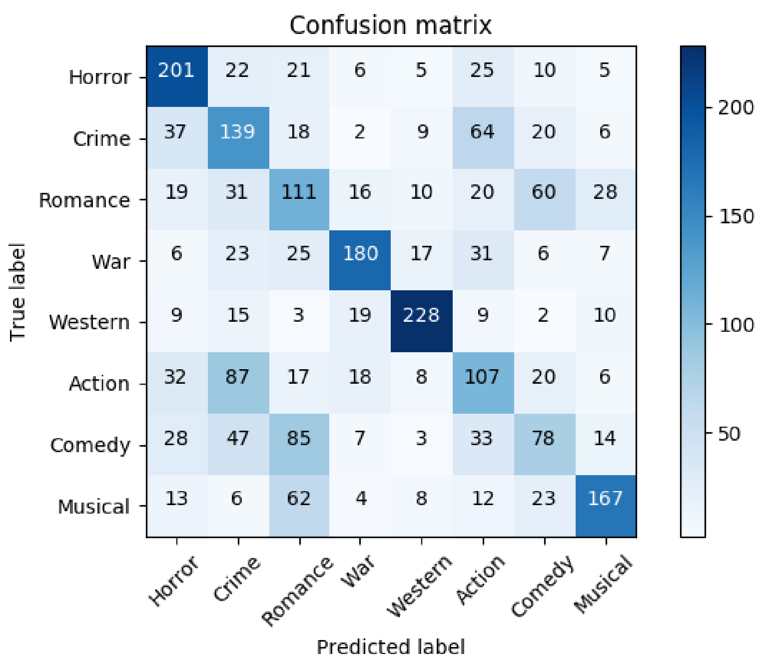


Figure 18: Confusion matrix.

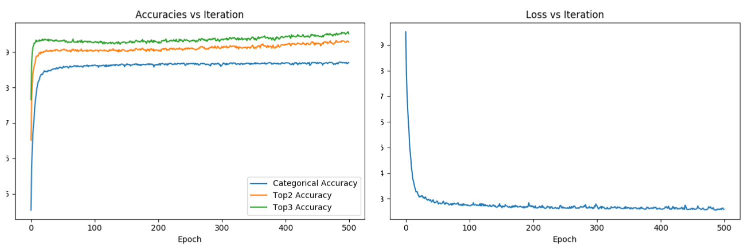
According to experiments, the best paremeters are as follows:

Parameter	Best value
Minumum document frequency	0.25
Method	COUNT
Layer configuration	(512,256)
Activation function	Relu
Loss Function	Categorical hinge loss

1.4 Best model with 500 epochs

The result for that experiment is on the table below.

Loss	Categorical Accuracy	Top 2 Accuracy	Top 3 Accuracy
0.955	0.522	0.9915	0.999



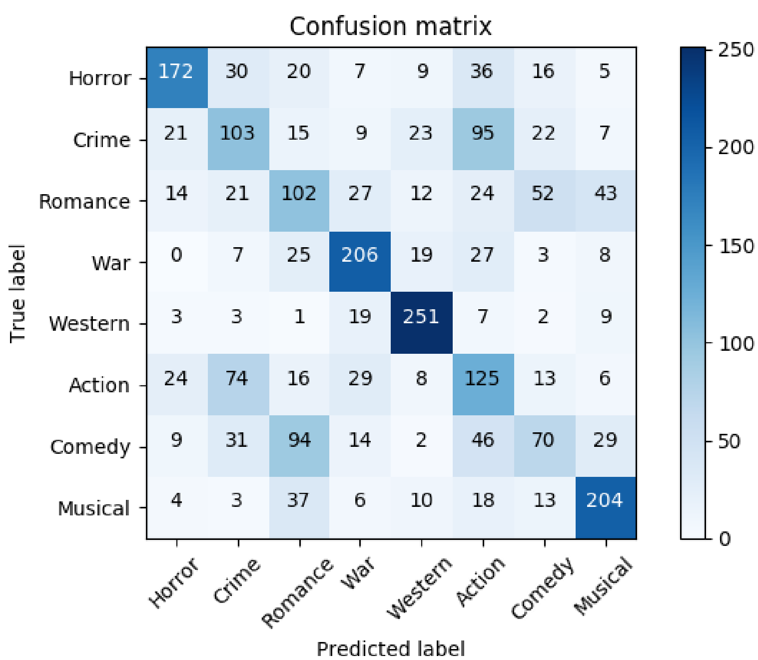


Figure 19: Confusion matrix.

2 Brain Storming Questions

2.1 What if we were very lucky to get these accuracy results depending on training and test set distribution? What is the correct way to measure the accuracy?

There is a possibility to that we were very lucky to get these accuracy results. The correct way to measure accuracy applying cross validation. If we choose one training and test, we could get lucky in that training and test set. However, the model cannot generalize well. Instead of that, we should use cross validation. We split the dataset in to groups, and choose traning and test set, after training and testing, we compare the results in groups. In that way, measured accuracy is more accurate than choosing just one training and test set.

2.2 Why did we use top 2 and top 3 accuracy and they are significantly higher than categor- ical accuracy?

Because in top 2 and top 3 accuracy, we compare the true label and one of the top 2 or top 3 higher predictions. For example, our model can misclassify 'Crime' class as 'Action' by very little chance, in that situation, if we use the categorical accuracy, we will say "misclassification", but when we use top 2 or top 3 accuracy, we can say "true prediction". That is why accuracy is significantly higher when we use top 2 or top 3 accuracy instead of categorical accuracy.

Looking at the confusion matrix of best model with 500 epochs, the model had a tendency to misclassify similar classes, such as "Action" and "Crime", "Romance" and "Comedy". It is expected result since we use bag-of-word approach. It may be better classified if we use better approach to encode words, such as doc2vec method.

2.3 What are the disadvantages of using Bag-of-Words approach?

In bag-of-words approach, we just count the words and based on that count we encode the word. However, bag-of-words approach do not care the ordering of the words. That is really big problem. For example, consider two sentences, "Course is good" and "Is course good", in bag-of-words approach, both sentences we the have encoded vector, however, the semantic is totally different. Also, when the vocabulary goes very big, our feature vector will be very big. With very big feature vector, models can overfit easily.

2.4 What makes the difference between tf-idf and count? What is the importance of idf?

The difference is that tf-idf looks all documents for each term, whereas count looks just for that document. This difference comes from idf. IDF means inverse document frequency. It gives us a number according to how many times the term appear in all documents. When we use tf-idf instead of count, we will assign accurate values for terms with higher occurences and terms with lower occurences.

2.5 What was the effect of changing min df?

Min df variable determines which words we should add to our vocabulary. When we change it, our vocabulary changes, because we omit the words with smaller df than our min df variable. When we make min df variable small, our vocabulary will contain much more words, because we will add more words to our vocabulary. When our vocabulary is dense, we can encode sentences more accurate. Therefore, we can get better feature vectors. The training experiments support the claim. When we lower the min df variable, accuracies get higher.

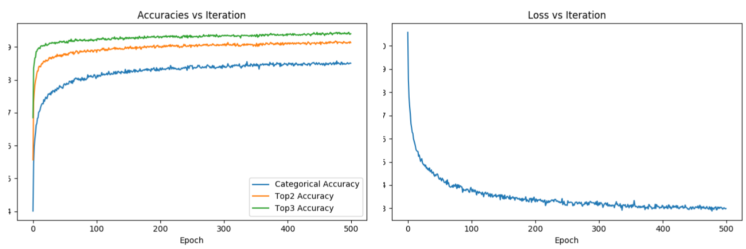
3 Bonuses

I tried Dropout and Batch normalization layers with higher epoch numbers. Also, I tried leaky relu activation function. Here I provide the results of the bonus experiments and graphs of them.

3.1 Best model with Dropout rate = 0.5 and Batch Normalization

The result for that experiment is on the table below.

Loss	Categorical Accuracy	Top 2 Accuracy	Top 3 Accuracy
1.091	0.454	0.984	0.997



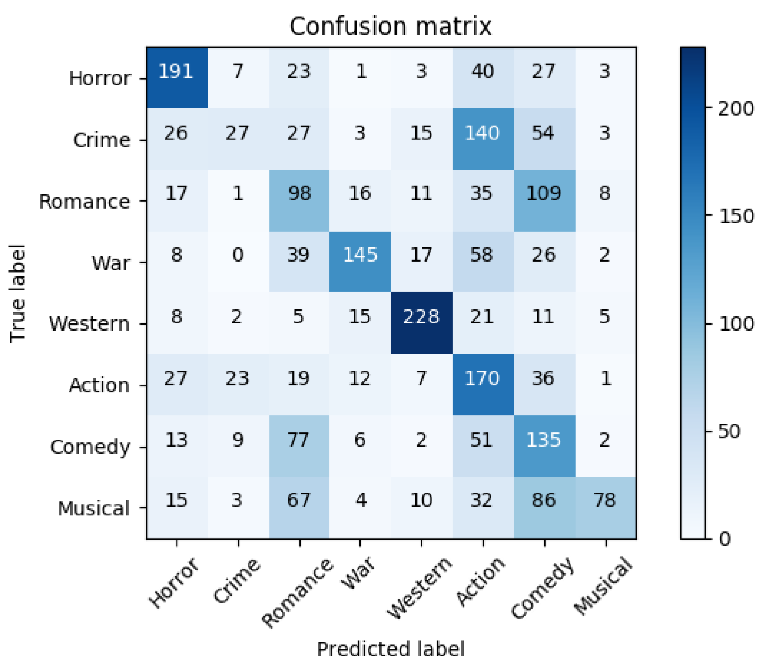
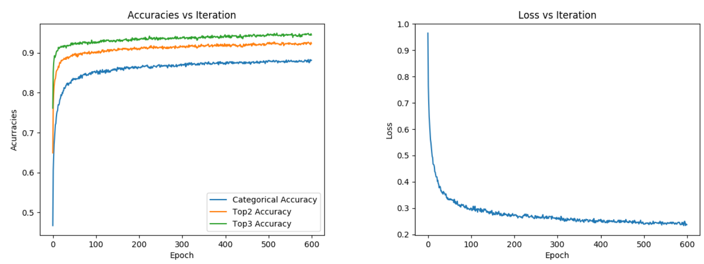


Figure 20: Confusion matrix.

3.2 Best model with Dropout rate = 0.2 and Batch Normalization with 600 epochs

The result for that experiment is on the table below.

Loss	Categorical Accuracy	Top 2 Accuracy	Top 3 Accuracy
1.355	0.322	0.986	0.998



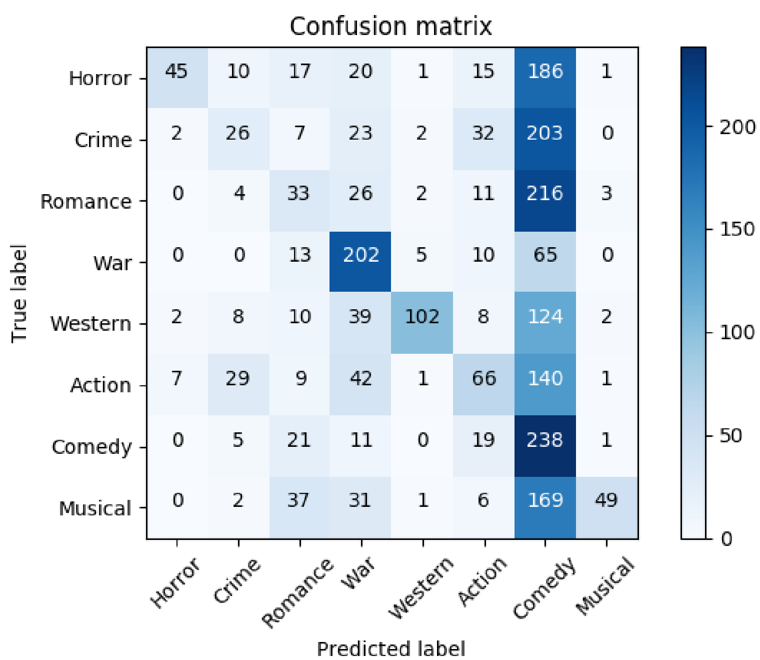
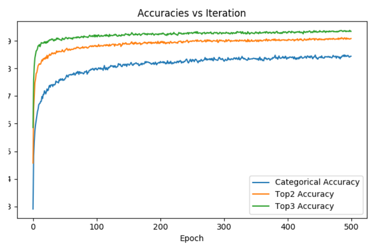


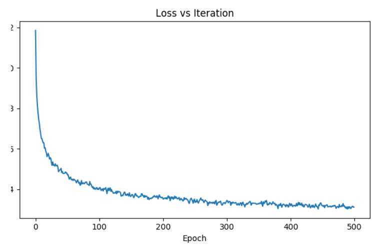
Figure 21: Confusion matrix.

3.3 (512, 256, 128) with Dropout rate = 0.5 and Batch Normalization

The result for that experiment is on the table below.

Loss	Categorical Accuracy	Top 2 Accuracy	Top 3 Accuracy
1.100	0.449	0.981	0.998





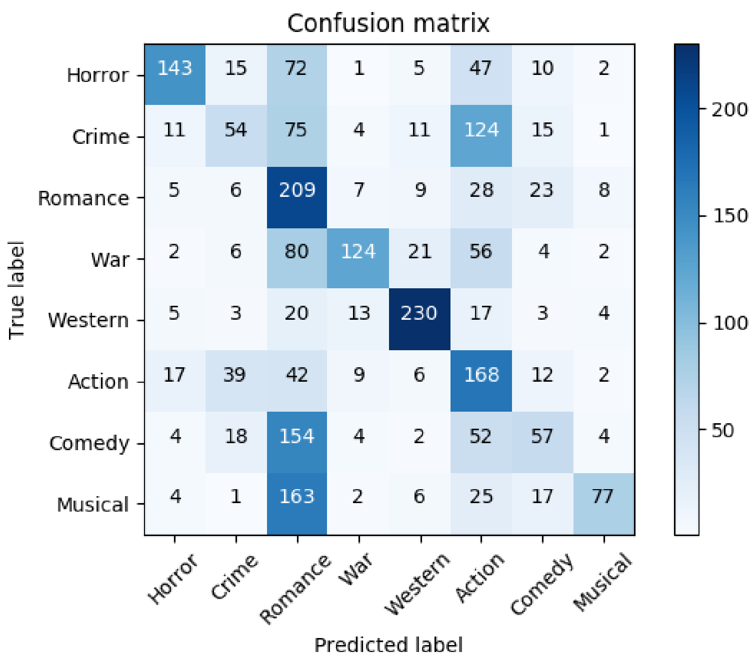
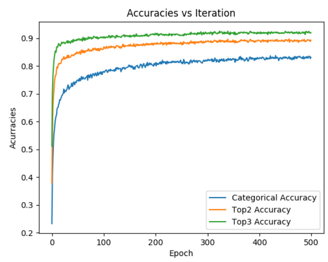


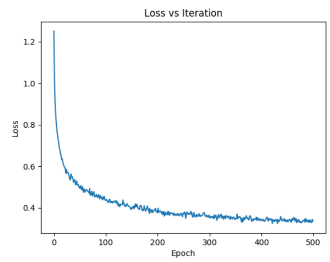
Figure 22: Confusion matrix.

3.4 (512, 256, 256, 128) with Dropout rate = 0.5 and Batch Normalization

The result for that experiment is on the table below.

Loss	Categorical Accuracy	Top 2 Accuracy	Top 3 Accuracy
1.045	0.476	0.972	0.994





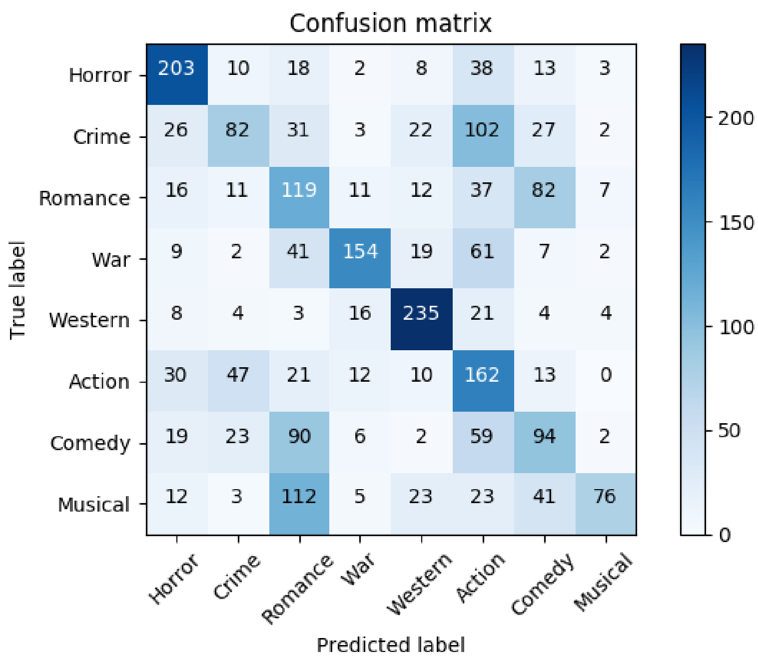
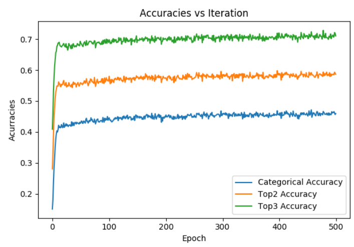


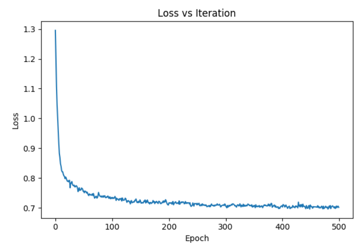
Figure 23: Confusion matrix.

3.5 (512, 256, 128, 64, 32) with Dropout rate = 0.5 and Batch Normalization

The result for that experiment is on the table below.

Loss	Categorical Accuracy	Top 2 Accuracy	Top 3 Accuracy
0.983	0.211	0.338	0.469





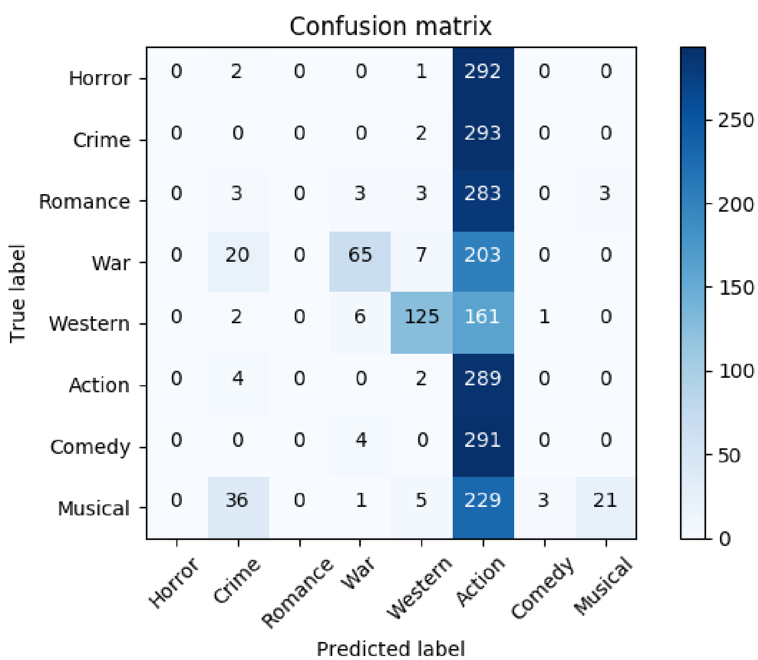


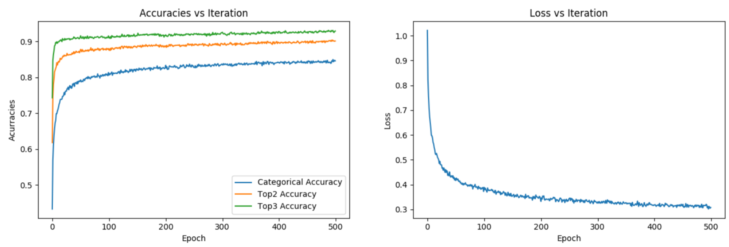
Figure 24: Confusion matrix.

Leaky Relu results

3.6 (512, 256) with Dropout rate = 0.5 and Batch Normalization with Leaky Relu activation

The result for that experiment is on the table below.

Loss	Categorical Accuracy	Top 2 Accuracy	Top 3 Accuracy
1.120	0.439	0.989	0.999



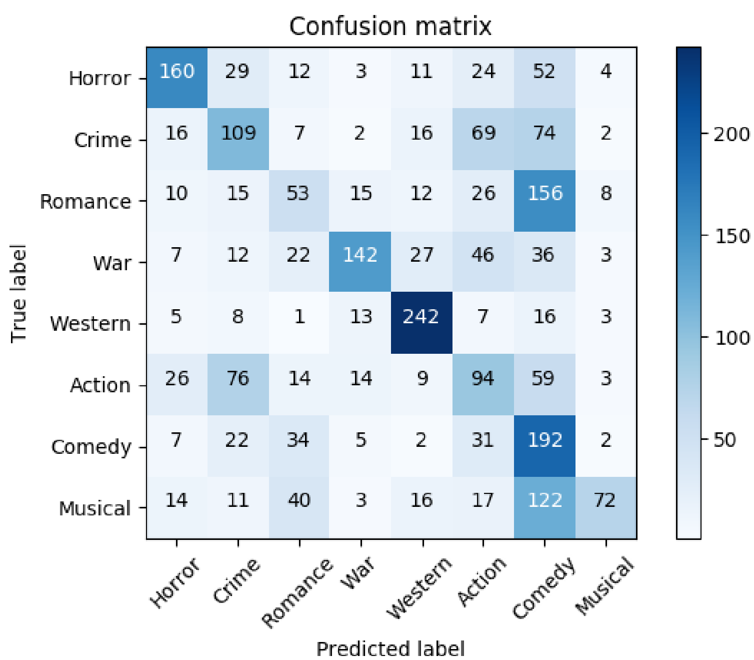
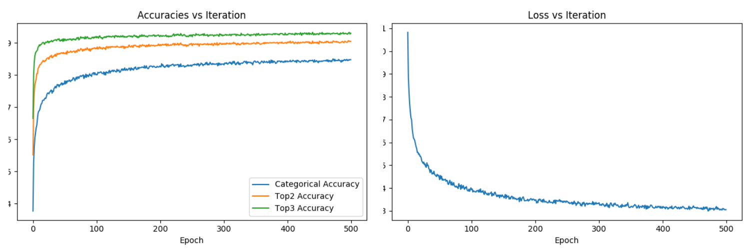


Figure 25: Confusion matrix.

3.7 (512, 256, 128) with Dropout rate = 0.5 and Batch Normalization with Leaky Relu activation

The result for that experiment is on the table below.

Loss	Categorical Accuracy	Top 2 Accuracy	Top 3 Accuracy
1.086	0.457	0.988	0.999



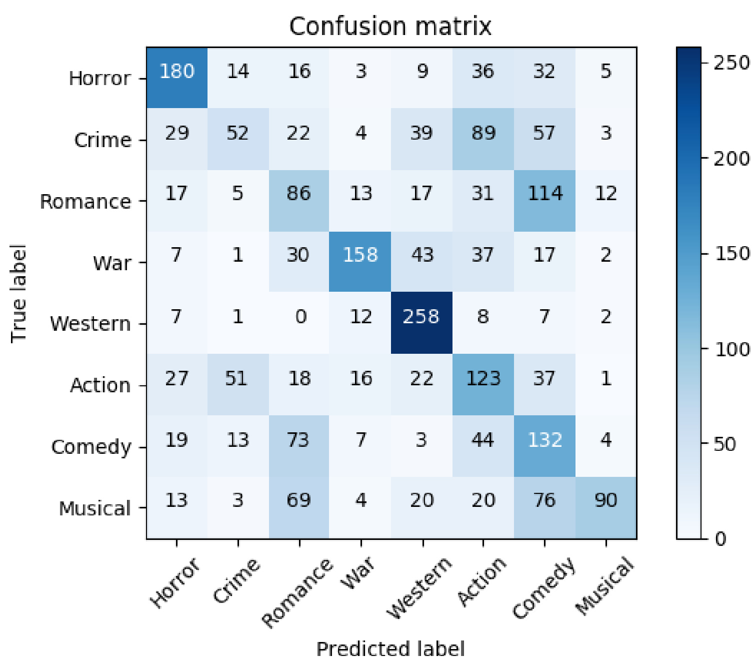
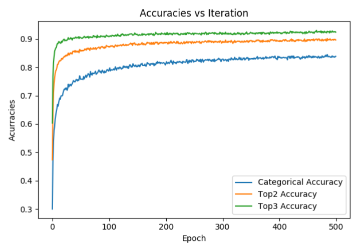


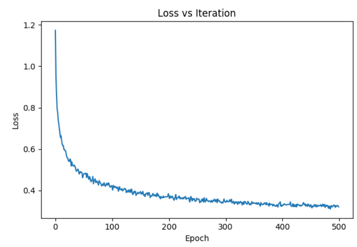
Figure 26: Confusion matrix.

3.8 (512, 256, 256, 128) with Dropout rate = 0.5 and Batch Normalization with Leaky Relu activation

The result for that experiment is on the table below.

Loss	Categorical Accuracy	Top 2 Accuracy	Top 3 Accuracy
1.097	0.450	0.985	0.999





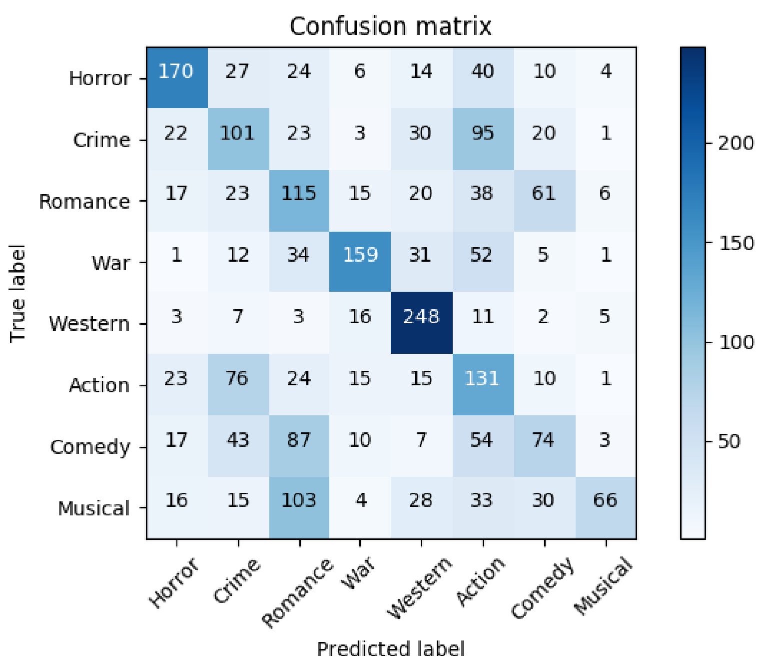
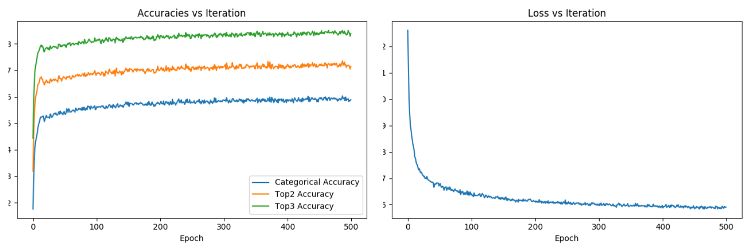


Figure 27: Confusion matrix.

3.9 (512, 256, 128, 64, 32) with Dropout rate = 0.5 and Batch Normalization with Leaky Relu activation

The result for that experiment is on the table below.

Loss	Categorical Accuracy	Top 2 Accuracy	Top 3 Accuracy
1.228	0.285	0.452	0.591



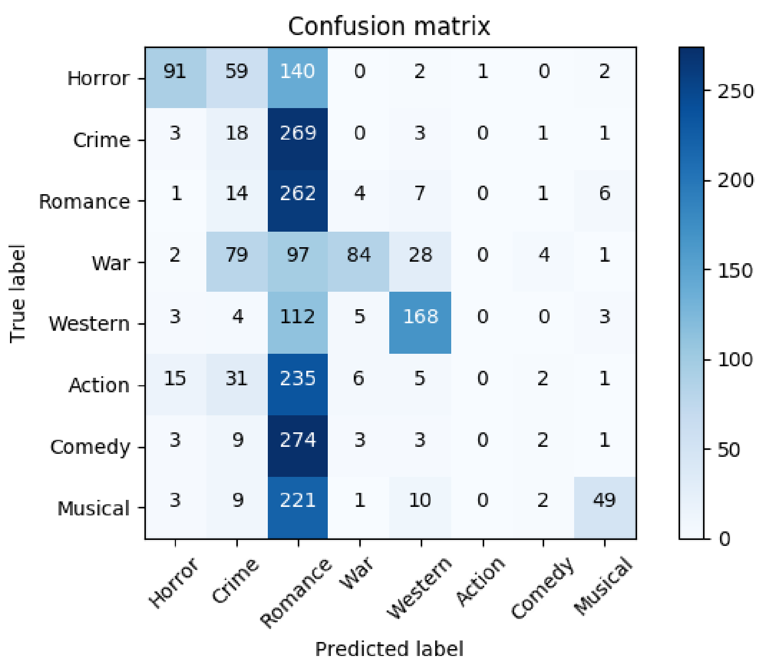
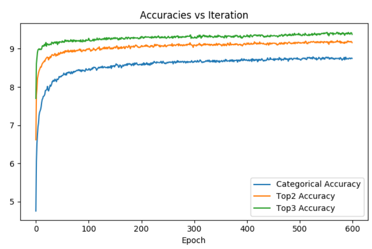


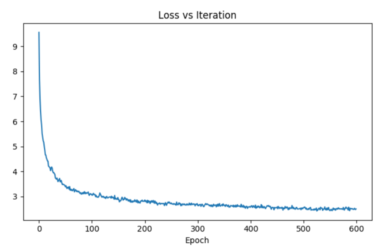
Figure 28: Confusion matrix.

3.10 (512, 256, 128, 64, 32) with Dropout rate = 0.5 and Batch Normalization with Leaky Relu activation

The result for that experiment is on the table below.

Loss	Categorical Accuracy	Top 2 Accuracy	Top 3 Accuracy
1.098	0.450	0.990	0.999





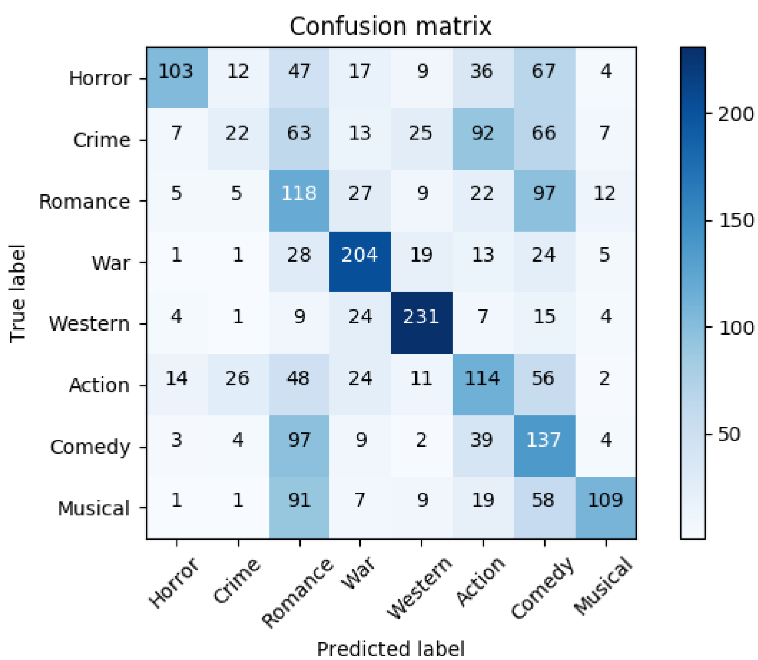


Figure 29: Confusion matrix.