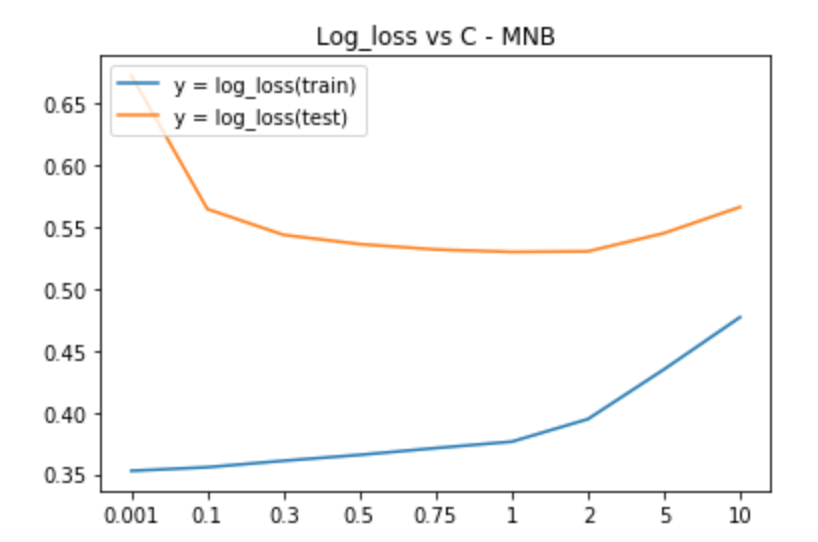
Machine Learning Laboratory at IIT - Fall 2018 Final Report

Transparent Text Classification

Ekrem Guzelyel­



# Introduction & Purpose

Data means prediction power. However, sometimes you can’t have enough data only with human sources. In order to classify movie reviews by IMDB, we use EDUs. Naturally, labeling all EDUs by hand consumes too much valuable time. Our job as Undergraduate Research Assistants at ML Lab at IIT is to build an efficient model that predicts the labels for the EDUs. This way we would be able to have a bigger dataset to train the actual text classification model. Tasks include:

* Labeling more EDUs on top of already labeled 2000 data points.
* Trying different combinations of different approaches to train a model that maximize the performance.

## Approaches

I have used 7 different models for prediction; all of which resulted with different accuracies. These models are with using Logistic Regression, Support Vector Machines (SVM), Multinomial Naïve Bayes (MNB), Long-Short Term Memory (LSTM), Stacked LSTM, LSTM on top of Convolutional Neural Networks (CNN), Gated Recurrent Units (GRU).

“Try every combination. One will behave better.” – Anneke Soraya Hidayat

We expect one model to give a better accuracy than the baseline LR model.

|  |  |  |
| --- | --- | --- |
|  | Train Accuracy | Test Accuracy |
| LR | Text | 123.45 |
| SVM | Text | 123.45 |
| MNB | Text |  |
| LSTM | Text |  |
| CNN+LSTM | Text |  |
| Stacked LSTM | Text |  |
| GRU | Text |  |

## Metrics

The main metrics for this task is comparing train and test accuracy, along with looking at precision and recall. For some cases that were needed to be investigated deeper, log-loss has also been used.

“If you don’t collect any metrics, you’re flying blind. If you collect and focus on too many, they may be obstructing your field of view.” – Scott M. Graffius

## Data

We used EDUs that are extracted from IMDB movie review dataset. The main metrics for this task is comparing train and test accuracy, along with looking at precision and recall. Want to insert a picture from your files or add a shape, text box, or table? You got it! On the Insert tab of the ribbon, just tap the option you need.

“Looks like we don’t have enough data.” – Ekrem Guzelyel

## Communication

Both of the undergraduate students worked separately on their own models.

“Did you push your changes to Github?” – Syed Hasan Rizvi

“No.” – Ekrem Guzelyel

## Logistic Regression

The main metrics for this task is comparing train and test accuracy, along with looking at precision and recall. Want to insert a picture from your files or add a shape, text box, or table? You got it! On the Insert tab of the ribbon, just tap the option you need.

“We use Logistic Regression as the baseline.” – Every ML Researcher Ever

## SVM

The main metrics for this task is comparing train and test accuracy, along with looking at precision and recall. Want to insert a picture from your files or add a shape, text box, or table? You got it! On the Insert tab of the ribbon, just tap the option you need.

“That’s an interesting idea.” – Caner Komurlu

## MNB

The main metrics for this task is comparing train and test accuracy, along with looking at precision and recall. Want to insert a picture from your files or add a shape, text box, or table? You got it! On the Insert tab of the ribbon, just tap the option you need.

“Stop, and take your time to breathe.” – Mustafa Bilgic

## LSTM

The main metrics for this task is comparing train and test accuracy, along with looking at precision and recall. Want to insert a picture from your files or add a shape, text box, or table? You got it! On the Insert tab of the ribbon, just tap the option you need.

“Did you try ConvLSTM layer?” – Anneke Soraya Hidayat

## CNN+LSTM

The main metrics for this task is comparing train and test accuracy, along with looking at precision and recall. Want to insert a picture from your files or add a shape, text box, or table? You got it! On the Insert tab of the ribbon, just tap the option you need.

“Ask Jay, he knows about CNNs better.” – Ruo Zhao

## GRU

The main metrics for this task is comparing train and test accuracy, along with looking at precision and recall. Want to insert a picture from your files or add a shape, text box, or table? You got it! On the Insert tab of the ribbon, just tap the option you need.

“Aim for simplicity in Data Science. Real creativity won’t make things more complex. Instead, it will simplify them.” – Damien Duffy Mingle

# Results

Data means prediction power. However, sometimes you can’t have enough data only with human sources. In order to classify movie reviews by IMDB, we use EDUs. Naturally, labeling all EDUs by hand consumes too much valuable time. Our job as Undergraduate Research Assistants at ML Lab at IIT is to build an efficient model that predicts the labels for the EDUs. This way we would be able to have a bigger dataset to train the actual text classification model. Tasks include:

* Labeling more EDUs on top of already labeled 2000 data points.
* Trying different combinations of different approaches to train a model that maximize the performance.

## What I Have Learned

The main metrics for this task is comparing train and test accuracy, along with looking at precision and recall. Want to insert a picture from your files or add a shape, text box, or table? You got it! On the Insert tab of the ribbon, just tap the option you need.

“Communication!” – Anneke Soraya Hidayat