IT3190E – Machine Learning

SENTIMENT ANALYSIS WITH IMDB DATASET

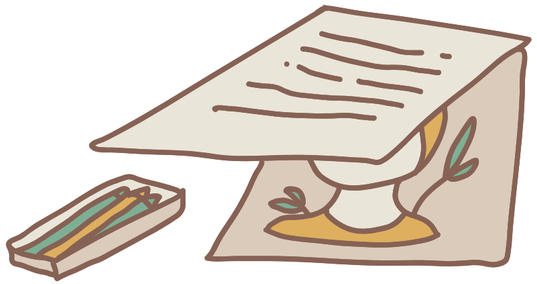
Outline

1. Problem Statement
2. Machine Learning Approach
3. Deep Learning Approach
4. Expanded Problem

Problem Statement

Dataset

IMDb Reviews is a large dataset for binary sentiment classification, consisting of 50,000 highly polar reviews (in English) with an even number of examples for training and testing purposes.
The dataset contains 50000 additional unlabelled data.The number of positive and negative sentiment is equal in the dataset (25000 each)



Example

Negative Review (0):
"After months of hype, we are left with a mess of a story that is mostly just CGI blobs, some light shows and some boring dialogue."

Positive Review (1):
"It was interesting to see how easily things can go really bad that even Thanos Crusade that was the big deal can seem so simple and trivial."

Machine Learning Approach

Data preprocessing

TF - IDF

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF

Term x within document y

$tf_{x,y}$ = frequency of x in y

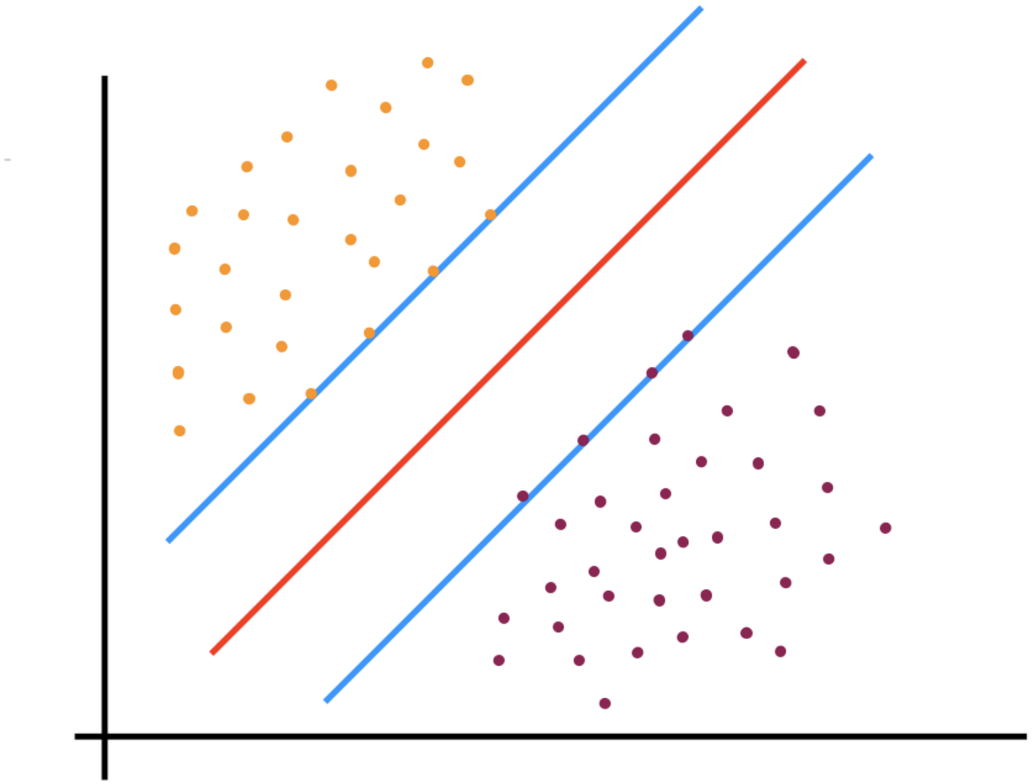
df_x = number of documents containing x

N = total number of documents

Term frequency - Inverse document frequency: is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus

Machine Learning approach

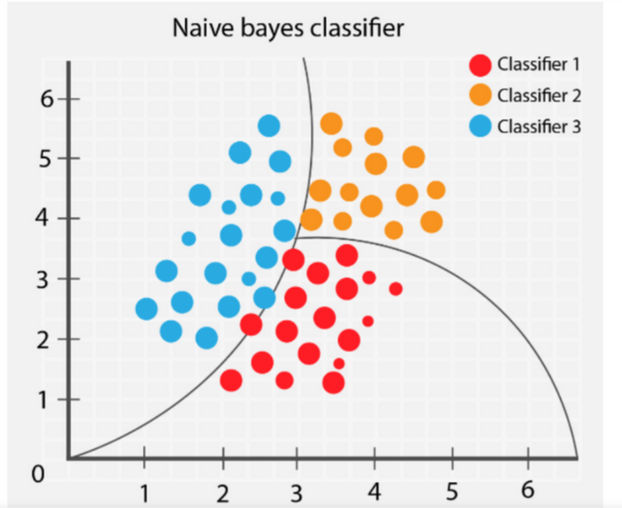
SVM



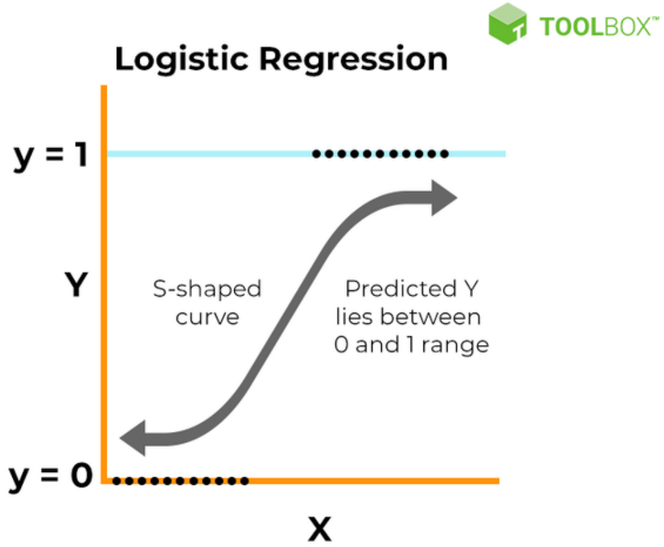
C = 100
kernel='linear'

Machine Learning approach

Simple Classifier

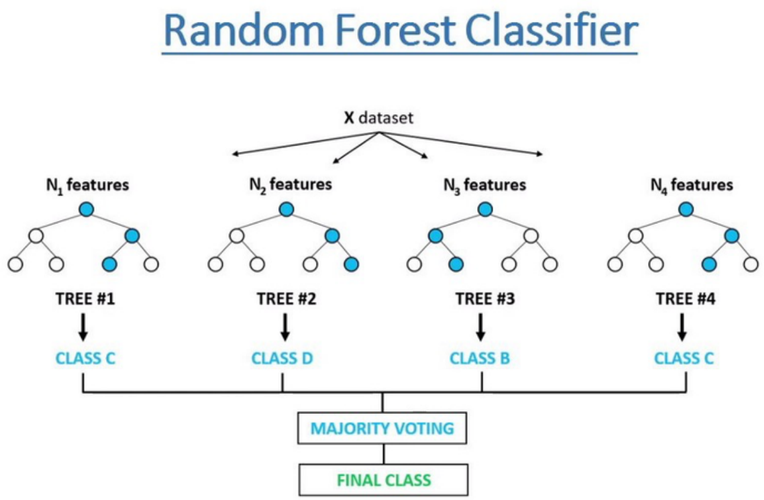


Naive Bayes



Logistic regression

C = 100



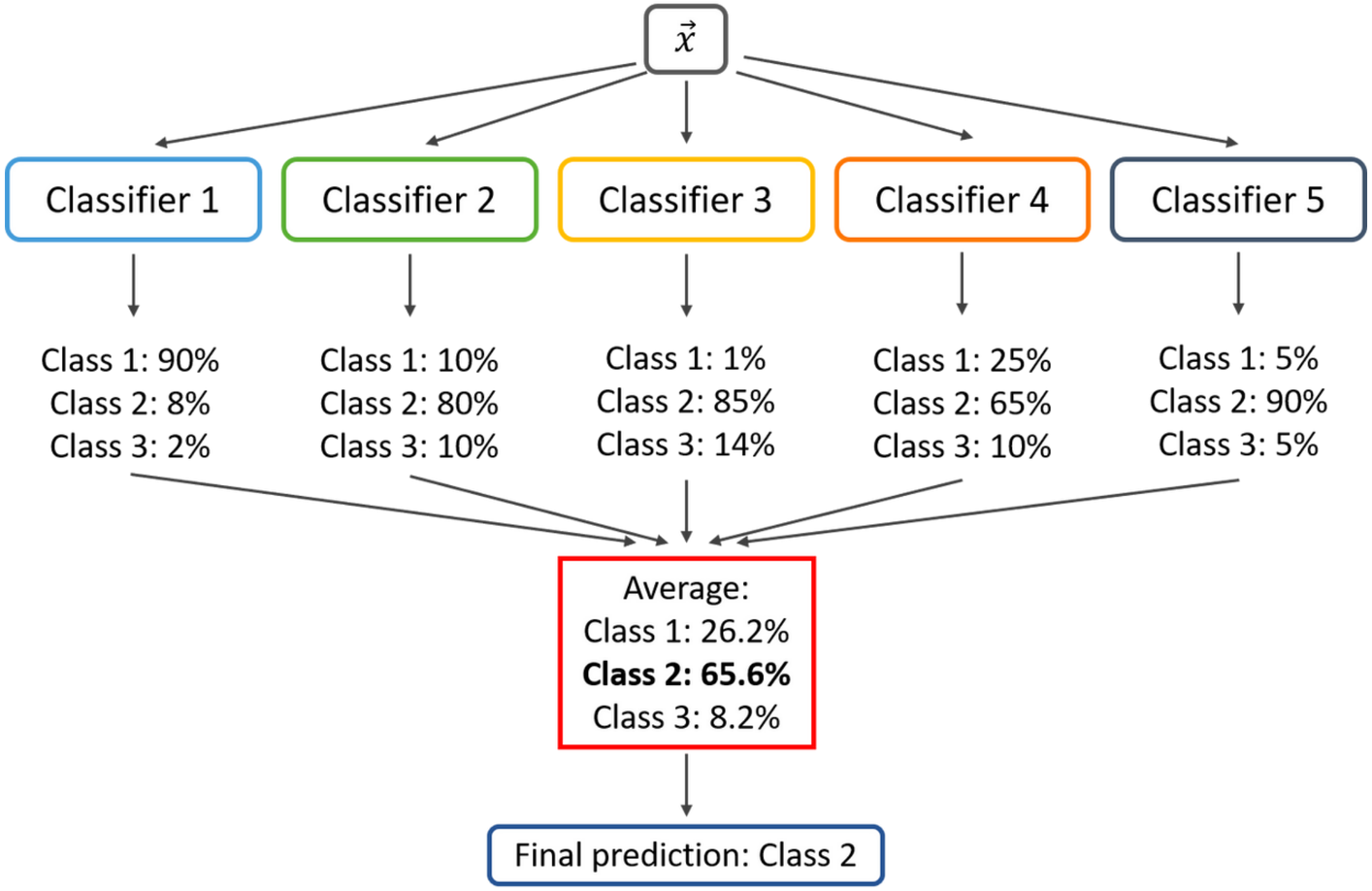
Random Forest

n_estimator = 250

criterion = 'entropy'

Machine Learning approach

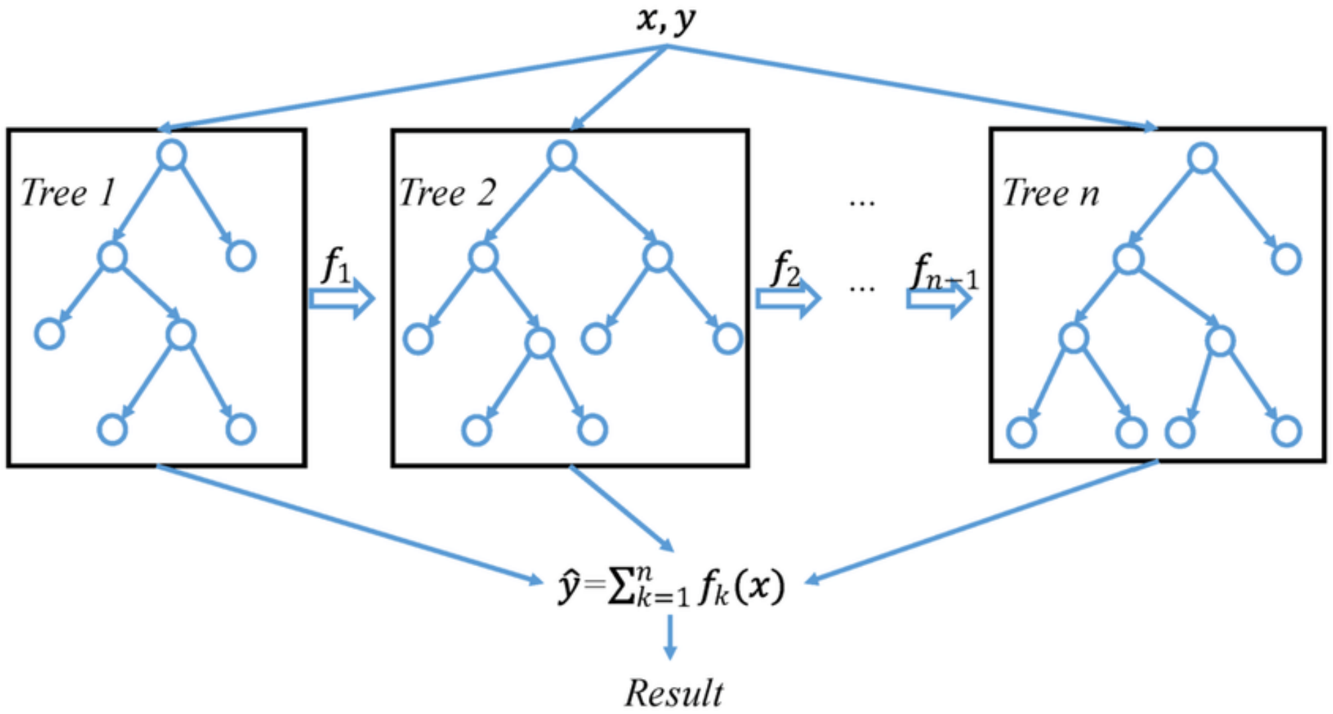
Voting Classifier



A machine learning model that trains on an ensemble of numerous models and predicts an output (class) based on their highest probability of chosen class as the output.

Machine Learning approach

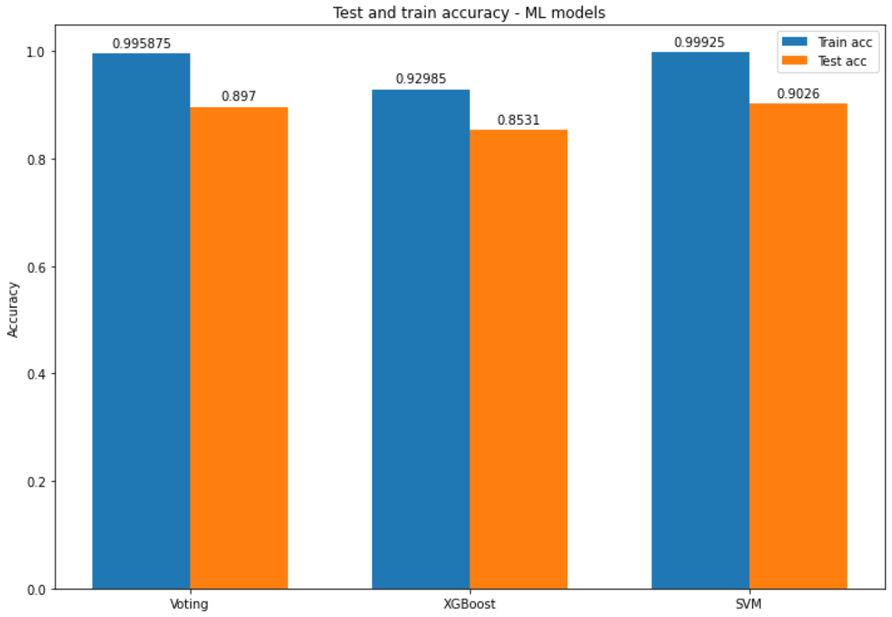
XGBoost

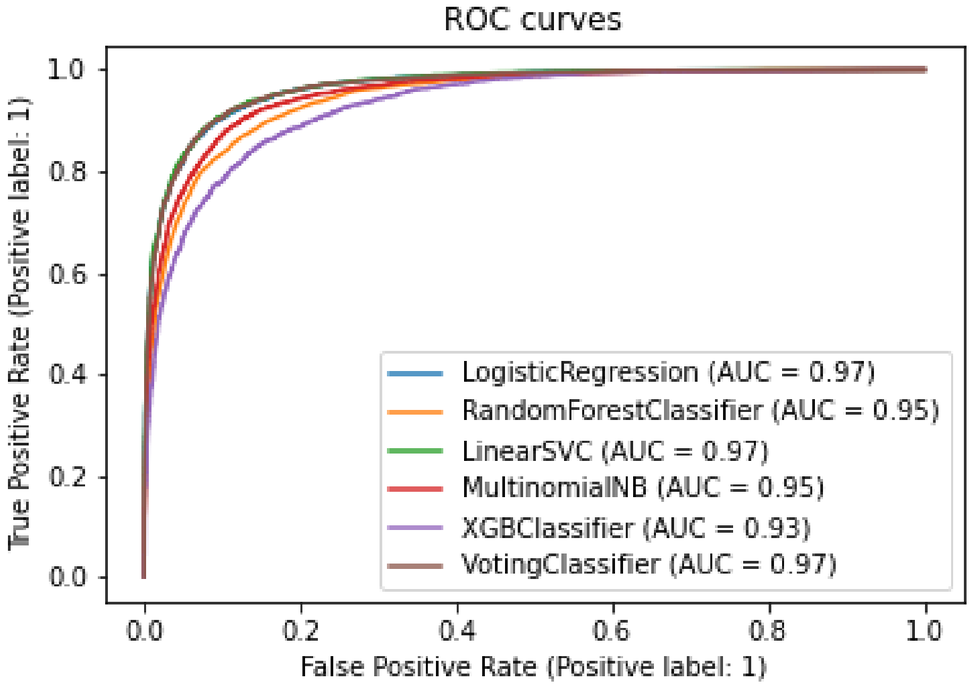


XGBoost is a popular and efficient open-source implementation of the gradient boosted trees algorithm. .

Machine Learning approach

Models evaluation





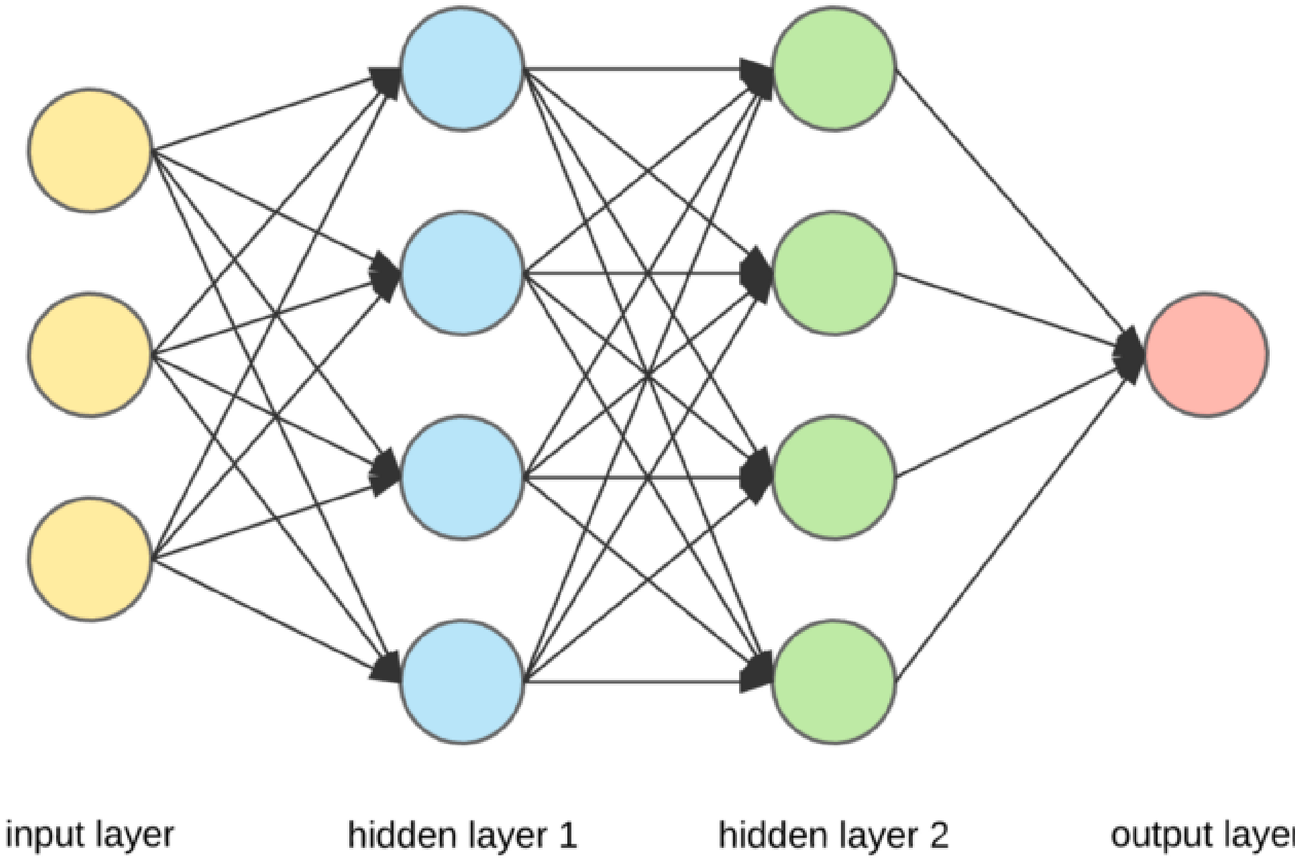
Deep Learning Approach

Neural Network

Neural network as a mathematical function

f: X => y

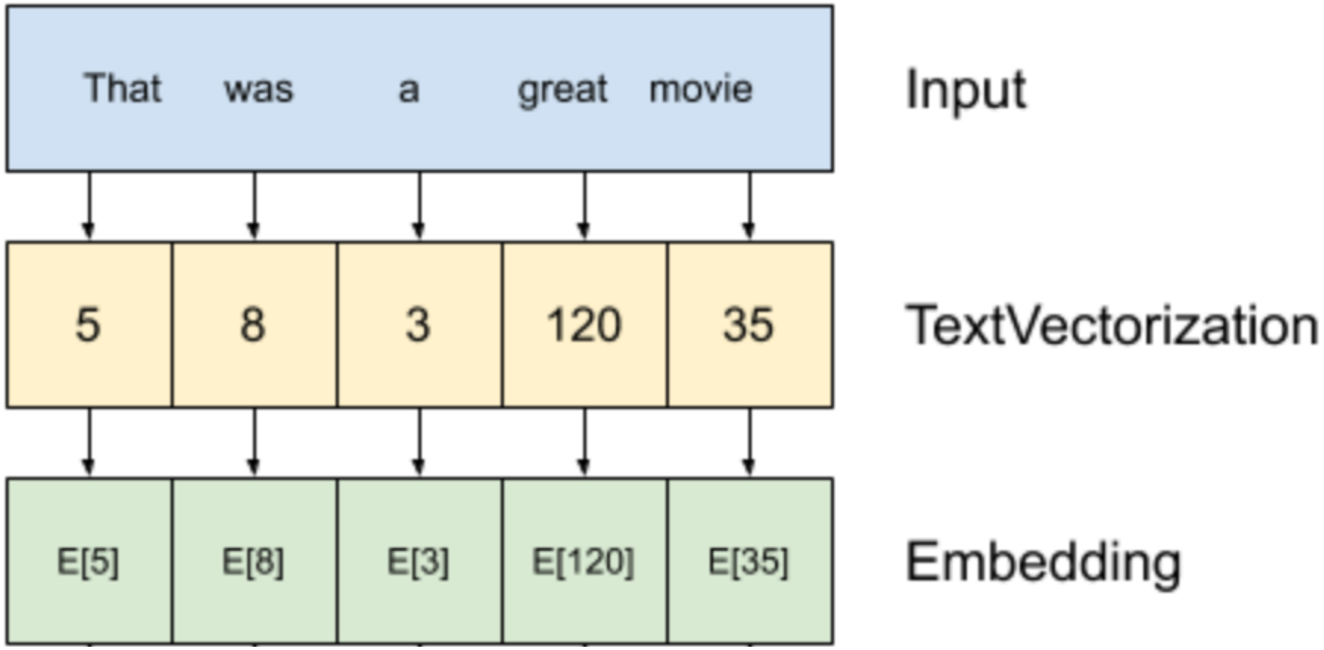
X: words?



Preprocessing: Word2Vec

lower and strip punctuation,
turn word into index

turns positive integers (indexes)
into dense vectors of fixed size

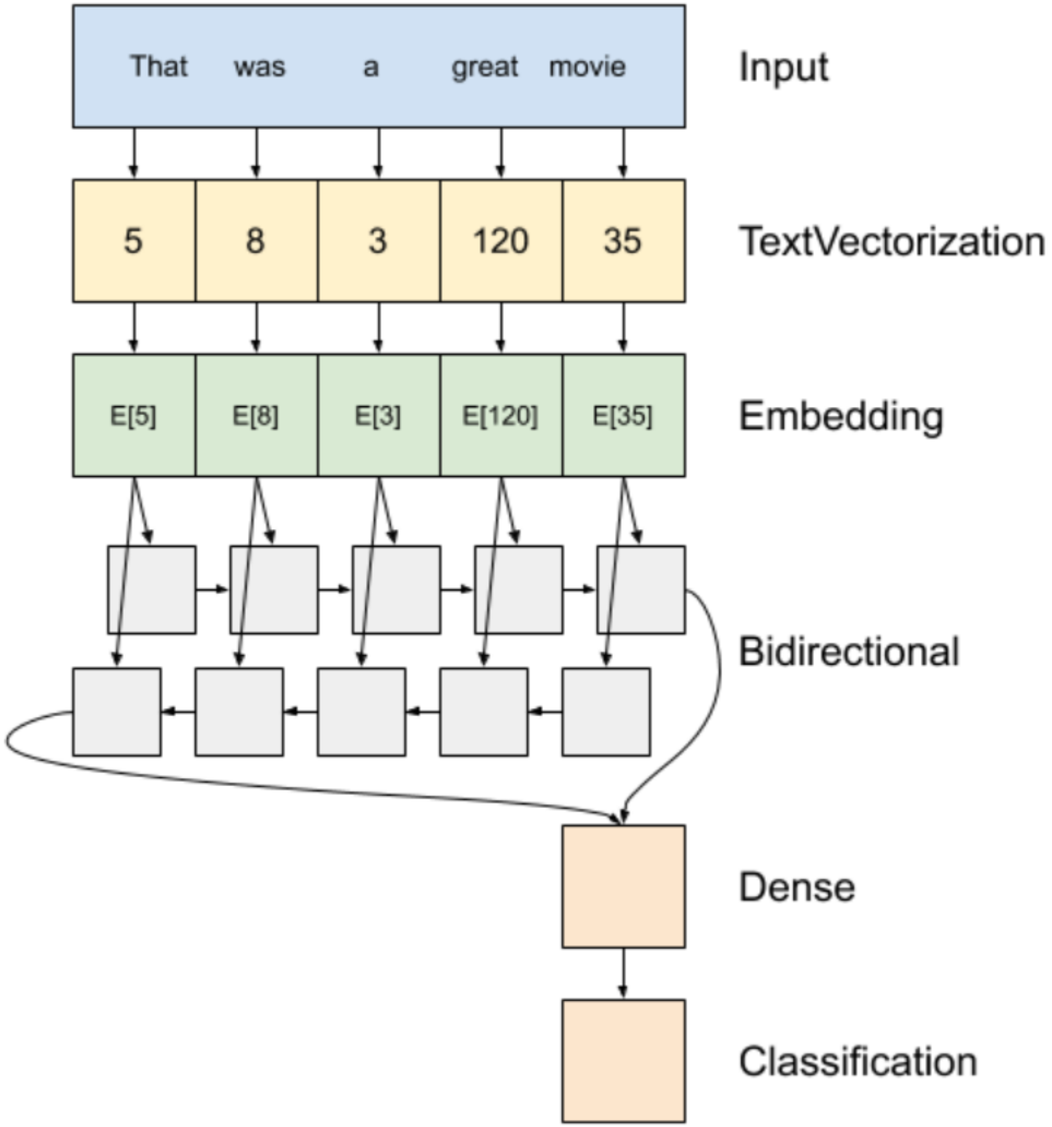


Word embeddings are a type of word representation that allows words with similar meaning to have a similar representation.

Bidirectional LSTM Neural Network

Word embedding

Feature Extraction



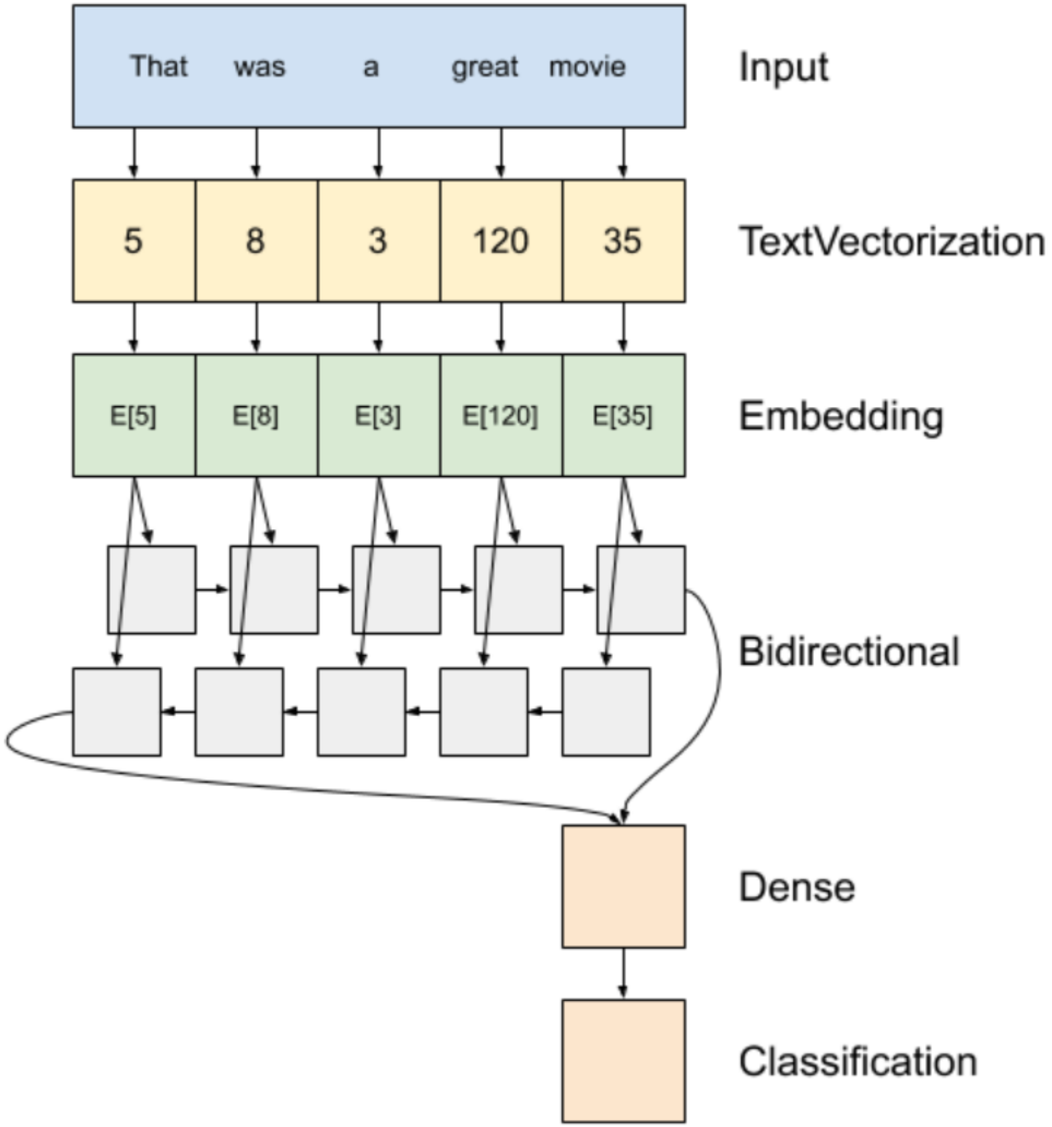
Deep Bidirectional LSTM Neural Network

Deep bidirectional LSTM

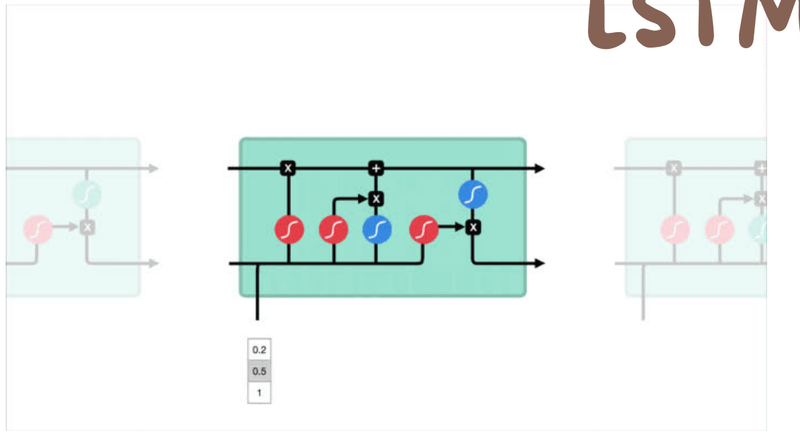
- Similar to bidirectional LSTM.

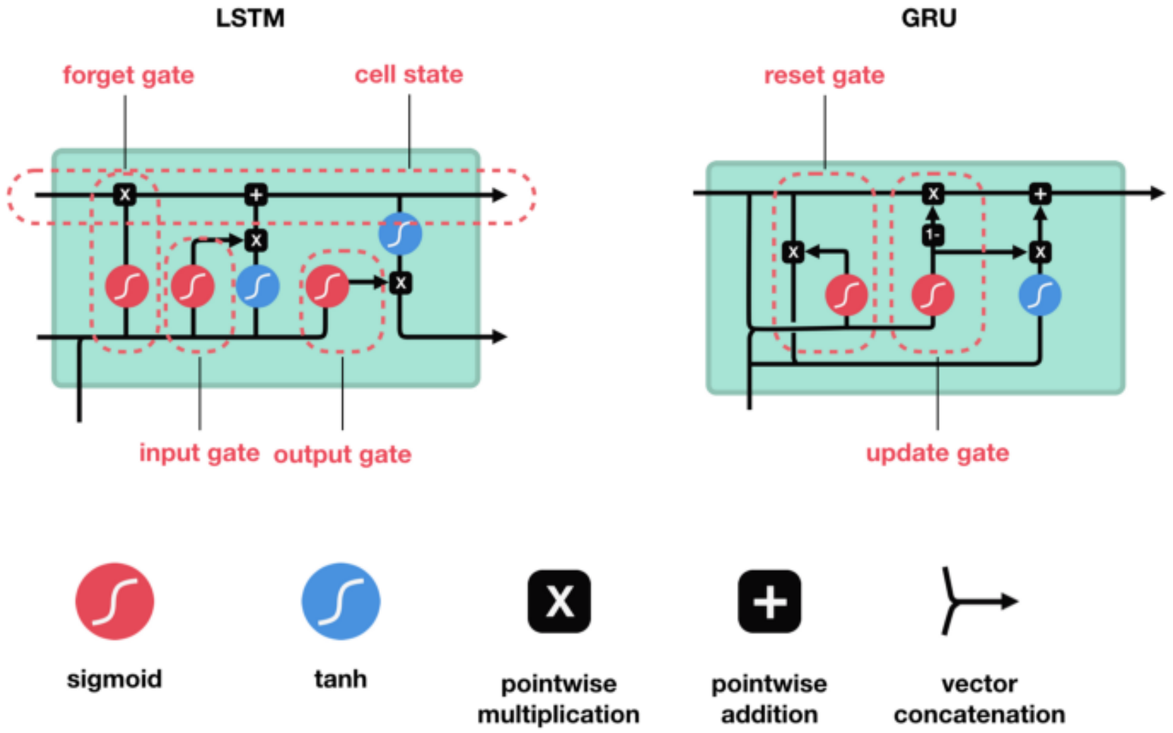
- Difference: multiple layers per time step.

- Higher learning capacity, need larger dataset



LSTM cell replacement: GRU





GRU - Gated recurrent unit

- Newer generation of Recurrent Neural networks

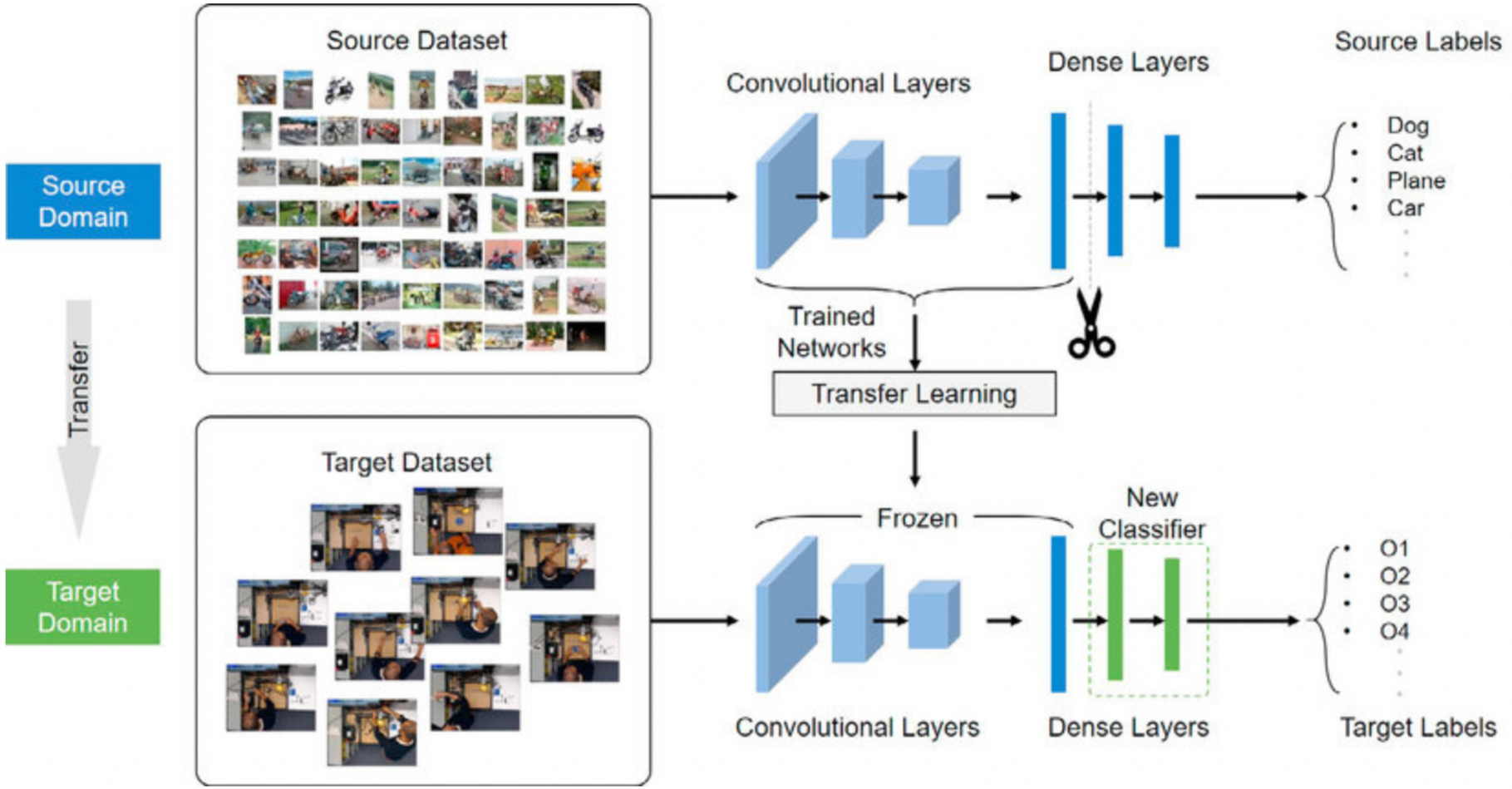
- Similar to an LSTM cell

- Used the hidden state to transfer information

Deep Learning Approach

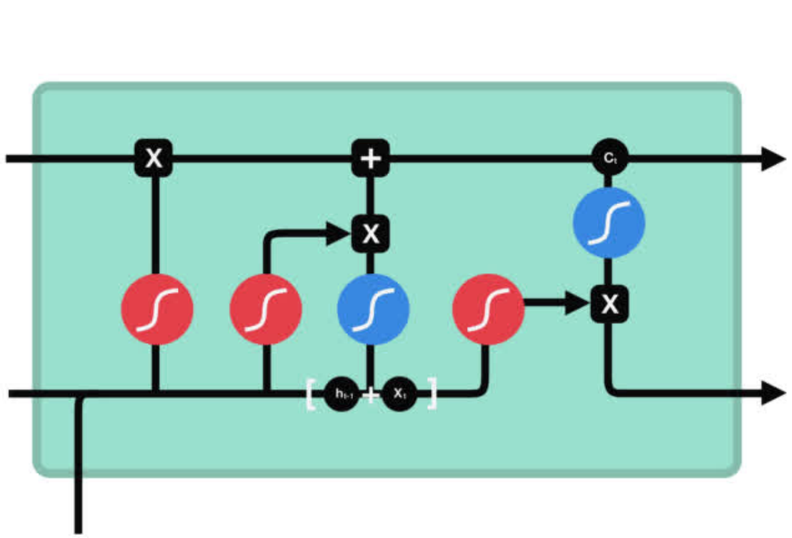
UMLFiT

Transfer Learning

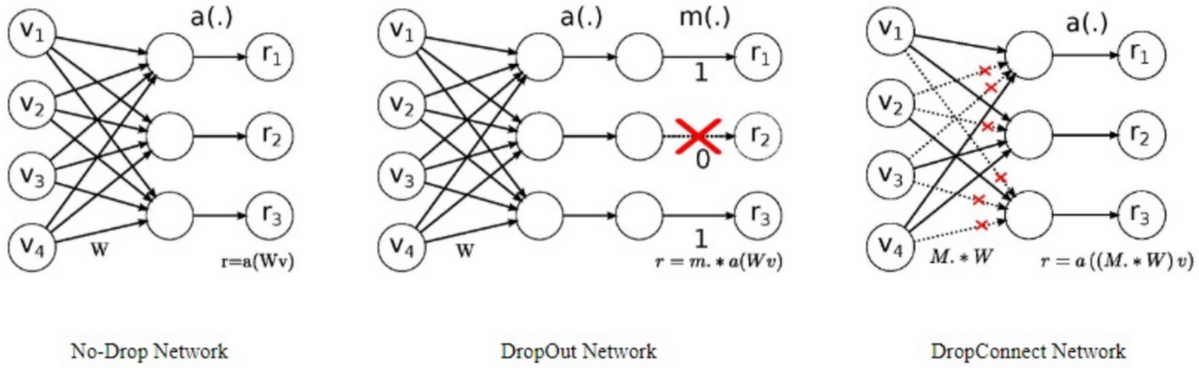


Transfer learning is the improvement of learning in a new task through the transfer of knowledge from a related task that has already been learned.

AWD-LSTM

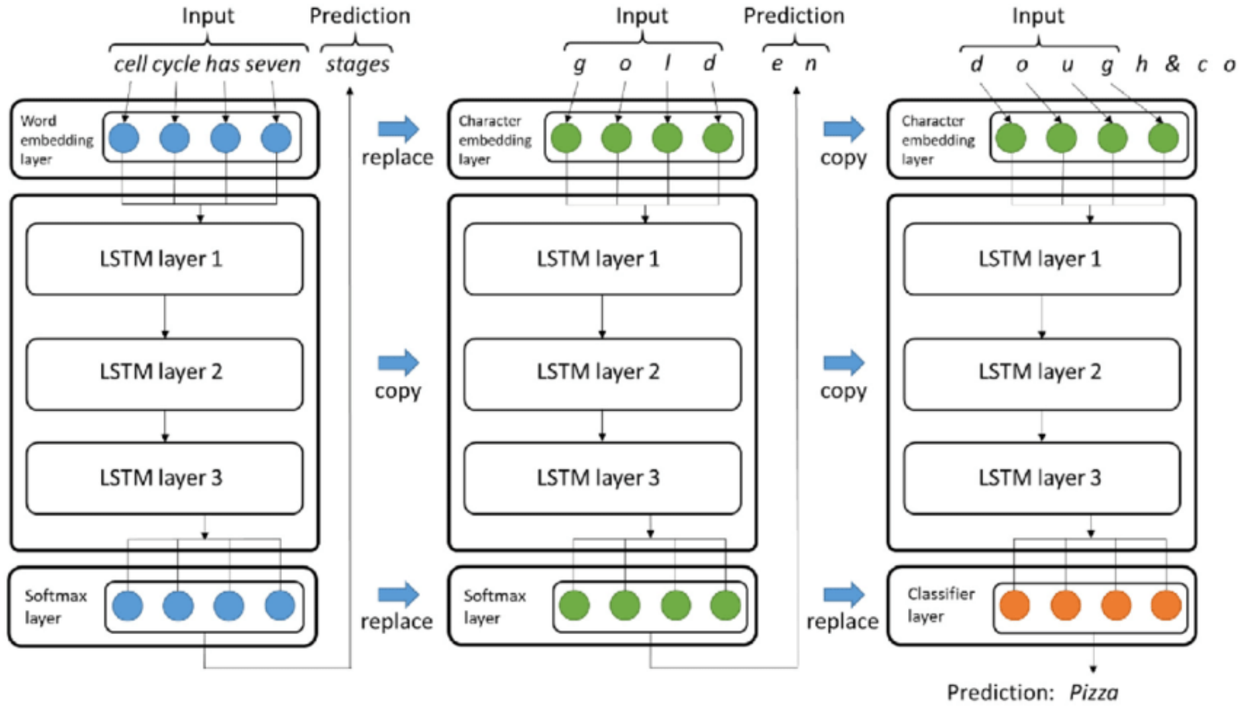


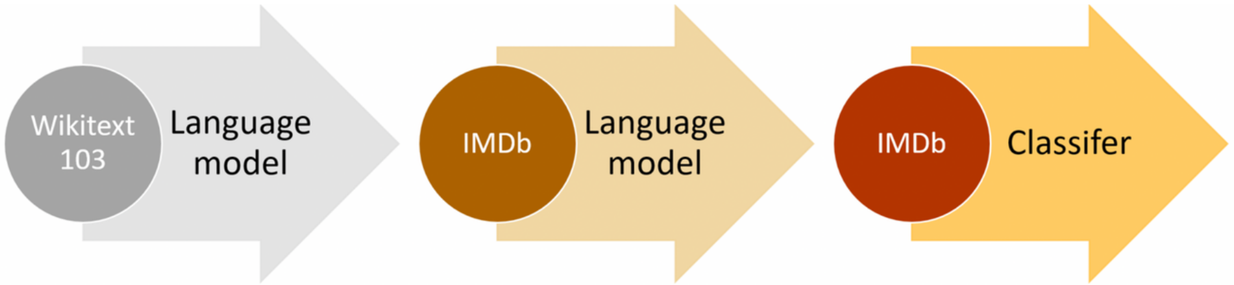
A regular LSTM with various tuned dropout hyperparameters



ASGD Weight-Dropped LSTM uses DropConnect and a variant of Average-SGD (NT-ASGD) along with several other wellknown regularization strategies.

UMLFiT - Universal Language Model Fine-tuning

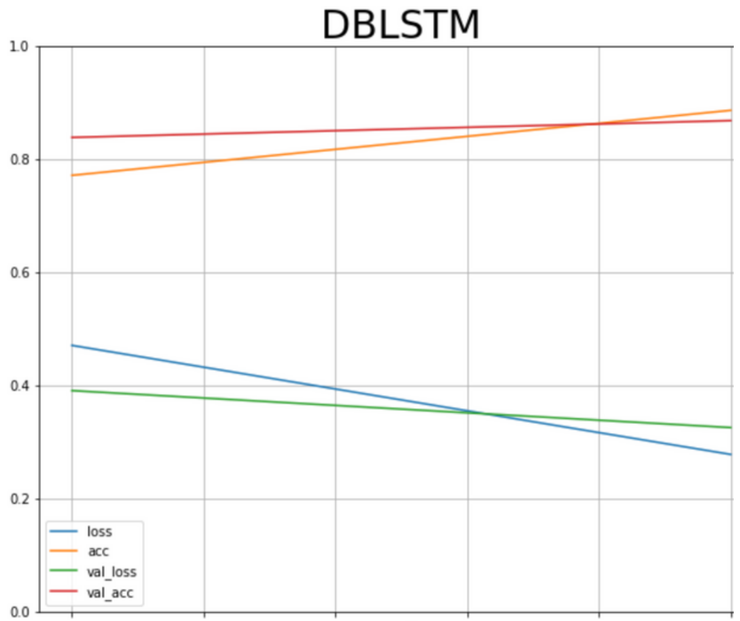


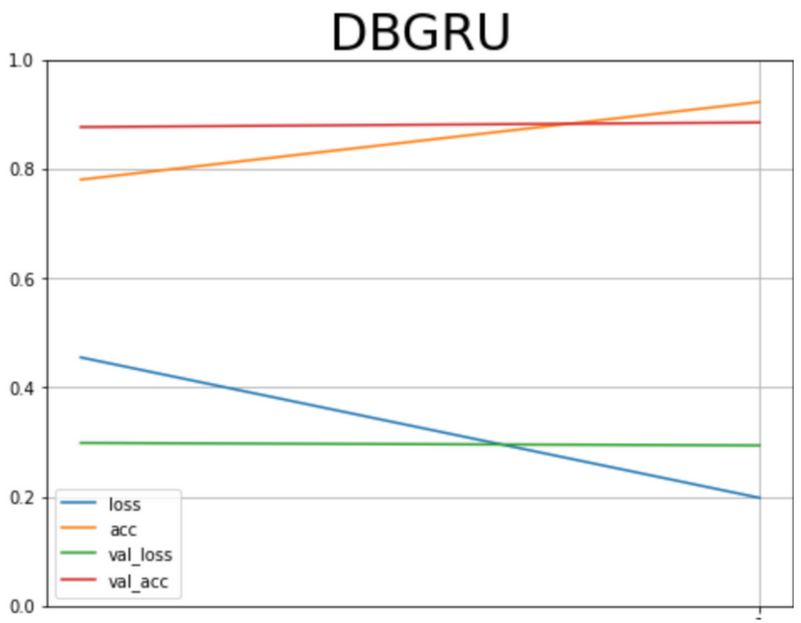


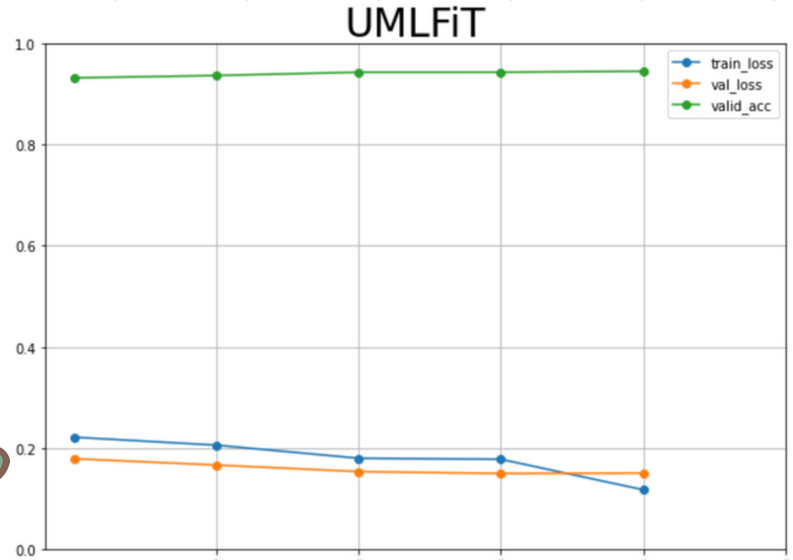
An architecture and transfer learning method, involves a 3-layers AWD-LSTM architecture for its representations.

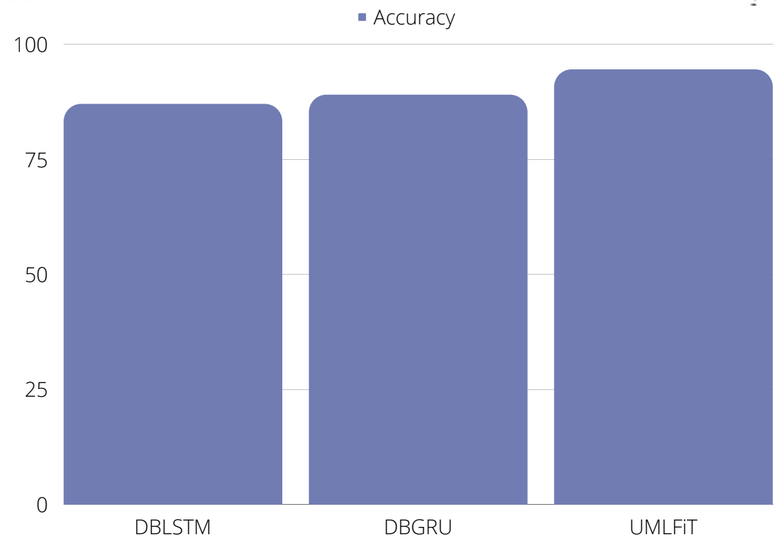
Three stages:
- General-domain LM pretraining
- Target task LM fine-tuning
- Target task classifier fine-tuning

Evaluation









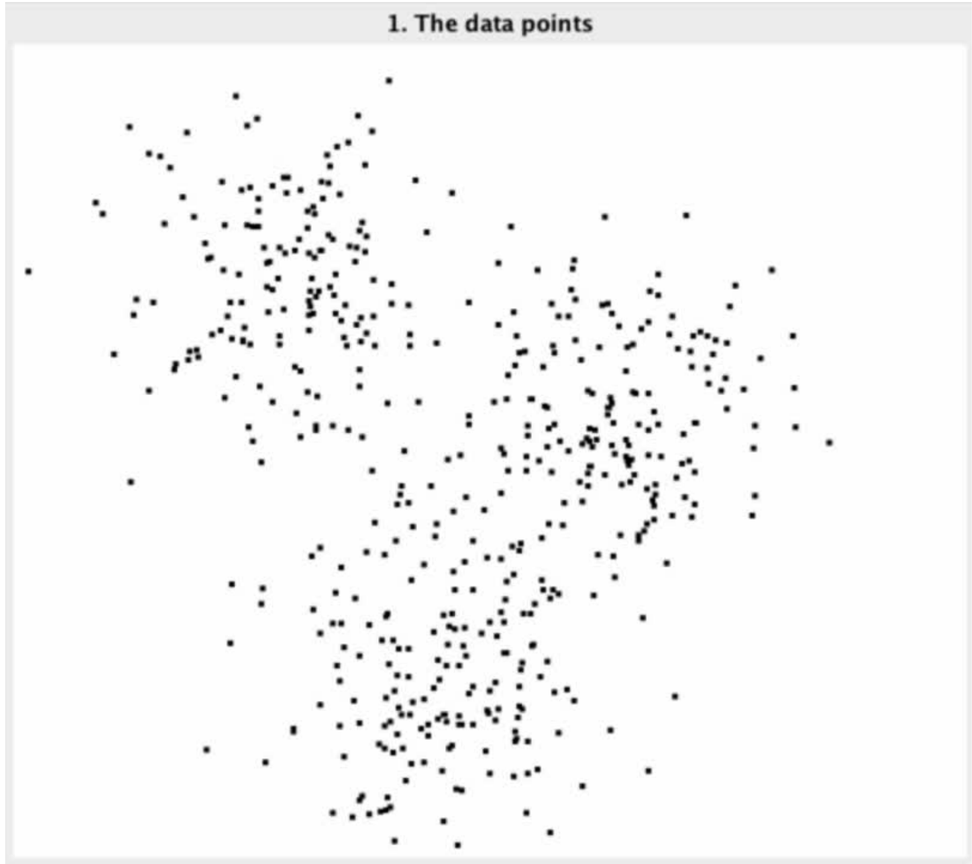
DBLSTM collected acc 87%
DBGRU collected acc 89%
after **2 epoches** train
(EarlyStop when val_loss increase avoid overfitting)

UMLFiT collected acc 94.5%
after **5 epoches** train
(3x unfrezze transfer layers)

Unsupervised Learning

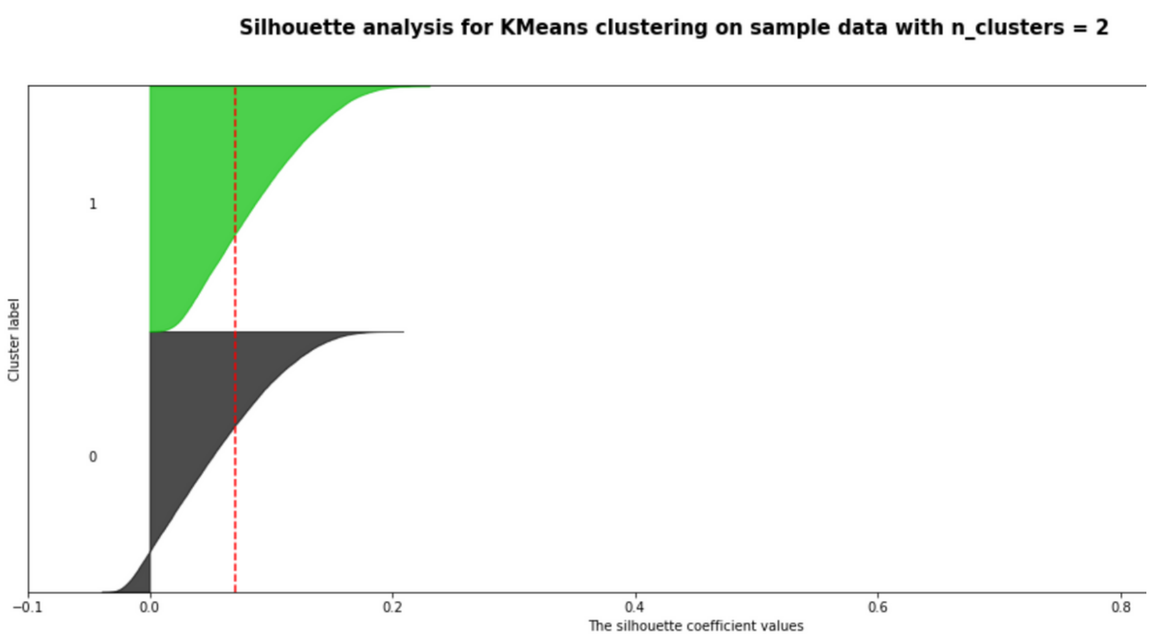
Pre-trained Word2Vec x Mini-batch K-means

Generating embedding vectors
using pre-trained model Word2Vec

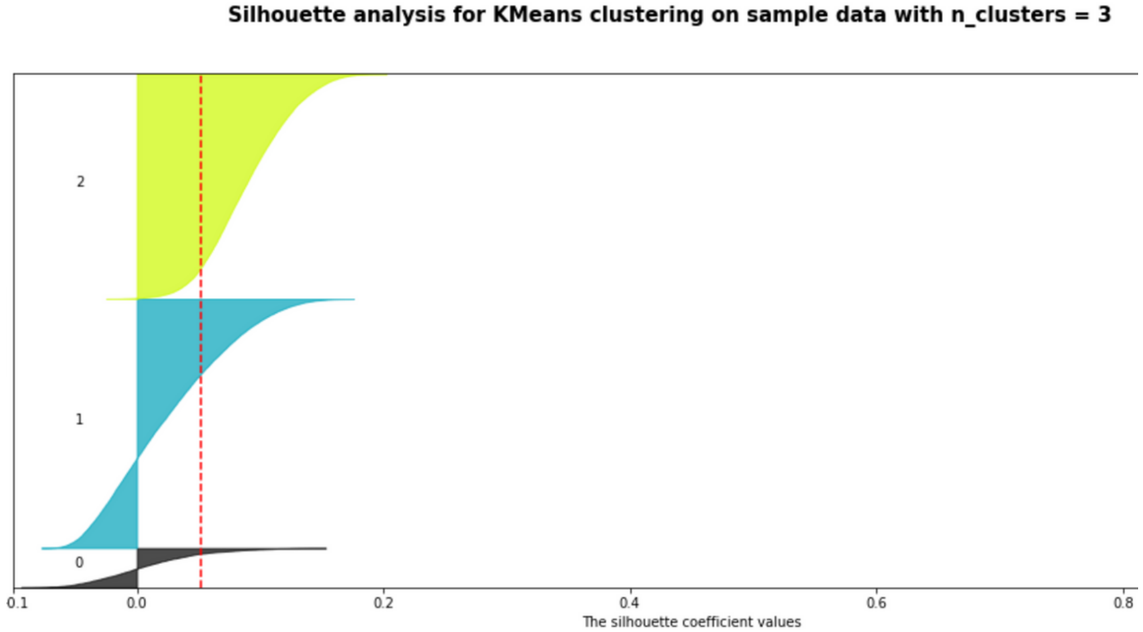


Apply Mini-batch K-means for clustering

Evaluation clustering



2 Clusters with hypothesis data has positive reviews and negative reviews



3 Clusters with hypothesis data has positive reviews, negative reviews and neutral reviews

Thank You!

Do you have any question?