

# Project Update Presentation

Semi-supervised Sequence Learning

Sabiha Sultana<sup>1</sup>, Piyakorn Munegan<sup>2</sup>,  
Sanjay Kanakkot Viswanathan<sup>3</sup>,  
Mohammed Rizwan Amanullah<sup>4</sup>

Department of Computing,  
Macquarie University

October 12, 2021

# Overview

- In this paper, we are replicating project which aims on unlabelled data to improve sequence learning with Recurrent network

## Previous approach - Bag of words on IMDB Datasets

Text	Sentiment
Looking for a REAL super bad movie? If you wanna have great fun, don't hesitate and check this one! Ferrigno is incredibly bad but is also the best of this mediocrity.	Negative
A professional production with quality actors that simply never touched the heart or the funny bone no matter how hard it tried. The quality cast, stark setting and excellent cinematography made you hope for Fargo or High Plains Drifter but sorry, the soup had no seasoning...or meat for that matter. A 3 (of 10) for effort.	Negative
The screen-play is very bad, but there are some action sequences that i really liked. I think the image is good, better than other romanian movies. I liked also how the actors did their jobs.	Negative

## Current approach

- (LM-LSTM) Language modelling. Predict what comes next in sequence.
- (SA-LSTM) Sequence auto encoder. Reads input sequence into vectors and predict the input sequence again.

# Sentiment analysis experiments with IMDB

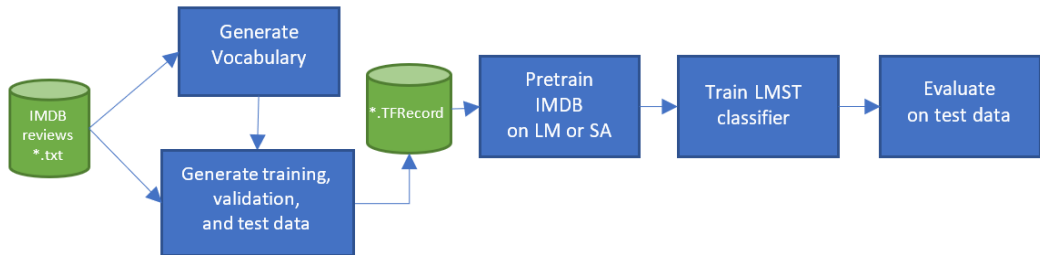
- Requirements: AWS(EC2) with Python 3, tensorflow v 1.15.5
- Dataset: IMDB movie review dataset
  - Training set: 25,000 labeled and 50,000 unlabeled reviews.
  - Test set: 25,000 labeled reviews.
  - The average length of each document is 241 words.

8\_7.txt - Notepad

File Edit Format View Help

Very good drama although it appeared to have a few blank areas leaving the viewers to fill in the action for themselves. I can

# Sentiment analysis experiments with IMDB cont.



- Accuracy of model LM-LSTM on 10% of the size of the original dataset is 0.52

# Output - Representation

```
I1010 09:40:11.504500 140117827979072 evaluate.py:96] Running batch 2150/2222...
INFO:tensorflow:Running batch 2150/2222...
I1010 09:40:12.019873 140117827979072 evaluate.py:96] Running batch 2150/2222...
INFO:tensorflow:Eval metric values:
I1010 09:40:12.021146 140117827979072 evaluate.py:109] Eval metric values:
INFO:tensorflow:accuracy = 0.525
I1010 09:40:12.021491 140117827979072 evaluate.py:113] accuracy = 0.525
INFO:tensorflow:Running batch 2160/2222...
I1010 09:40:12.494797 140117827979072 evaluate.py:96] Running batch 2160/2222...
INFO:tensorflow:Running batch 2170/2222...
I1010 09:40:12.985984 140117827979072 evaluate.py:96] Running batch 2170/2222...
INFO:tensorflow:Running batch 2180/2222...
I1010 09:40:13.452341 140117827979072 evaluate.py:96] Running batch 2180/2222...
INFO:tensorflow:Running batch 2190/2222...
I1010 09:40:13.978806 140117827979072 evaluate.py:96] Running batch 2190/2222...
INFO:tensorflow:Running batch 2200/2222...
I1010 09:40:14.470025 140117827979072 evaluate.py:96] Running batch 2200/2222...
INFO:tensorflow:Eval metric values:
I1010 09:40:14.471410 140117827979072 evaluate.py:109] Eval metric values:
INFO:tensorflow:accuracy = 0.518
I1010 09:40:14.471820 140117827979072 evaluate.py:113] accuracy = 0.518
INFO:tensorflow:Running batch 2210/2222...
I1010 09:40:14.922598 140117827979072 evaluate.py:96] Running batch 2210/2222...
INFO:tensorflow:Running batch 2220/2222...
I1010 09:40:15.430259 140117827979072 evaluate.py:96] Running batch 2220/2222...
INFO:tensorflow:Eval metric values:
I1010 09:40:15.573567 140117827979072 evaluate.py:109] Eval metric values:
INFO:tensorflow:accuracy = 0.516
I1010 09:40:15.574075 140117827979072 evaluate.py:113] accuracy = 0.516
WARNING:tensorflow:From evaluate.py:116: The name tf.train.get_global_step is deprecated. Please use tf.compat.v1.train.get_global_step instead.
I1010 09:40:15.574461 140117827979072 module_wrapper.py:139] From evaluate.py:116: The name tf.train.get_global_step is deprecated. Please use tf.compat.v1.train.get_global_step instead.
INFO:tensorflow:Finished eval for step 0
I1010 09:40:15.588693 140117827979072 evaluate.py:117] Finished eval for step 0
(base) ubuntu@ip-172-31-7-99:~/models/research/adversarial_text$
(base) ubuntu@ip-172-31-7-99:~/models/research/adversarial_text$ cd /tmp/models/imdb_eval
(base) ubuntu@ip-172-31-7-99:/tmp/models/imdb_eval$ ls
events.out.tfevents.1633858708.ip-172-31-7-99_graph.pbtxt
(base) ubuntu@ip-172-31-7-99:/tmp/models/imdb_eval$
```

Figure: Accuracy got after running Evaluate.py file

# Issue Analysis and Solutions

## The Issues Faced

1. Process Requirement for LSTM-RNN is huge.
2. Not Getting the Accuracy as expected.
3. AWS kills the process due to memory issues.
4. Eg : Took 2 days to process with half Data.
5. Installing dependencies  
Eg: **Tensor Flow 1.15.5, Wget,nltk,pinkt**
6. Filtering/ down sizing the file.

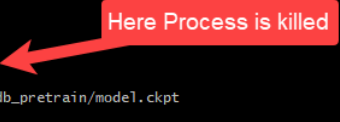
As a workaround during replication of the project we reduced the model size to 10 Percentage and set parameters to low and ran the process which resulted in getting **accuracy of 0.52** with nearly **7 hours of time.**

The folders neg, pos, unsup where filtered and **downsized to 10 percentage** of the raw data.

# Explanation - Issues

Figure: Error encountered when ran the file without scaling down

```
2021-10-09 14:31:38.922626: W tensorflow/core/framework/cpu_allocator_impl.cc:81] Allocation of 89095168 exceeds 10% of system memory.
2021-10-09 14:31:38.978610: W tensorflow/core/framework/cpu_allocator_impl.cc:81] Allocation of 89095168 exceeds 10% of system memory.
2021-10-09 14:31:39.034539: W tensorflow/core/framework/cpu_allocator_impl.cc:81] Allocation of 69605600 exceeds 10% of system memory.
2021-10-09 14:31:39.077938: W tensorflow/core/framework/cpu_allocator_impl.cc:81] Allocation of 69605600 exceeds 10% of system memory.
2021-10-09 14:31:39.133730: W tensorflow/core/framework/cpu_allocator_impl.cc:81] Allocation of 89095168 exceeds 10% of system memory.
INFO:tensorflow:Running local_init_op.
1009 14:31:41.073074 139933257013056 session_manager.py:500] Running local_init_op.
INFO:tensorflow:Done running local_init_op.
1009 14:31:41.168809 139933257013056 session_manager.py:502] Done running local_init_op.
INFO:tensorflow:global_step/sec: 0
1009 14:31:45.563056 139931913549568 supervisor.py:1099] global_step/sec: 0
INFO:tensorflow:Saving checkpoint to path /tmp/models/imdb_pretrain/model.ckpt
1009 14:31:45.563828 139931905156864 supervisor.py:1117] Saving checkpoint to path /tmp/models/imdb_pretrain/model.ckpt
killed
(base) ubuntu@ip-172-31-0-194:~/models/research/adversarial_text$
```



**Scaled-down** - We reduced the parameter on the code to proceed further.

- rnn-cell size, batch-size, max-steps are the parameters changed.

# Explanation - Issues Solution

## Pretrain IMDB Language Model

```
$ PRETRAIN_DIR=/tmp/models/imdb_pretrain
$ python pretrain.py \
  --train_dir=$PRETRAIN_DIR \
  --data_dir=$IMDB_DATA_DIR \
  --vocab_size=87007 \
  --embedding_dims=256 \
  1 --rnn_cell_size=1024 \
  --num_candidate_samples=1024 \
  2 3 --batch_size=256 \
  --learning_rate=0.001 \
  --learning_rate_decay_factor=0.9999 \
  4 --max_steps=100000 \
  --max_grad_norm=1.0 \
  --num_timesteps=400 \
  --keep_prob_emb=0.5 \
  --normalize_embeddings
```

changed parameters for new models

```
$ python pretrain.py \
> --train_dir=$PRETRAIN_DIR \
> --data_dir=$IMDB_DATA_DIR \
> --vocab_size=87007 \
> --embedding_dims=256 \
> --rnn_cell_size=64 \
> --num_candidate_samples=64 \
> --batch_size=1 \
> --learning_rate=0.001 \
> --learning_rate_decay_factor=0.9999 \
> --max_steps=10000 \
> --max_grad_norm=1.0 \
> --num_timesteps=400 \
> --keep_prob_emb=0.5 \
> --normalize_embeddings
```

Figure: On left the Default parameter, On right Parameters we used for replication



# New Data Generation

- As part of data creation, we have scrapped amazon.in website using selenium and BeautifulSoup for product listing and product reviews.
- We have extracted product listing and product reviews information for ten categories.
  - After extraction, we are manually tagging the product listing and reviews.
  - For product listing we are manually tagging the company names and for product review which having positive and negative reviews.
- **Issues faced:** For certain product categories, the alignment of the page is changing as a result, we had to find common **xpath** to find and extract product listing for all the ten categories.

# New Data Generation: Representation

## Sample Output file after web scrapping

```
graph.pbtxt  x  pretrain.py  x  dbpedia_rep.sh  x  IMDb8 Scrapping.ipynb  x  Product_names.txt  x
1 Philips MG5740/15, 12-in-1, Face, Hair and Body - Multi Grooming Kit. Dual Cut Blades for Maximum Precision, 80 Mins Run Time (Silver)
2 VEGA Men X3 Beard Trimmer For Men With Quick Charge, 90 Mins Run-time, Waterproof, For Cord & Cordless Use And 40 Length Settings, (VHTH-24)
3 Beurer HR 2000 Precision Cordless Nose, Ear & Eyebrow Trimmer Extra comb attachment with 3/6 mm, High-quality vertical stainless steel blade
  ,Battery-powered with 3 years warranty.
4 NOVA NG 1152 Cordless Rechargeable: 60 Minutes Runtime Multi Grooming Trimmer for Men (Blue)
5 PHILIPS BT1232/15 Skin-friendly Beard Trimmer - DuraPower Technology, Cordless Rechargeable with USB Charging, Charging indicator, Travel lock, No Oil
  Needed, Blue
6 MI Cordless Beard Trimmer 1C, with 20 length settings, 60 MINutes of usage, & USB Fast charging, black
7 Philips BT3211/15 corded & cordless Beard Trimmer with Fast Charge; 20 settings; 60 min run time
8 BSC Cordless Beard & Hair Trimmer for Men | 4 Trimming Combs upto 12 mm Length | 1.5hrs Run Time | 2 hrs Ultra Fast USB Charging | 2 yrs Warranty |
  Washable and Easy To Use
9 Philips BT3203/15 cordless rechargeable Beard Trimmer - 10 length settings; 45 min run time
10 Mi Corded & Cordless Waterproof Beard Trimmer with Fast Charging - 40 length settings
11
12 Nova NHT 1073 USB Rechargeable and Cordless: 60 Minutes Runtime Professional Hair Clipper for Men
13 Nova NHT 1076 Cordless: 30 Minutes Runtime Trimmer for Men (Black)
14 Philips BT3203/15 cordless rechargeable Beard Trimmer - 10 length settings; 45 min run time
15 Philips BT3221/15 corded & cordless Titanium blade Beard Trimmer - 20 length settings; 90 min run time
16 Mi Corded & Cordless Waterproof Beard Trimmer with Fast Charging - 40 length settings
17 Syska HT200 PRO BeardPro Cordless Rechargeable Trimmer - 10 Length Settings; 45 min Runtime (Black)
18 SYSKA HT3333K Corded & Cordless Stainless Steel Blade Grooming Trimmer with 60 Minutes Working Time; 10 Length Settings (Black)
19 Nova NHT-1045 Rechargeable Cordless: 30 Minutes Runtime Beard Trimmer for Men (Black)
20 Philips BG1025/15 Showerproof Body Groomer for Men
21 Philips QP2525/10 Cordless OneBlade Hybrid Trimmer and Shaver with 3 Trimming Combs (Lime Green)
22 Philips Nose Trimmer Nt3650/16, Cordless Nose, Ear & Eyebrow Trimmer with Protective Guard System, Fully Washable, Including AA Battery, 2 Eyebrow
  Combs, Pouch (Gray)
```

The End