SHUBHAMITRA DAS

Sobha Marvella,Bellandur, Bangalore, 560103 | C: 984-581-1138 | shubhamitra@gmail.com

Summary

Machine Learning practitioner for more than a year focussing on Natural Language Processing (NLP). Having equipped with 16 years of software engineering experience, my current passion is to solve business problems that require automation of cognitive processes.
I am currently on a career break of 2 years and have been preparing myself in the field of Machine Learning.

Past 16 years :

Nokia (Former Alcatel-Lucent Bell Labs) : Built modules on Performance Management for Optical Network Management Systems
Aricent (Former Hughes Software) : Built Services for Mediation Platform.

Data Science Knowledge and Skills

Machine Learning

- Machine Learning with Structured data.
- Dealt with numeric, categorical, time series data
- Handled Multiclass and Multilabel problems.
- Major emphasis on Deep Learning with unstructured text data (NLP)

NLP papers studied:

- Self Attention for classification (https://arxiv.org/pdf/1703.03130.pdf)
- Transformer network (Attention is all you need)
- Embedding from Language Model ELMo(https://arxiv.org/pdf/1802.05365v2.pdf)

Libraries used:

- Scikit-Learn
- Gensim, Glove, Word2Vec
- XGBoost
- Matplotlib
- Keras
- AllenNLP with Pytorch
- Pandas, Numpy

Hands-on Projects on Data Science

As a part of my studies, went through various online resources like, MOOCs, Youtube videos, Arxiv papers and Blogs.
Participated in few Kaggle (www.kaggle.com) competitions which gave me a peek into the fascinating array of real problems in various industries that Machine Learning/Deep Learning can solve. Below is a write-up on my learning on one of the competitions in NLP.

1. Toxicity-Challenge :
This competition (https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge) is a multi-class classification challenge where users are asked to classify text collected from social media into 6 different categories.
The scoring used to rank the competitors was ROC-AUC score.

2. Google Colaboratory :I ran all my models in Google Colaboratory (https://colab.research.google.com/) which is a free cloud offering from Google and provides a setup for running Machine Learning and Deep Learning projects. It is equipped with iPython Jupyter Notebook with Python 3 and K80 GPUs. This can be run for 12 hrs continuously. For the Keras models that I ran, used CuDNNLSTM and CuDNNGRU(Fast GPU version of native LSTM and GRU).

3. Imbalanced Dataset: The dataset is an imbalanced one, out of 159571 records present in dataset 143346 records are clean, which means they were not classified to any of the target categories.

4. Pre-processing/Hyperparameter Tuning : For most of the Deep Learning models/embeddings that I tried pre-processing the data did not influence the score much .
Hyperparameter tuning too had minor impact on the score.

5. Embedding : For the Embedding layer used various pre-trained embedding vectors mentioned below. Resnet seems to give the best result for a single model.
1. crawl-300d-2M.vec
2. glove.840B.300d.w2vec.txt
3. wiki.en.vec
4. glove.twitter.27B.200d.txt
5. GoogleNews-vectors-negative300.bin
6. numberbatch-en.txt

Concatenated Glove[300] and Resnet[300] vectors with a total dimension of 600 for the Embedding Layer, did not give a better result.

Added NLTK POS Tag information to the concatenated vector of dimension 601 :

Embedding Vector = Glove + Resnet + POS (300 + 300 + 1).
This yielded a slightly better result in the private leaderboard.

6. Data Augmentation : To tackle the problem of class imbalance in the Dataset, used a Data Augmentation Library available SMOTE ,but this did not help the score.

Found an effective and innovative way of Data Augmentation from another kaggler : (https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/discussion/48038)

This is basically translating English Text to French/Spanish/German ,and then back to English using Google Translate API. The newly translated text is then added to our training Dataset, this did help in improving the score. In the recent QANet paper from Google, I found Google did mention about it for improving Machine Comprehension.

7. Blending : Averaging/Weighted Averaging of the output files from the different single models had the best score on the LeaderBoard. It is due to the fact that errors from the different models average out and provide a more generic model.
My final score on the Private Leaderboard with 75% Test data was 0.9862 (AUC Score).

8. Code : More details and code in (https://github.com/shubhamitradas/Toxicity-Challenge

9. Currently working on adding ELMO Embedding with AllenNLP Library for the Toxicity Challenge.

Previous Experience

Technical Specialist
Nokia (Former Alcatel-Lucent, 2004 to 2016) — Bangalore

OMS (Optical Management System) is a network management solution for optical networks for both SONET and SDH. Extensively worked in the Software development in both NMS (Network Management System)/EMS (Element Management System) ,have been mostly involved in design and development of new features for Performance Management.

Senior Software Engineer
Aricent(Former Hughes Software Systems, 2000 to 2004) — Bangalore

Worked in the HSS Mediation product which provides complete mediation solution for telecom operators. It provides online service provisioning including receipt of service requests from downstream application, generation of commands for network elements and processing of responses from network elements.

Education and Training

B.E(Electrical Engineering) 74% 1999
National Institute Of Technology — Silchar, India

Trainings and Certification

1. Completed Deep Learning Course by Andrew NG from Coursera .