Toxic Comments Prediction for Social Media Platform

James Liao 490851

Content

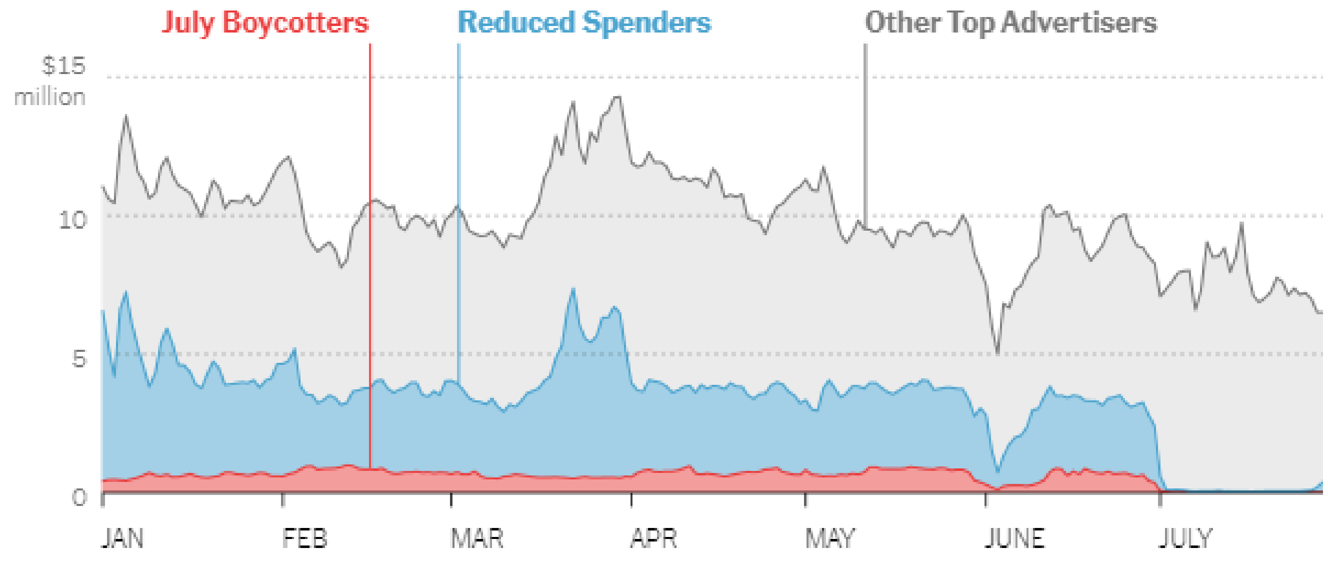
- Introduction
- Problem Description
- Database Background and Data Preprocessing
- Model Building
- Model Result
- Conclusion

Introduction

- Research about toxic comments prevention is increasing
- What is toxic comments? Comments with malicious intent/ Discrimination on sexual, race, personality…etc
- Social media platform is facing increasing number of disrespectful posting
- Our goal: Establish deep neural network-based model to predict and filter toxic comments (CNN, CNN-RNN, and CNN-LSTM)

Problem Description

Estimated spending of Facebook's top 100 advertisers



Note: "Reduced spenders" are companies that did not officially announce boycotts, but decreased their spending in July by at least 90 percent compared to June. • Source: Pathmatics • By Eleanor Lutz

- Several companies announced stop collaborating with Facebook due to its promotion and ignorance of spreading hated comments.
- Almost 50% of existing advertisers stop using Facebook to advertise their products.
- The price of ignore toxic comments is huge -> revenue loss

#StopHateForProfit

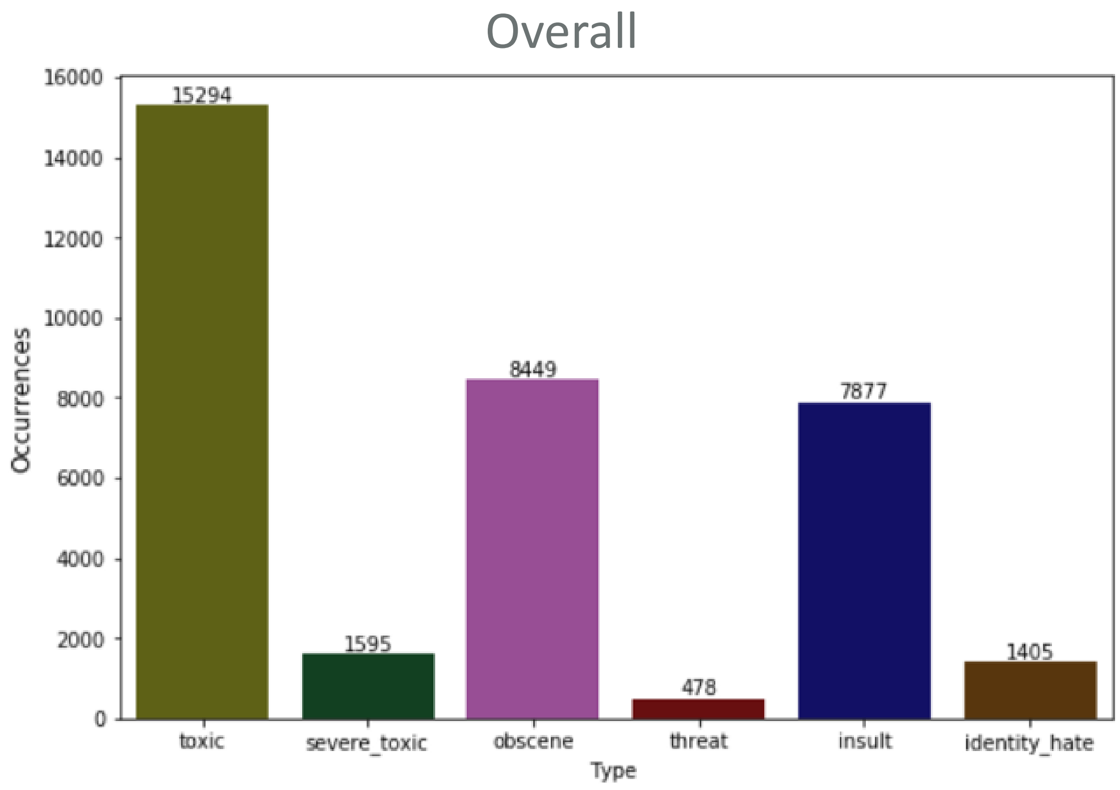
Dataset Background & Model Pre-Processing

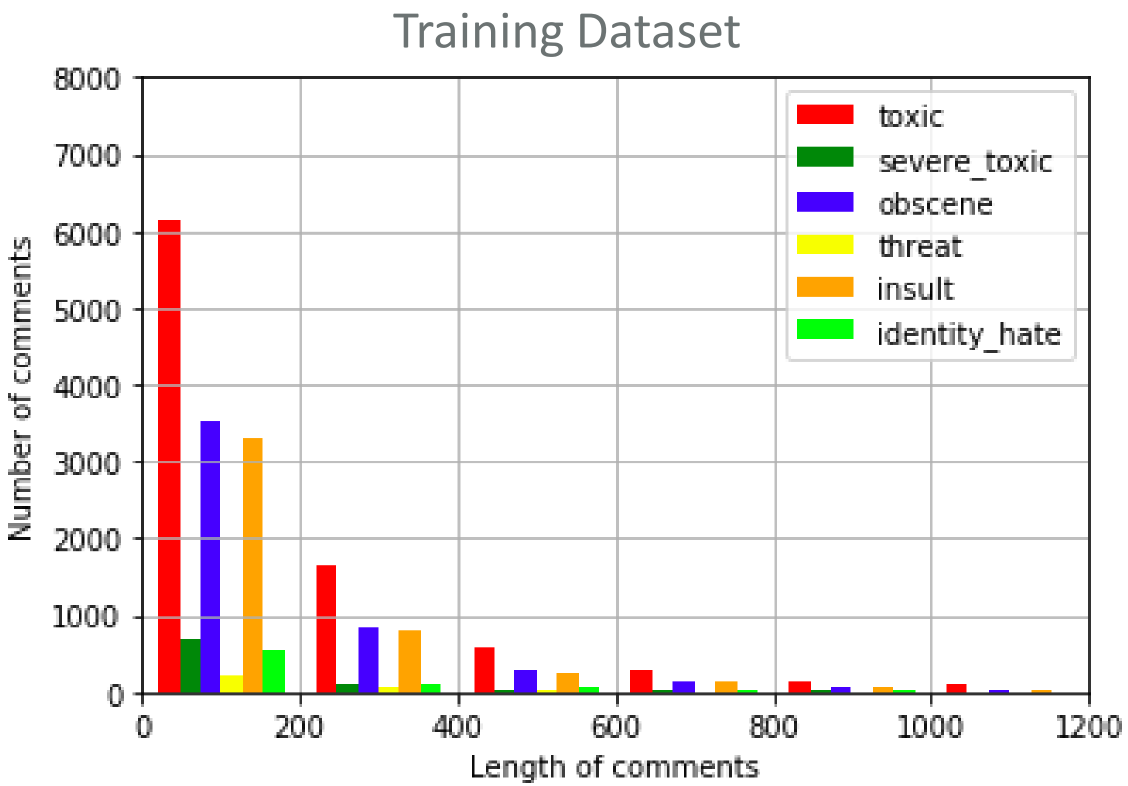
Dataset Background

- Data Download Source: Kaggle competition- Google Jigsaw team
- Data collected source: Wikipedia toxic comments
- 159, 571 rows
- Columns: “text_id”, “comment_text”, “toxic”, “severe_toxic”, “obscene”, “threat”, “insult”, “identity_hate”

Dataset Background & Model Pre-Processing

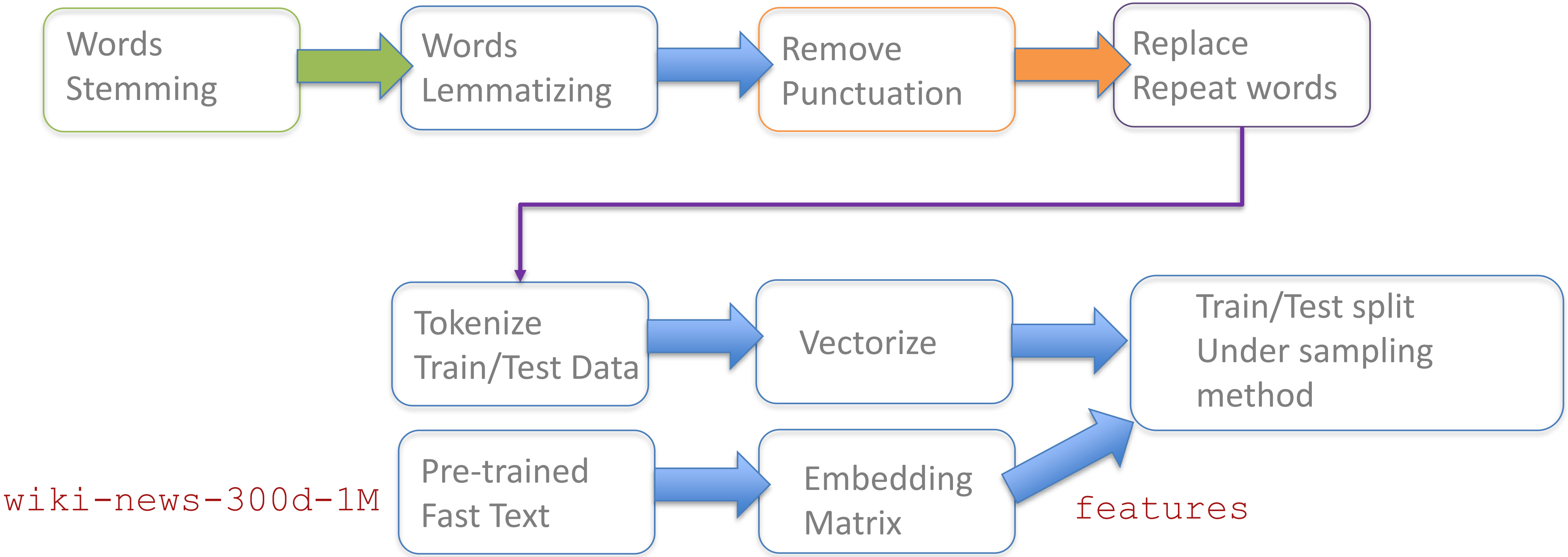
Dataset Background





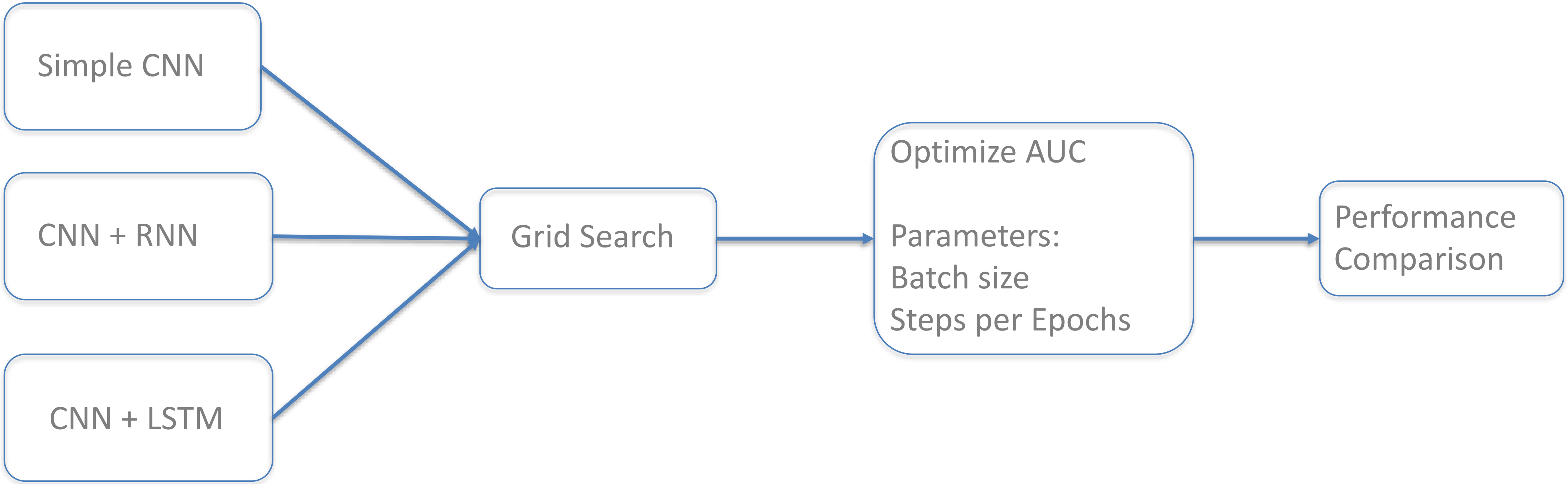
Dataset Background & Model Pre-Processing

Model Pre-Processing



Model Building

General Steps



Model Building

Grid Search Result- Testing AUC

CNN

Steps/ Batch	30	35	40	45	50
20	0.90	0.9596	0.95	0.93	0.91
30	0.89	0.92	0.92	0.90	0.90
50	0.89	0.87	0.89	0.86	0.87
100	0.78	0.79	0.81	0.79	0.8
150	0.75	0.85	0.79	0.80	0.70

CNN+LSTM

Steps/ Batch	30	35	40	45	50
20	0.93	0.92	0.92	0.91	0.93
30	0.92	0.94	0.94	0.92	0.89
50	0.94	0.9684	0.965	0.93	0.88
100	0.85	0.91	0.88	0.89	0.9
150	0.82	0.82	0.83	0.88	0.87

Model Building

Model Comparison

CNN + RNN

Layer (type)	Output Shape	Param #
embedding_31 (Embedding)	(None, 50, 300)	55831200
simple_rnn_1 (SimpleRNN)	(None, 50, 60)	21660
conv1d_25 (Conv1D)	(None, 50, 128)	38528
max_pooling1d_22 (MaxPooling1D)	(None, 16, 128)	0
global_max_pooling1d_22 (GlobalMaxPooling1D)	(None, 128)	0
batch_normalization_22 (BatchNormalization)	(None, 128)	512
dense_44 (Dense)	(None, 50)	6450
dropout_22 (Dropout)	(None, 50)	0
dense_45 (Dense)	(None, 6)	306

Total params: 55,898,656
Trainable params: 55,898,400
Non-trainable params: 256

CNN + LSTM

Layer (type)	Output Shape	Param #
embedding_32 (Embedding)	(None, 50, 300)	55831200
lstm_layer (LSTM)	(None, 50, 60)	86640
conv1d_26 (Conv1D)	(None, 50, 128)	38528
max_pooling1d_23 (MaxPooling1D)	(None, 16, 128)	0
global_max_pooling1d_23 (GlobalMaxPooling1D)	(None, 128)	0
batch_normalization_23 (BatchNormalization)	(None, 128)	512
dense_46 (Dense)	(None, 50)	6450
dropout_23 (Dropout)	(None, 50)	0
dense_47 (Dense)	(None, 6)	306

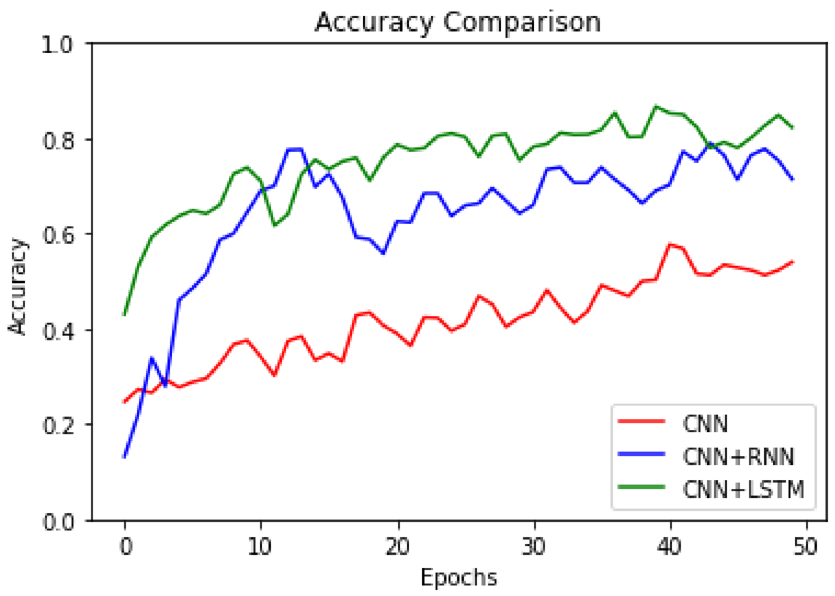
Total params: 55,963,636
Trainable params: 55,963,380
Non-trainable params: 256

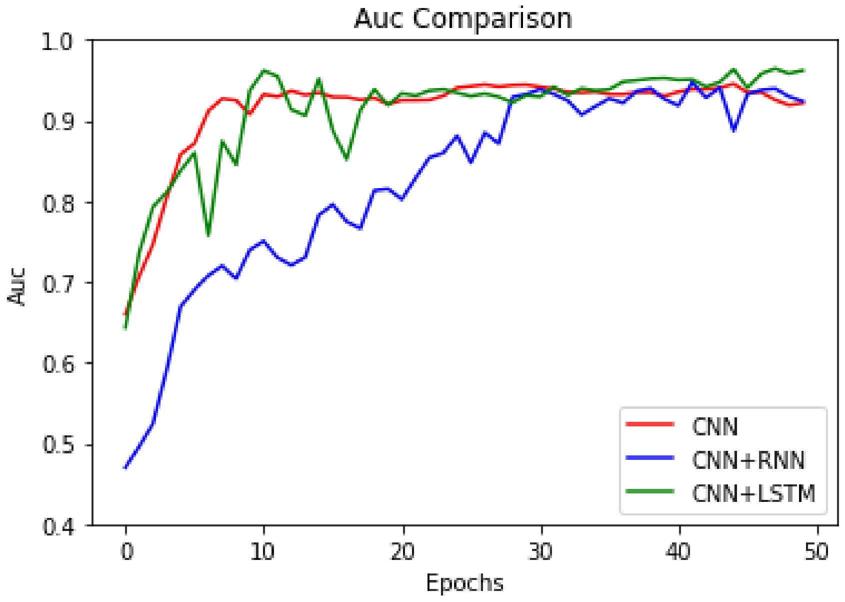
CNN

Layer (type)	Output Shape	Param #
embedding_33 (Embedding)	(None, 50, 300)	55831200
conv1d_27 (Conv1D)	(None, 50, 128)	192128
max_pooling1d_24 (MaxPooling1D)	(None, 16, 128)	0
global_max_pooling1d_24 (GlobalMaxPooling1D)	(None, 128)	0
batch_normalization_24 (BatchNormalization)	(None, 128)	512
dense_48 (Dense)	(None, 50)	6450
dropout_24 (Dropout)	(None, 50)	0
dense_49 (Dense)	(None, 6)	306

Total params: 56,030,596
Trainable params: 56,030,340
Non-trainable params: 256

Model Result





	CNN	CNN + RNN	CNN + LSTM
AUC	0.9596	0.961	0.9684
Accuracy	0.8657	0.8840	0.9077

Conclusion and Future Works

Brief Summary

- In this project, we establish the three neural networks-based classification model to predict the toxic text messages
- CNN+LSTM has best accuracy and AUC score
- The featured model could be implemented in social media platforms such as Facebook, Twitter, or Reddit

Future Works

- More advanced model used (Bi-GRU, Bi-LSTM, Attention Layers….etc)
- K-fold validation for training dataset

References

- Banerjee, Ling, et al, Comparative effectiveness of convolutional neural network (CNN) and recurrent neural network (RNN) architectures for radiology text report classification, Artificial Intelligence In Medicine, 2019

- Sharma, Chaurasia, et al, Sentimental Short Sentences Classification by Using CNN Deep Learning Model with Fine Tuned Word2Vec, Procedia Computer Science, 2020

- Kohli, Kuehler, & Palowith, and Paying attention to toxic comments online, Havard Business Review

- Li, Wang, & Xu, Chinese Text Classification Model Based on Deep Learning, Future Internet, 2018

- Sari, Rini,& Malik, Text Classification Using Long Short-Term Memory with GloVe Features, Jurnal Ilmiah Teknik Elektro Komputer dan Informatika, 2020

- Wang, Hsiao, & Chang, Automatic paper writing based on a RNN and the TextRank algorithm, Applied Soft Computing Journal, 2020