Toxic Comments Classification

ECE 539 Project-Fall 2018

Deepan Das(MS-ECE, UW Madison)

Executive Summary: *Problem at hand*

Objective:

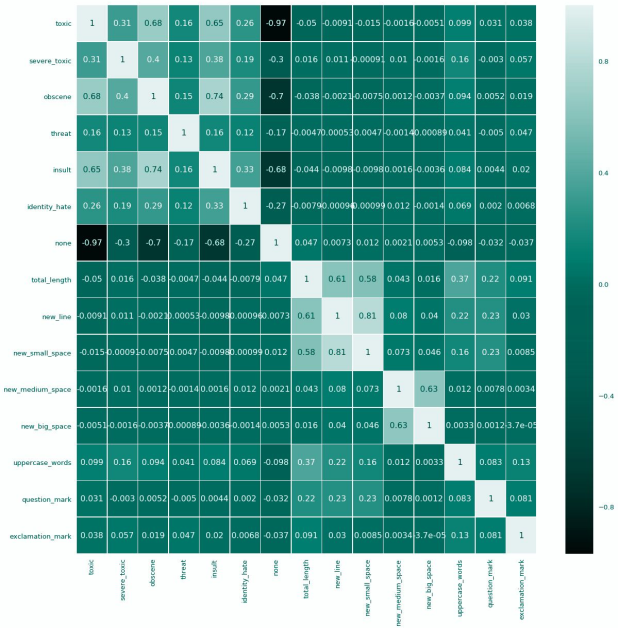
Explore and develop models for **toxic comment detection** based solely on **textual information.**

Self-Assessment:

- Implemented 3 distinct models
- Used open-source code to build these models, but developed on them later with visible improvements

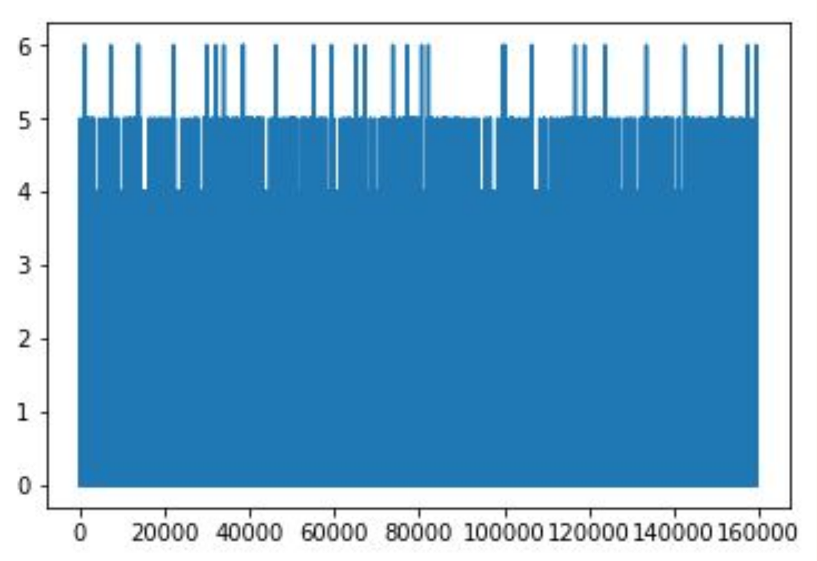
Dataset: EDA and Insights

Correlation between attributes

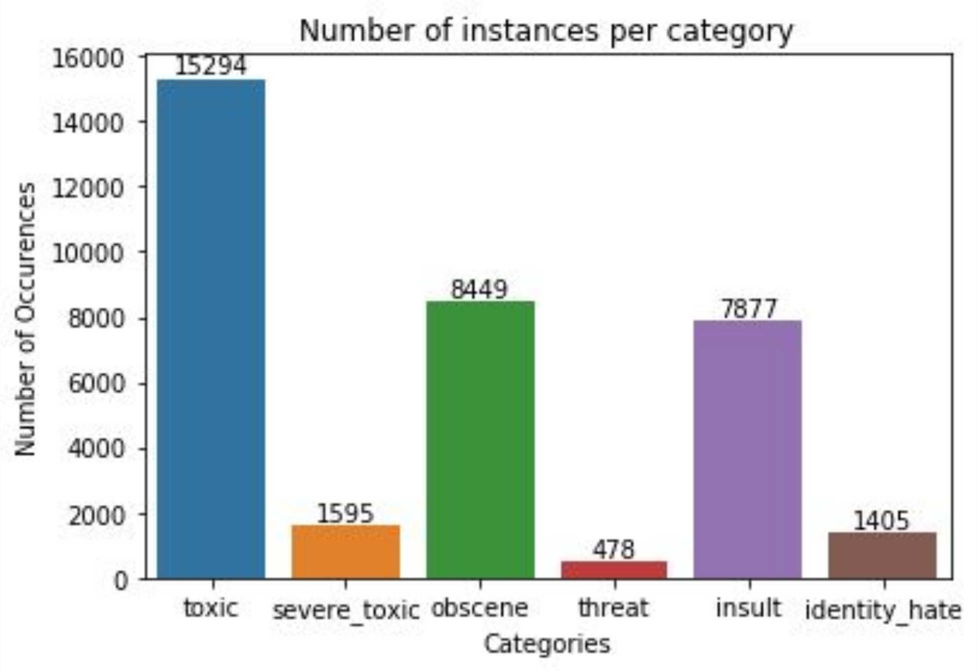


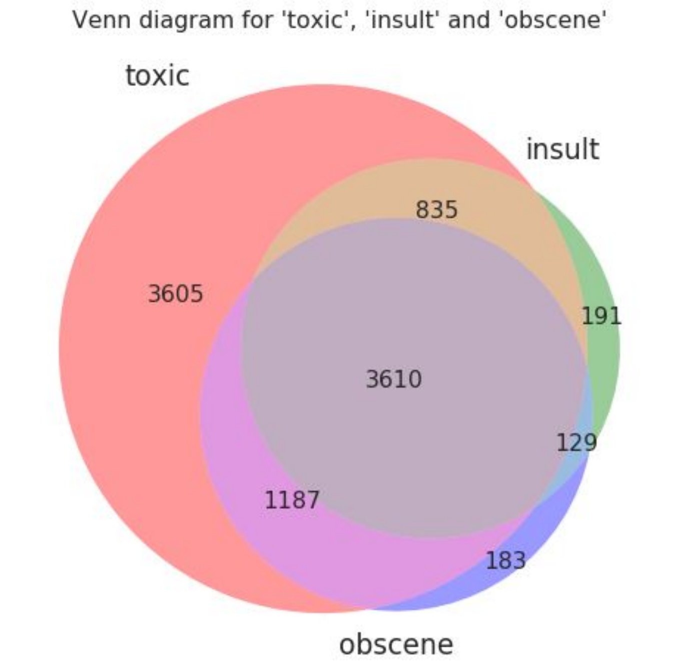
	toxic	severe_toxic	obscene	threat	insult	identity_hate	none	total_length	new_line	new_small_space	new_medium_space	new_big_space	uppercase_words	question_mark	exclamation_mark
toxic	1	0.31	0.68	0.16	0.65	0.26	-0.97	-0.05	-0.0091	-0.015	-0.0016	-0.0051	0.099	0.031	0.038
severe_toxic	0.31	1	0.4	0.13	0.38	0.19	-0.3	0.016	0.011	-0.00091	0.01	-0.0016	0.16	-0.003	0.057
obscene	0.68	0.4	1	0.15	0.74	0.29	-0.7	-0.038	-0.0021	-0.0075	0.0012	-0.0037	0.094	0.0052	0.019
threat	0.16	0.13	0.15	1	0.16	0.12	-0.17	-0.0047	0.00053	0.0047	-0.0014	0.00089	0.041	-0.005	0.047
insult	0.65	0.38	0.74	0.16	1	0.33	-0.68	-0.044	-0.0098	-0.0098	0.0016	-0.0036	0.084	0.0044	0.02
identity_hate	0.26	0.19	0.29	0.12	0.33	1	-0.27	-0.0079	0.00096	0.00099	0.012	-0.0014	0.069	0.002	0.0068
none	-0.97	-0.3	-0.7	-0.17	-0.68	-0.27	1	0.047	0.0073	0.012	0.0021	0.0053	-0.098	-0.032	-0.037
total_length	-0.05	0.016	-0.038	-0.0047	-0.044	-0.0079	0.047	1	0.61	0.58	0.043	0.016	0.37	0.22	0.091
new_line	-0.0091	0.011	-0.0021	0.00053	-0.0098	0.00096	0.0073	0.61	1	0.81	0.08	0.04	0.22	0.23	0.03
new_small_space	-0.015	-0.00091	-0.0075	0.0047	-0.0098	0.00099	0.012	0.58	0.81	1	0.073	0.046	0.16	0.23	0.0085
new_medium_space	-0.0016	0.01	0.0012	-0.0014	0.0016	0.012	0.0021	0.043	0.08	0.073	1	0.63	0.012	0.0078	0.0034
new_big_space	-0.0051	-0.0016	-0.0037	0.00089	-0.0036	-0.0014	0.0053	0.016	0.04	0.046	0.63	1	0.0033	0.0012	-3.7e-05
uppercase_words	0.099	0.16	0.094	0.041	0.084	0.069	-0.098	0.37	0.22	0.16	0.012	0.0033	1	0.083	0.13
question_mark	0.031	-0.003	0.0052	-0.005	0.0044	0.002	-0.032	0.22	0.23	0.23	0.0078	0.0012	0.083	1	0.081
exclamation_mark	0.038	0.057	0.019	0.047	0.02	0.0068	-0.037	0.091	0.03	0.0085	0.0034	-3.7e-05	0.13	0.081	1

Number of classes each comment belongs to



Dataset: EDA and Insights





Dataset: EDA and Insights

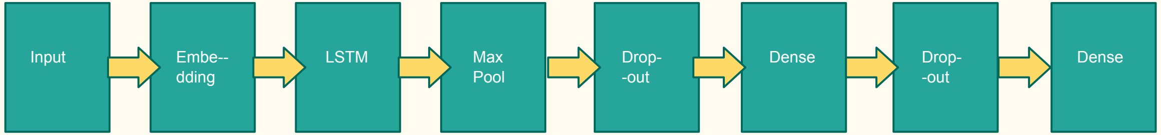
- Choice of **Embedding** essential
- No CNNs !
- Tendency for comments with more sentences to be toxic.

Approach I: Logistic Regression

- Feature used: **TF-IDF**
- Used Python’s in-built Scikit Library Functions
- **Kaggle Score: 0.970**

Approach II: Bi-LSTM, Random Embedding

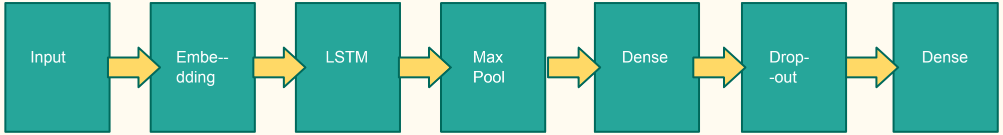
- Training Accuracy: **97.01%**
- Validation(Testing) Accuracy: **96.89%**
- **Random Embedding Layer(128-d)**



[Inspiration: Wei-Yeoung Seow(Top 100 Kaggler). **Change in Embedding layer.]**

Approach III: Bi-LSTM, GloVe Embedding

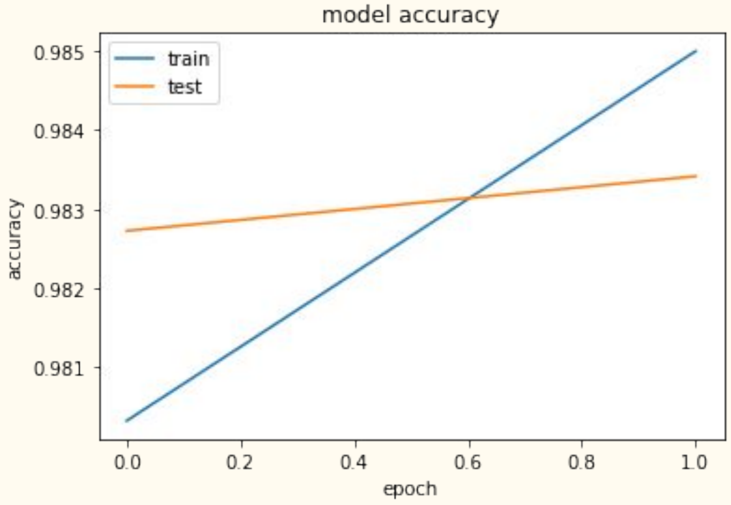
- Training Accuracy: **98.50%**
- Validation(Testing) Accuracy: **98.35%**
- **GloVe 600b20D Embedding**, Gigaword+Wikipedia Embedding

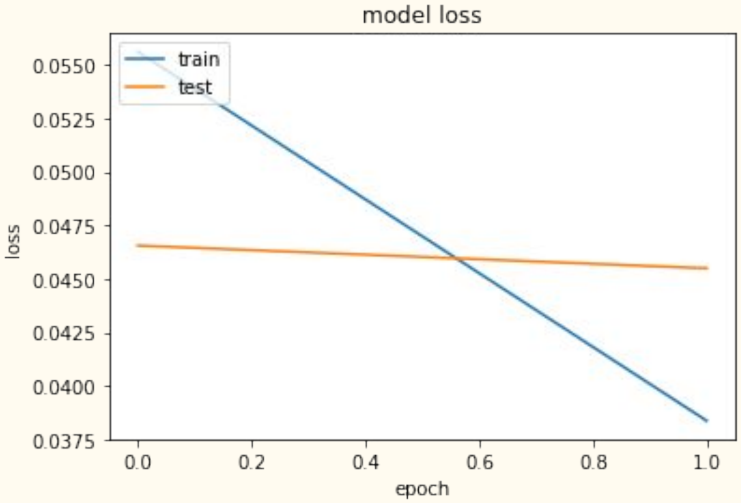


Inspiration: **Jeremy Howard, FastAI(USF)[Change in Embedding and Dropout to improve performance]**

Approach III: Bi-LSTM, GloVe Embedding

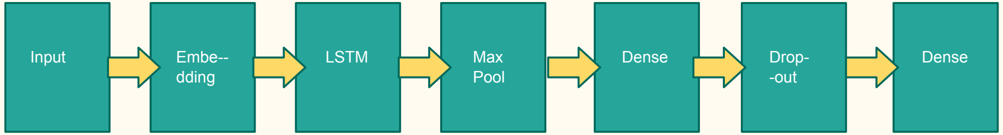
Problem at hand: Overfitting by the **2nd Epoch**





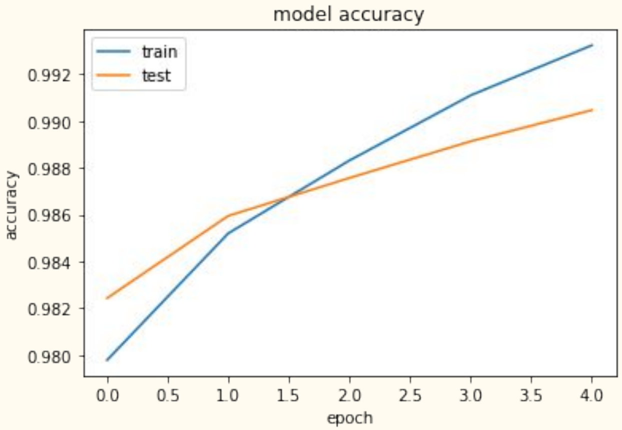
Approach IV: Bi-LSTM, GloVe, Data Augmentation

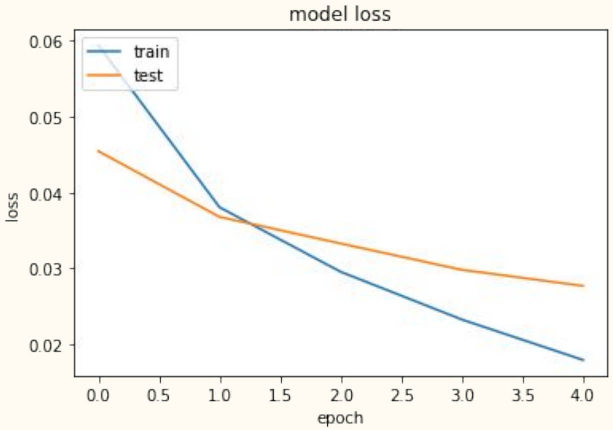
- Training Accuracy: **99.11%**
- Testing Accuracy: **98.91%**
- TextBlob's Machine translation tool.



***Independent Development**

Approach IV: Bi-LSTM, GloVe, Data Augmentation





Approach V: Hierarchical Attention Network

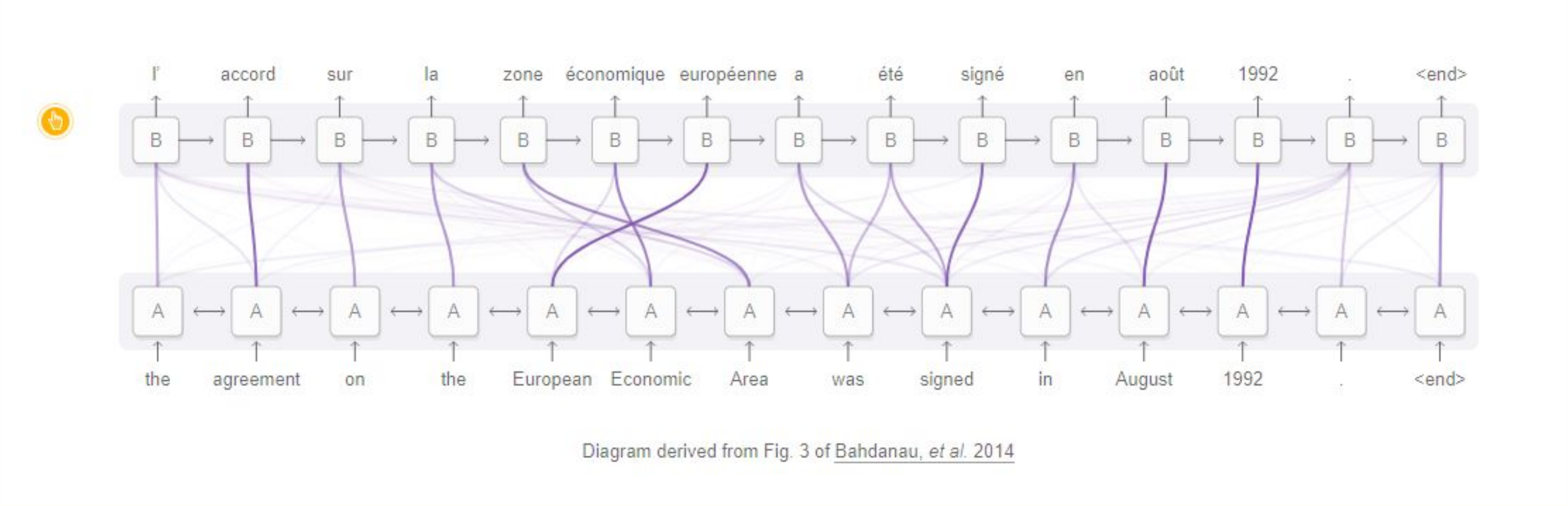
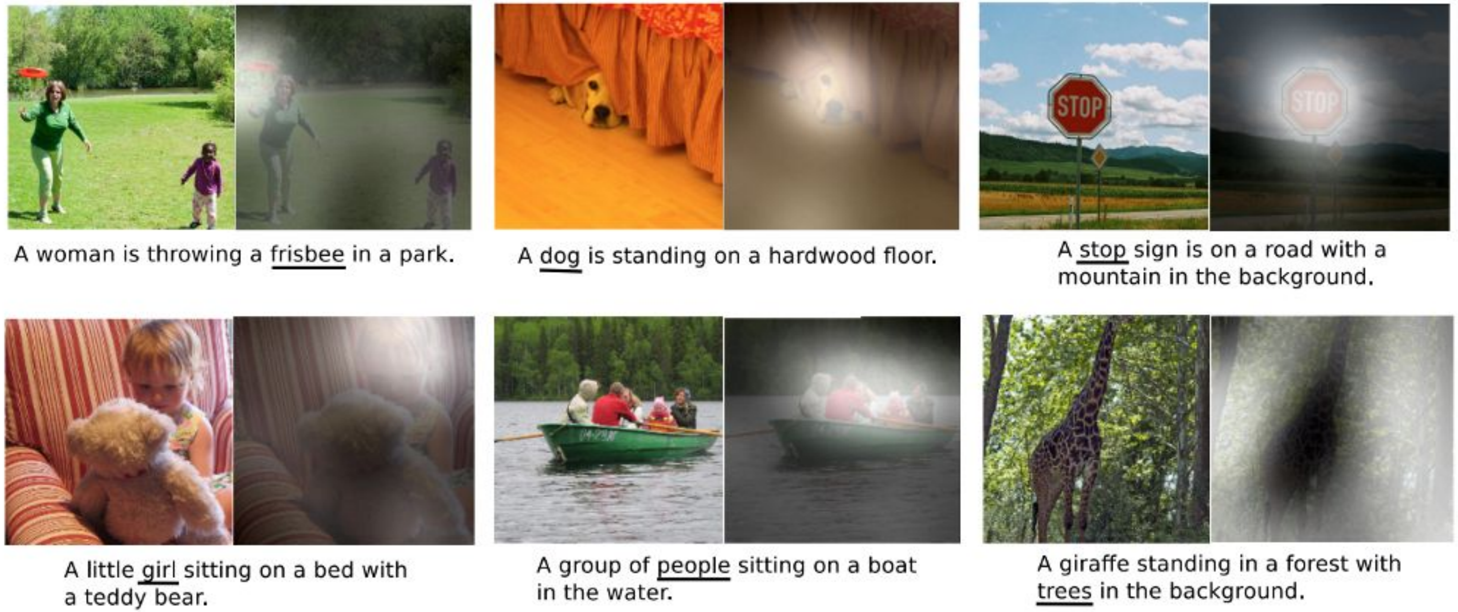


Diagram derived from Fig. 3 of Bahdanau, *et al.* 2014

Source: https://distill.pub/2016/augmented-rnns/#attentional-interfaces

Approach V: Hierarchical Attention Network



A woman is throwing a frisbee in a park.

A dog is standing on a hardwood floor.

A stop sign is on a road with a mountain in the background.

A little girl sitting on a bed with a teddy bear.

A group of people sitting on a boat in the water.

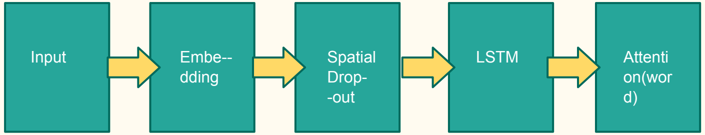
A giraffe standing in a forest with trees in the background.

Figure 6: Attending to objects in an image during caption generation. The white regions indicate where the attention mechanism focused on during the generation of the underlined word. From Xu, Kelvin, et al. "Show, attend and tell: Neural image caption generation with visual attention." International Conference on Machine Learning. 2015.

Source: https://medium.com/datalogue/attention-in-keras-1892773a4f22

Approach V: Hierarchical Attention Network

Word-level
Attention Model



Input

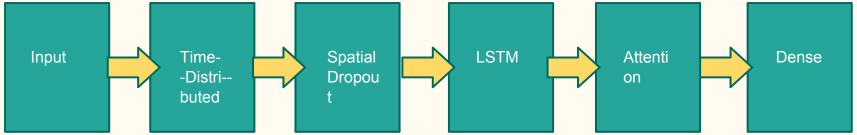
Embe--
dding

Spatial
Drop-
-out

LSTM

Attenti
on(wor
d)

Sentence-level
Attention Model



Input

Time-
-Distri--
buted

Spatial
Dropou
t

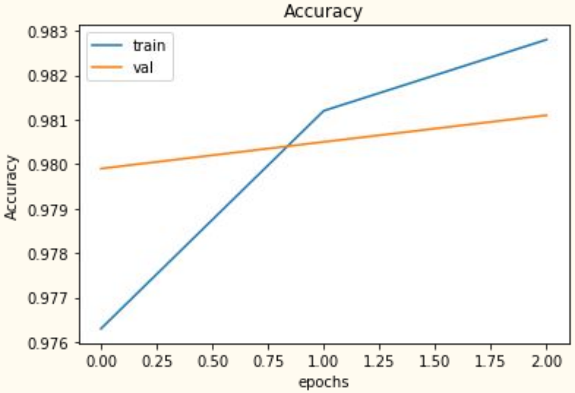
LSTM

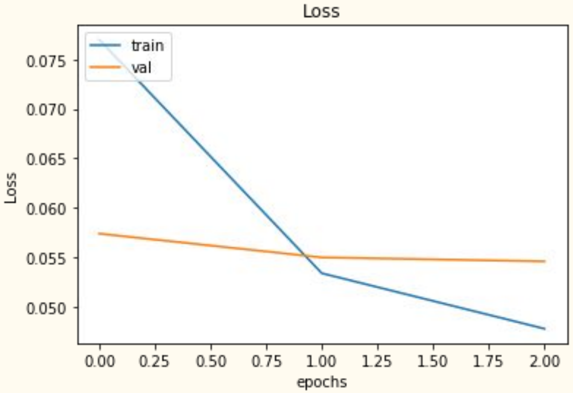
Attenti
on

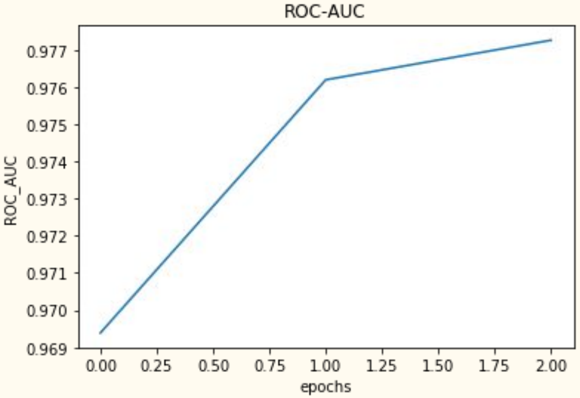
Dense

Approach V: Hierarchical Attention Network

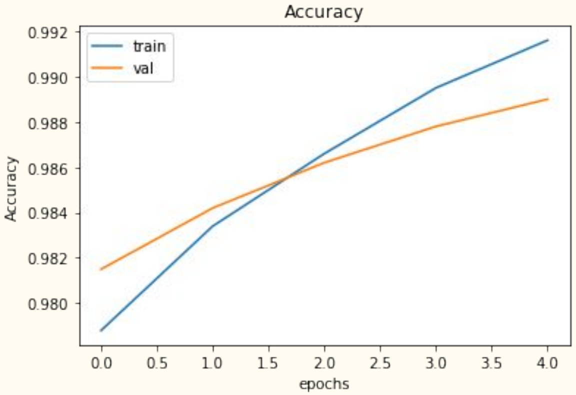
Hello Again, Overfitting!

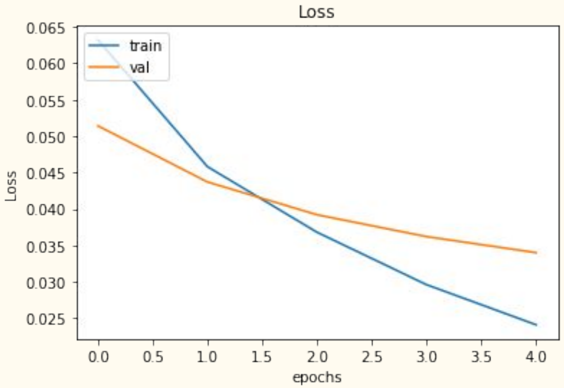


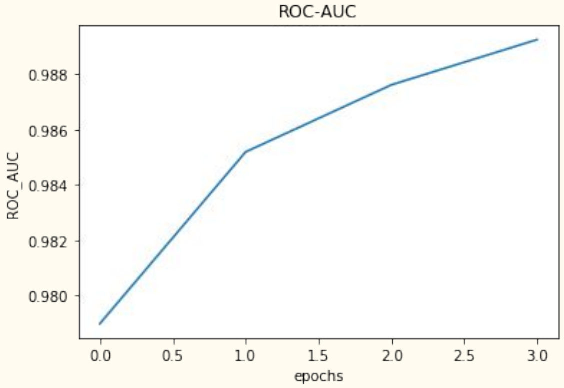




Approach V: Data Augmentation







Results

Method	Tra_Loss	Val_Loss	Tra_Accuracy	Val_Accuracy
Bi-LSTM+Random Embedding	0.0448	0.0489	97.01%	96.89%
Bi-LSTM+GloVe Embedding	0.0384	0.0455	98.50%	98.34%
Bi-LSTM+GloVe+Data Augmentation	0.0232	0.0298	99.11%	98.91%
Hierarchical Attention Network(HAN) + GloVe	0.0478	0.0546	98.28%	98.11%
HAN+Data Augmentation	0.0241	0.0340	99.16%	98.90%

Acknowledgments

- Prof. Yu Hen Hu
- Kaggle
- Euler
- Jigsaw/Conversation AI Toxic Comment Challenge Discussion Forum
- Blogposts: Colah’s Blog/Medium/Jeremy Howard/Andrej Karpathy