DAT 550 – Mini project

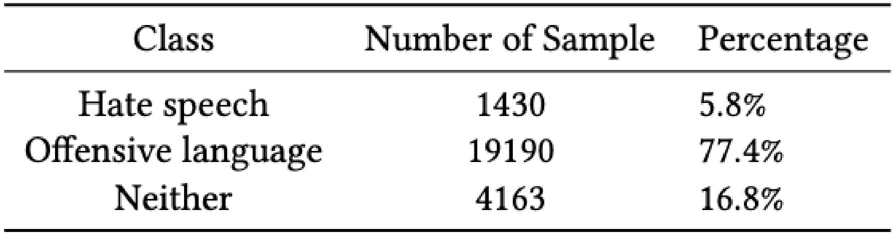
Hate Speech Detection

Group Members :

Javeria Habib
Fadwa Maatug
Khoa Le Nguyen

Dataset

- Dataset used is from Cornell University.
- consists of 24k tweets labeled by the members of CrowdFlower.
- Labelled as three classes
  - 0 – Hate Speech
  - 1 – Offensive Language
  - 2 – Neither



Class	Number of Sample	Percentage
Hate speech	1430	5.8%
Offensive language	19190	77.4%
Neither	4163	16.8%

[1] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. In Proceedings of the 11th International AAAI Conference on Weblogs and Social Media (ICWSM '17).
[2] Figure Eight. 2019. CrowdFlower. Retrieved Apr 24, 2019 from https://www.figure-eight.com/

Preprocessing

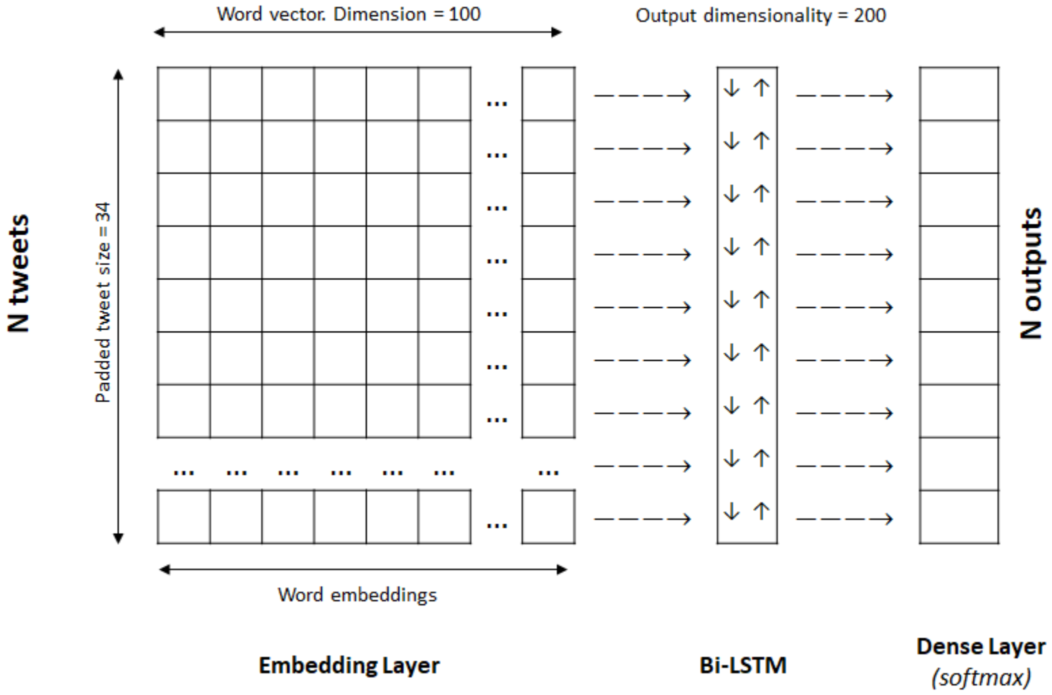
- Convert the text to lower case
- Remove all tweet mentions (any words starting with "@")
- Replace the contraction words with their proper forms.
- Remove all the remaining symbols.
- Replace the words "amp“ and "rt“
- (Optional) Remove all stop words.

Embedding

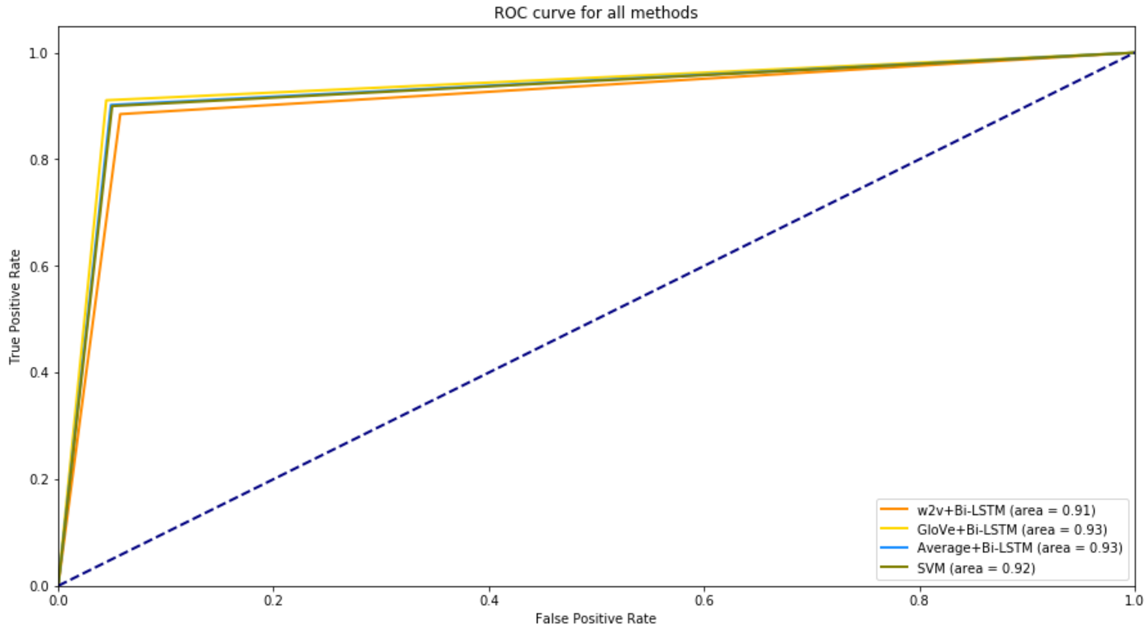
- Word2Vec
- Global Vector (GloVe)
- Combining Word2Vec and GloVe

$$Average\_vector = \frac{GloVe\_Vector + Word2Vec\_Vector}{2}$$

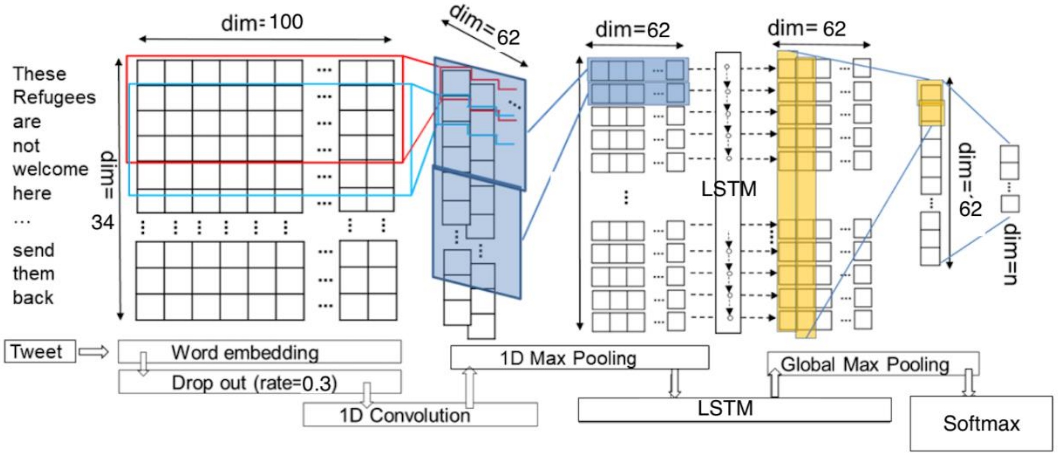
Bidirectional-LSTM



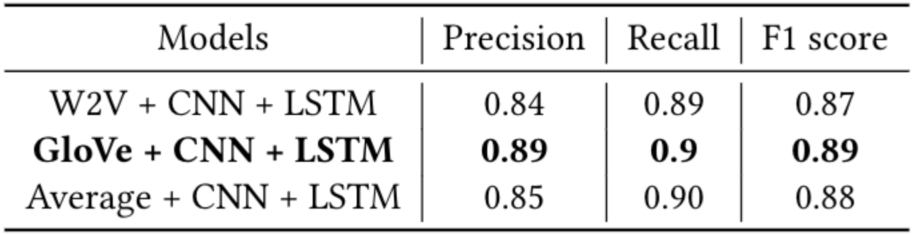
Bi-LSTM comparison



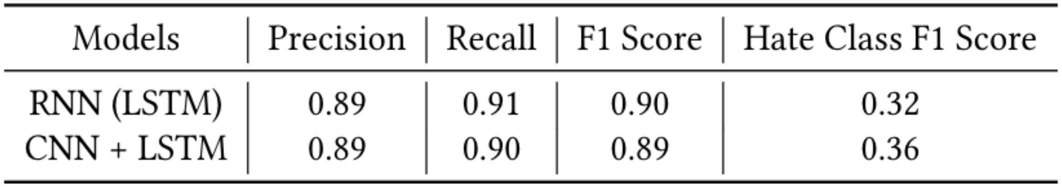
CNN + LSTM


dim=100
dim=62
dim=62
dim= 62
These
Refugees
are
not
welcome
here
…
send
them
back
dim=
34
LSTM
dim=62
dim=n
Tweet
Word embedding
Drop out (rate=0.3)
1D Convolution
1D Max Pooling
LSTM
Global Max Pooling
Softmax

CNN comparison

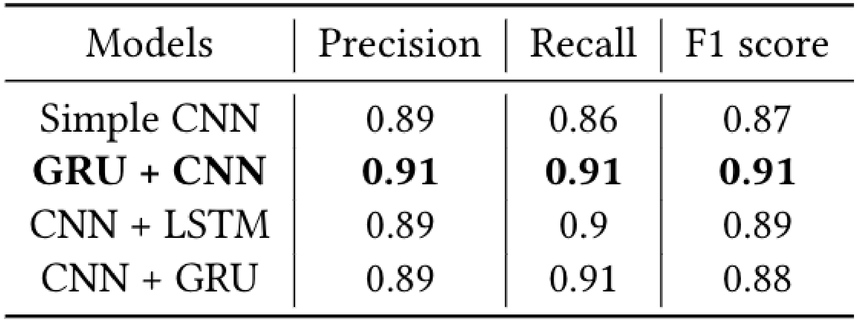


Models	Precision	Recall	F1 score
W2V + CNN + LSTM	0.84	0.89	0.87
GloVe + CNN + LSTM	**0.89**	**0.9**	**0.89**
Average + CNN + LSTM	0.85	0.90	0.88

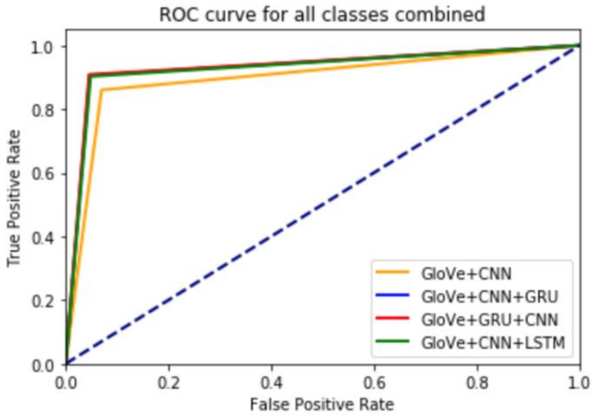


Models	Precision	Recall	F1 Score	Hate Class F1 Score
RNN (LSTM)	0.89	0.91	0.90	0.32
CNN + LSTM	0.89	0.90	0.89	0.36

CNN Models comparison



Models	Precision	Recall	F1 score
Simple CNN	0.89	0.86	0.87
GRU + CNN	**0.91**	**0.91**	**0.91**
CNN + LSTM	0.89	0.9	0.89
CNN + GRU	0.89	0.91	0.88



Data Augmentation

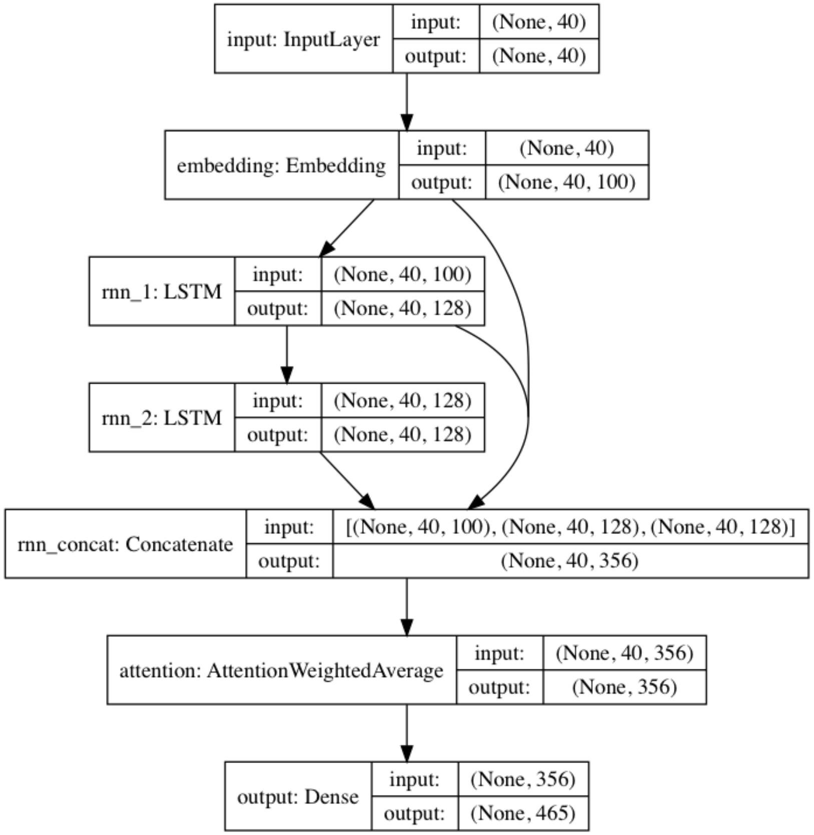
Imbalanced classes problem

- Only 5.8% samples are labeled as "hate speech"
- Not optimized result
- Overfit on majority class

Using Text Generation models to solve this problem

- Recurrent Neural Network (RNN) - textgenrnn
- Generative Adversarial Network (GAN) - LeakGAN

textgenrn



input: InputLayer | input: (None, 40) | output: (None, 40)

embedding: Embedding | input: (None, 40) | output: (None, 40, 100)

rnn_1: LSTM | input: (None, 40, 100) | output: (None, 40, 128)

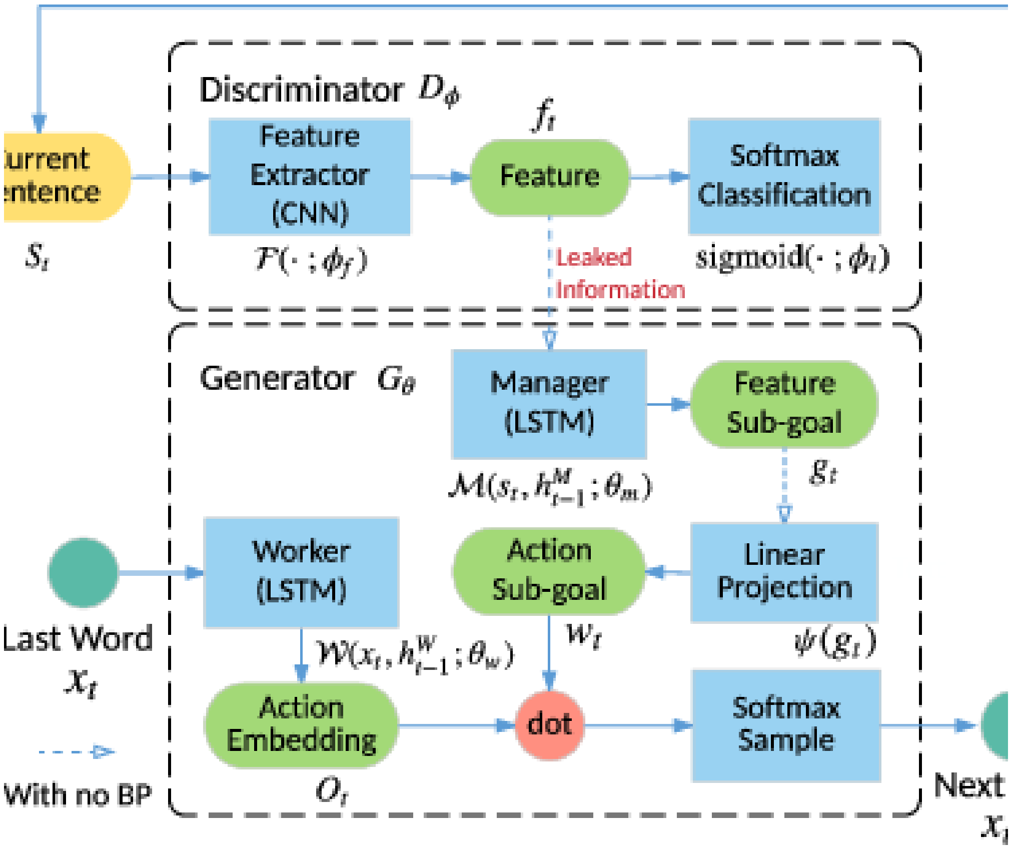
rnn_2: LSTM | input: (None, 40, 128) | output: (None, 40, 128)

rnn_concat: Concatenate | input: [(None, 40, 100), (None, 40, 128), (None, 40, 128)] | output: (None, 40, 356)

attention: AttentionWeightedAverage | input: (None, 40, 356) | output: (None, 356)

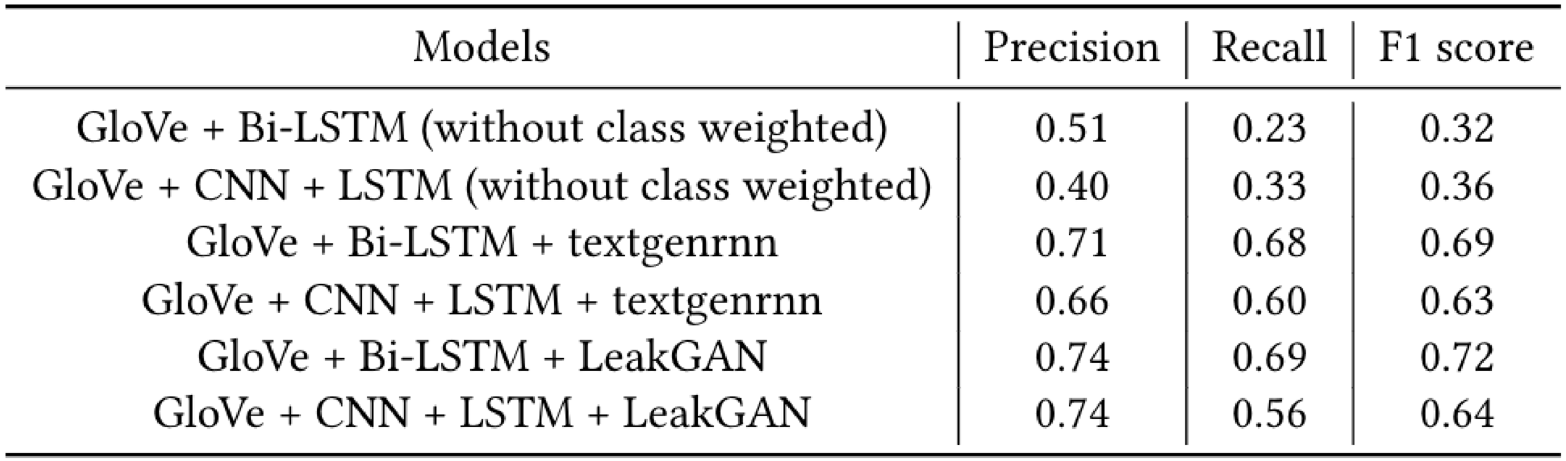
output: Dense | input: (None, 356) | output: (None, 465)

LeakGAN



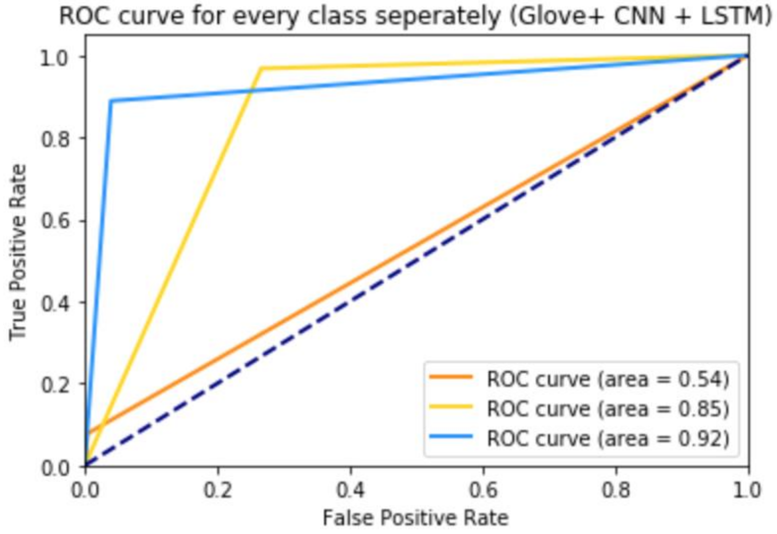
Discriminator D_ϕ
f_t
urrent
ntence
S_t
Feature
Extractor
(CNN)
$\mathcal{F}(\cdot\,;\phi_f)$
Feature
Softmax
Classification
sigmoid$(\cdot\,;\phi_l)$
Leaked
Information
Generator G_θ
Manager
(LSTM)
$\mathcal{M}(s_t, h^M_{t-1};\theta_m)$
Feature
Sub-goal
g_t
Last Word
x_t
Worker
(LSTM)
$\mathcal{W}(x_t, h^W_{t-1};\theta_w)$
Action
Sub-goal
w_t
Linear
Projection
$\psi(g_t)$
Action
Embedding
O_t
dot
Softmax
Sample
Next
x_t
With no BP

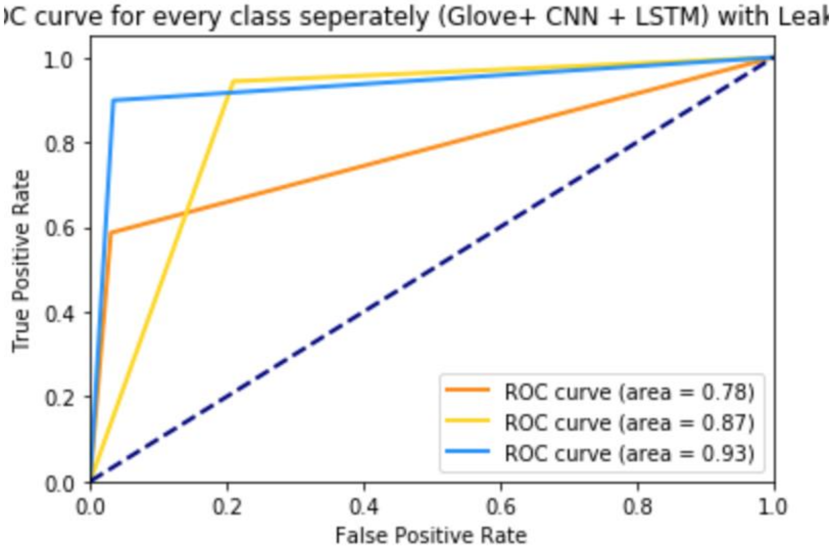
Results



Models	Precision	Recall	F1 score
GloVe + Bi-LSTM (without class weighted)	0.51	0.23	0.32
GloVe + CNN + LSTM (without class weighted)	0.40	0.33	0.36
GloVe + Bi-LSTM + textgenrnn	0.71	0.68	0.69
GloVe + CNN + LSTM + textgenrnn	0.66	0.60	0.63
GloVe + Bi-LSTM + LeakGAN	0.74	0.69	0.72
GloVe + CNN + LSTM + LeakGAN	0.74	0.56	0.64

Results


ROC curve for every class seperately (Glove+ CNN + LSTM)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.54)
ROC curve (area = 0.85)
ROC curve (area = 0.92)


)C curve for every class seperately (Glove+ CNN + LSTM) with Leak
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.78)
ROC curve (area = 0.87)
ROC curve (area = 0.93)