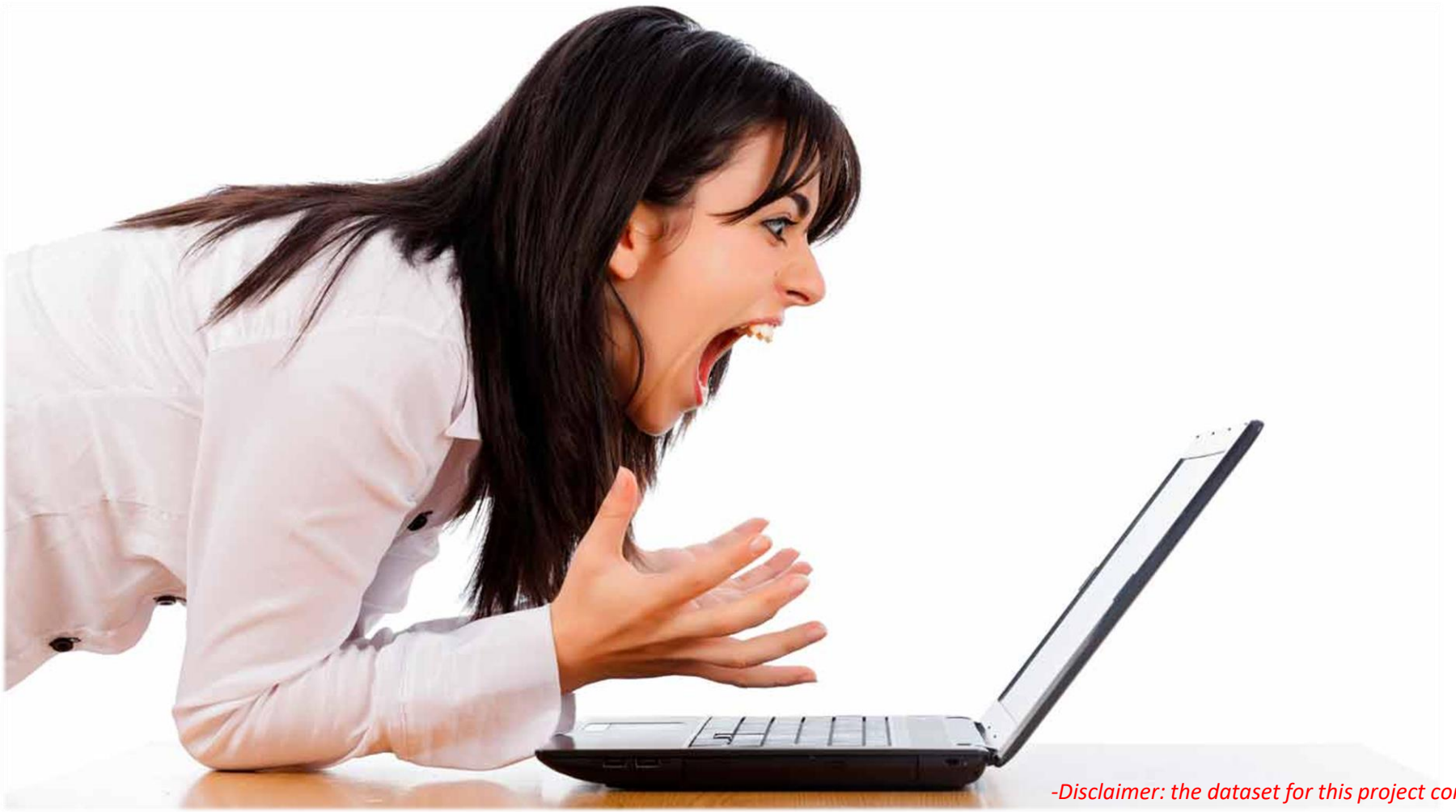
WIKIPEDIA TOXICITY CLASSIFICATION



Mustafa Aytuğ KAYA

-Disclaimer: the dataset for this project contains text that may be considered profane, vulgar, or offensive.

Dataset

- Wikipedia is a multilingual online encyclopedia based on open collaboration
- Some contents are toxic, having:
  - improper language
  - severe profane language
  - reflecting hatred towards a group of people
  - obscenity
  - threat
  - insult

```
1 train[train.iloc[:,2:8].sum(axis='columns')>5].iloc[4:,1:9]
```

	comment_text	toxic	severe_toxic	obscene	threat	insult	identity_hate
13964	I am going to murder ZimZalaBim ST47 for being...	1	1	1	1	1	1
22158	FUCK YOU!!!!!!!!!!!! YOU FUCKING NIGGER BAG OF...	1	1	1	1	1	1
29968	u motherfukkin bitch i want to rape you smelly...	1	1	1	1	1	1
32098	Fuck All Asyriac Nation \n\nQamishli belong to...	1	1	1	1	1	1
33951	GO FUCK YOURSELF BITCH. I HATE YOUR SOULD. M...	1	1	1	1	1	1
38513	AM GOING TO RAPE YOU IN THE ASS YOU FAT BITCH ...	1	1	1	1	1	1
38578	fuck you honkey, why you hatin' on blacks? You...	1	1	1	1	1	1

Dataset

- Multi-Label Dataset
- Imbalance in some labels as high as 1:300!

Data Cleaning / Text Pre-processing

- URL addresses were removed by Python built-in library Regex.
- Then accented characters were normalized into ASCII characters.
- Next, contractions are expanded via a dictionary for text standardization.
- Remaining special characters and symbols were removed with regular expressions code.
- Text was lemmatized.
- Stop words were removed, except for Recurrent Neural Network models.
- Extra white spaces and text was converted to lowercase.
- Spacy was used in this project since as well as Keras text preprocessing.

Baseline Models

- Multinomial Naive Bayes
- Logistic regression
- Support Vector Classification
- Classify each comment for each of the six classes.

```
1  print('Accuracy Score is {}, precision score is {},\n\
2  Recall score is {}, roc_auc score is {}'.format(acc,prc,rec,auc))
```

```
Accuracy Score is 0.963, precision score is 0.154,
Recall score is 0.6523, roc_auc score is 0.8091
```

```
9  print('Accuracy Score is {}, precision score is {},\n\
10 Recall score is {}, roc_auc score is {}'.format(acc,prc,rec,auc))
11
```

```
Accuracy Score is 0.9564, precision score is 0.9159,
Recall score is 0.5975, roc_auc score is 0.7959
```

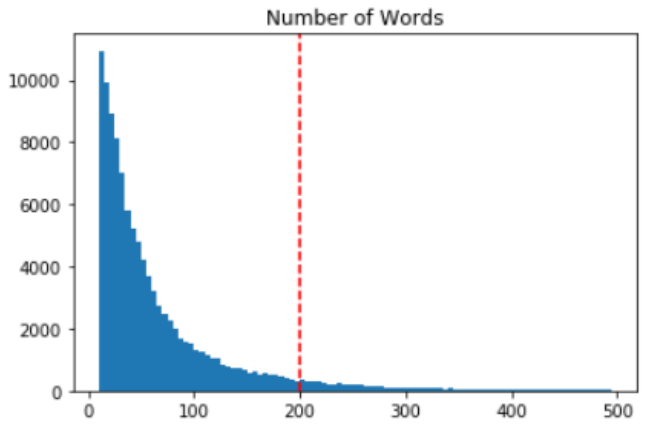
```
11 print('Accuracy Score is {}, precision score is {},\n\
12 Recall score is {}, roc_auc score is {}'.format(acc,prc,rec,auc))
```

```
Accuracy Score is 0.9971, precision score is 0.3333,
Recall score is 0.0204, roc_auc score is 0.5101
```

not too good...

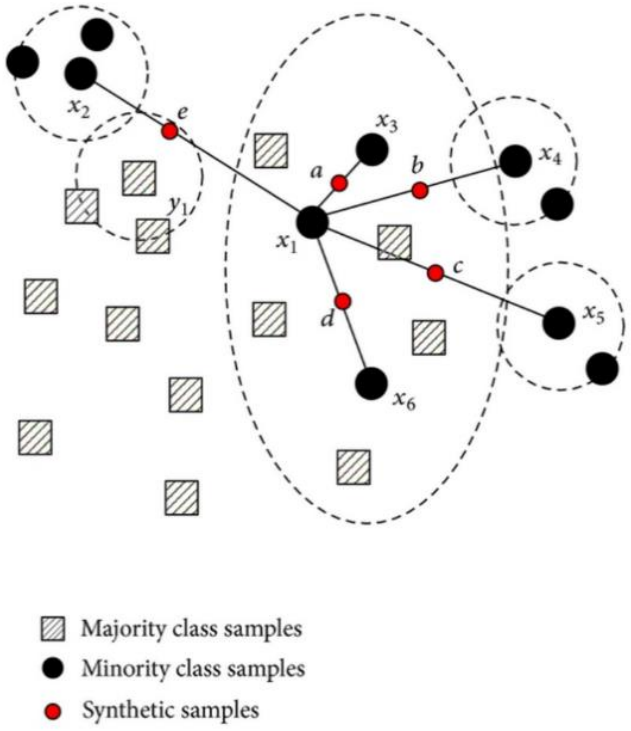
Neural Network Models

- With Regex, Spacy and Keras text preprocessing.
- Only Keras preprocessing
- With Regex, Spacy and Keras text preprocessing, leaving stop words.



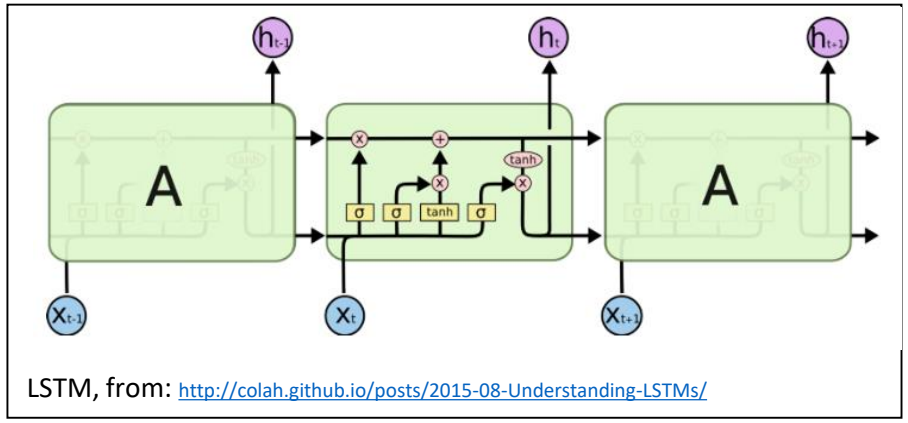
Class Imbalance

- Used ADASYN to oversample padded sequences to compensate for underrepresented classes.
- Used in conjunction with class weights.

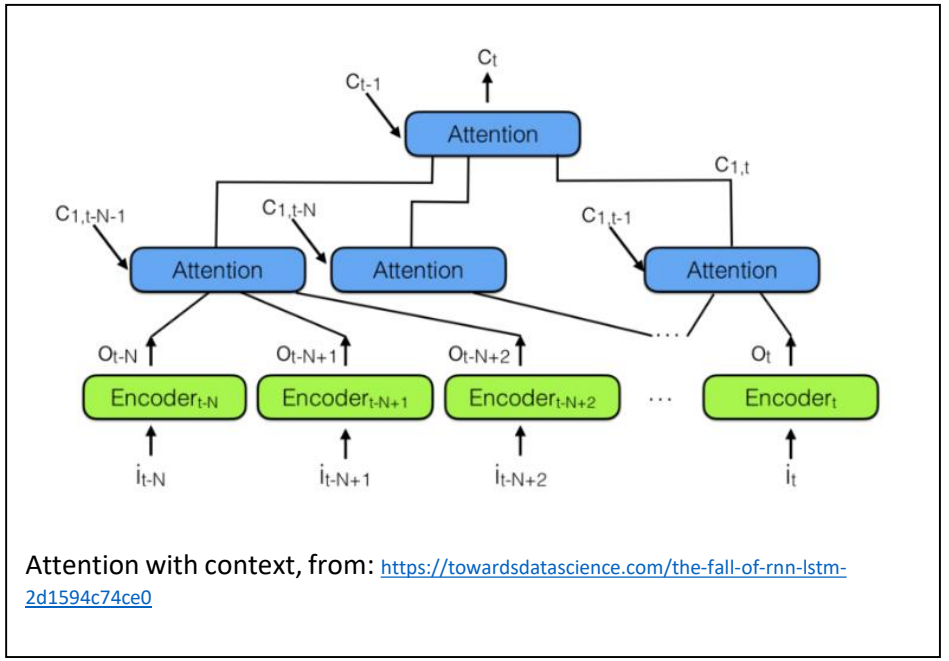


Master Model

- Embeddings (300 D)
- 2xBidirectional CudNNLSTM + Attention with context + Dense Layers



LSTM, from: http://colah.github.io/posts/2015-08-Understanding-LSTMs/



Attention with context, from: https://towardsdatascience.com/the-fall-of-rnn-lstm-2d1594c74ce0

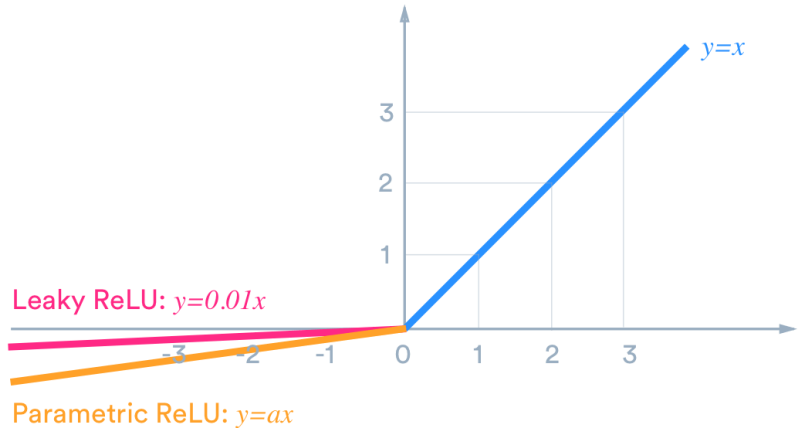
Master Model

- Leaky ReLu was used for activation.

Layer (type)	Output Shape	Param #
input_1 (InputLayer)	(None, 200)	0
embedding_5 (Embedding)	(None, 200, 300)	6000000
bidirectional_9 (Bidirection	(None, 200, 140)	208320
spatial_dropout1d_9 (Spatial	(None, 200, 140)	0
bidirectional_10 (Bidirectio	(None, 200, 140)	118720
spatial_dropout1d_10 (Spatia	(None, 200, 140)	0
attention_with_context_5 (At	(None, 140)	19880
dense_15 (Dense)	(None, 70)	9870
dropout_13 (Dropout)	(None, 70)	0
dense_16 (Dense)	(None, 70)	4970
dropout_14 (Dropout)	(None, 70)	0
dense_17 (Dense)	(None, 70)	4970
dropout_15 (Dropout)	(None, 70)	0
dense_18 (Dense)	(None, 6)	426

Total params: 6,367,156
Trainable params: 6,367,156
Non-trainable params: 0

None



Threat and Identity Hate Model

- Those models were concatenated to augment the master model.

Layer (type)	Output Shape	Param #
input_1 (InputLayer)	(None, 200)	0
embedding_4 (Embedding)	(None, 200, 300)	6000000
bidirectional_6 (Bidirection	(None, 200, 160)	244480
spatial_dropout1d_6 (Spatial	(None, 200, 160)	0
bidirectional_7 (Bidirection	(None, 200, 160)	154880
spatial_dropout1d_7 (Spatial	(None, 200, 160)	0
attention_with_context_4 (At	(None, 160)	25920
dropout_9 (Dropout)	(None, 160)	0
dense_12 (Dense)	(None, 80)	12880
dropout_10 (Dropout)	(None, 80)	0
dense_13 (Dense)	(None, 80)	6480
dropout_11 (Dropout)	(None, 80)	0
dense_14 (Dense)	(None, 1)	81

Total params: 6,444,721
Trainable params: 6,444,721
Non-trainable params: 0

None

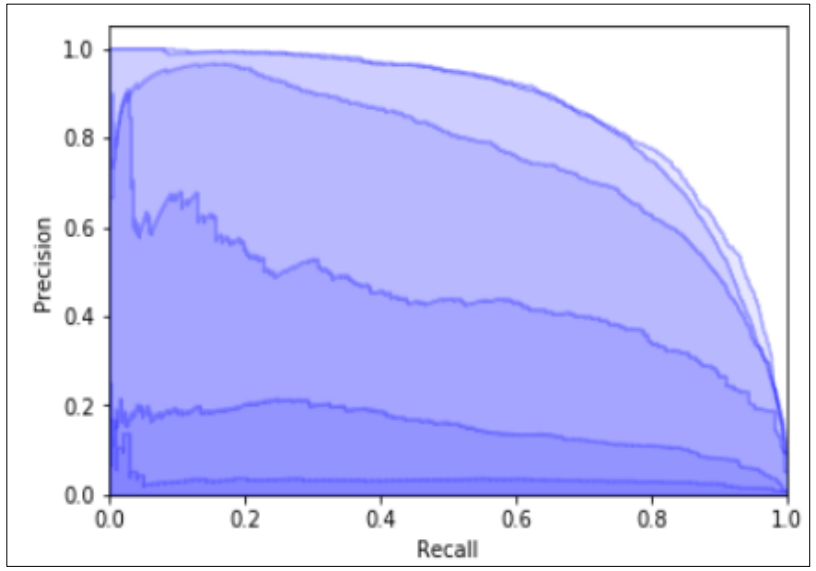
Layer (type)	Output Shape	Param #
input_1 (InputLayer)	(None, 200)	0
embedding_4 (Embedding)	(None, 200, 300)	6000000
bidirectional_6 (Bidirection	(None, 200, 160)	244480
spatial_dropout1d_6 (Spatial	(None, 200, 160)	0
bidirectional_7 (Bidirection	(None, 200, 160)	154880
spatial_dropout1d_7 (Spatial	(None, 200, 160)	0
attention_with_context_4 (At	(None, 160)	25920
dropout_9 (Dropout)	(None, 160)	0
dense_12 (Dense)	(None, 80)	12880
dropout_10 (Dropout)	(None, 80)	0
dense_13 (Dense)	(None, 80)	6480
dropout_11 (Dropout)	(None, 80)	0
dense_14 (Dense)	(None, 1)	81

Total params: 6,444,721
Trainable params: 6,444,721
Non-trainable params: 0

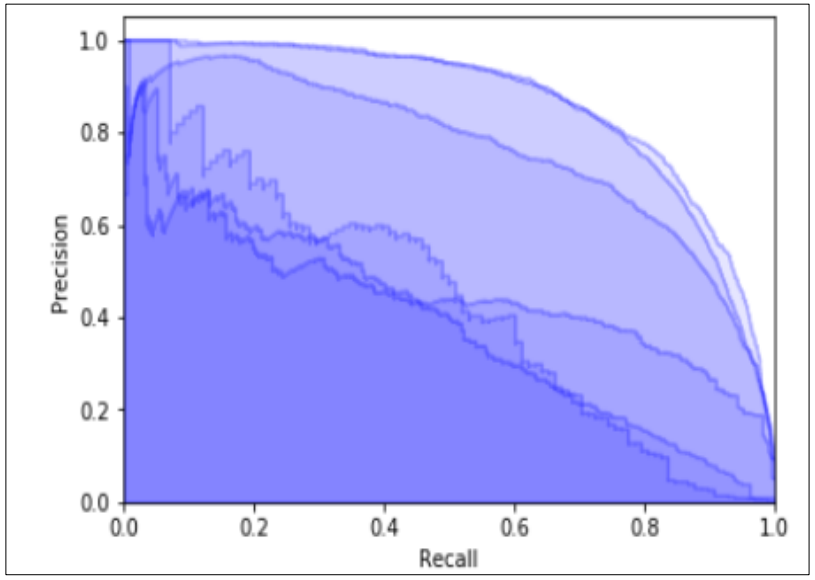
None

Precision Recall Curves

- ADASYN augmented models boosted the master model significantly.

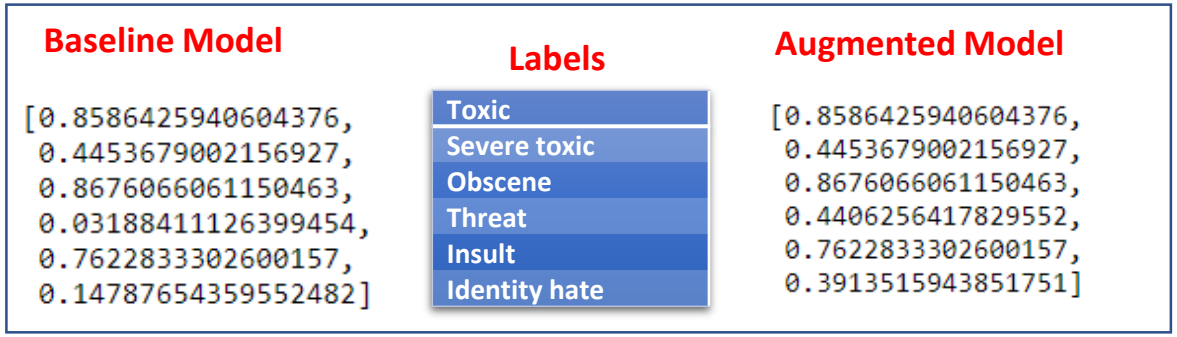


master model



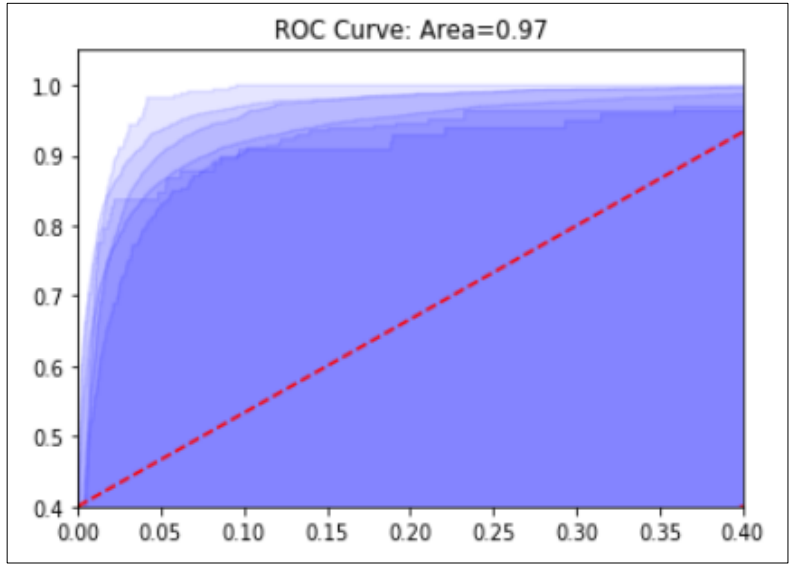
augmented model

Precision – Recall AUC



ROC AUC, Accuracy

- Due to extreme imbalance ROC is very misleading.




```
1 from sklearn.metrics import accuracy_score
2 accs=[accuracy_score(y_val[:,n],np.round(predictions3[:,n])) for n in range(6)]
3 np.mean(accs)
```

```
0.9813668489003308
```

Problem Areas

- Frequent occurrence of mislabeled instances.

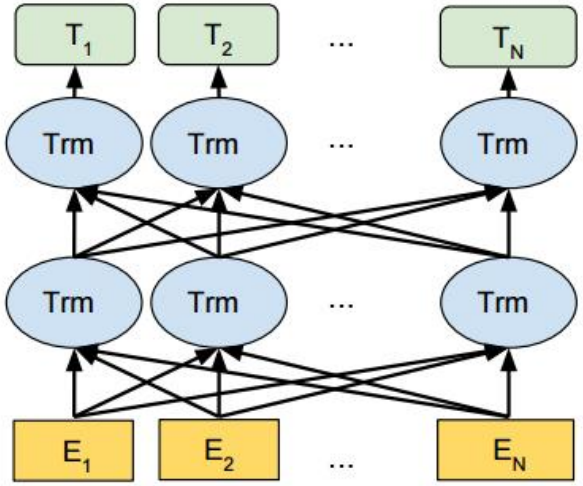
"You Zionist Jewbastard Khazar Turks just love filibusters that draw out this tragedy to no conclusion. That's righ
t, only YOU are allowed a say on the issue. YOU have the right to editorialise anything to YOUR content, media mogul jackass
es! Stay out of London, New York, Washington and Hollywood! Get the fuck out of America and stop dragging us into your stup
id affairs with Muslims! You deserved 9/11 and I hope more of you die from suicide bombings by economically tortured Muslim
s, just keep it in the Middle East. Helen Clark did well to not take your shite! I swear, I'll fucking kill you all if I ev
er go to Israel. I'll take nukes signed by each and every Jew of the Manhattan Project and level you to nothing; in a eulogy
to Theodore Herzl. What irony, to die by the products of your own hands, that had me fear for my life in the fucking Cold Wa
r. Mad scientists and loan sharks, fucking trash with no goddamn decency to Europe and America! Wanderer gypsies, no sense
of love for your own people enough to stick together on your own land. Can't even settle down and do your own thing away fro
m others. Leave us and Rachel Corrie alone! You have no respect for the dead! To you, she was just another Goy puppet! Yo
u will pay and I hope to personally see you die. Perhaps a Tay-Sachs bio-weapon to plague and infest the lot of you until de
ath do we part. Take your Michael Medveds, Adam Sandlers and Bob Dylans and shove them up your arses! No more Disraelis and
Kerrys or Rothschilds and Greenspans will ruin our lives! You bring disrepute to White males, by rewriting history; fucking
up churches and schools in culture war. You killed McKinley and the Romanovs. Fuck your Sigmund Freud, Ron Jeremy and Ruth
Westheimer pervert terrorists and kill yourselves NOW! You fucking Howard Stern/Jerry Seinfeld/Eugene Levy garbage are no be
tter than any other Semite like the Muslims themselves! You anti-Semitic hypocrites have no feature but greed and barbarism!
I read the massacre at Clifford's Tower in York Castle and hope for another! Your ritual child abuse still mutilates baby ge
nitals, just like labia removal. You are savages without civilisation and bloodsucking leeches holding onto hosts as all vir
uses do; so you are fake friends! No more Bugsy Siegel Hollowcau$e Industry and kosher racketeering!\n\nGENOCIDIST DAVID SLE
W GOLIATH OF PALESTINE! REJECTED ONES, YOU HAVEN'T MONOPOLY ON SUFFERING!",

```
1 train_transformed.iloc[47012:47013]
```

	Unnamed: 0	id	comment_text	toxic	severe_toxic	obscene	threat	insult	identity_hate
47012	47012	7d9a493d304f1b70	You Zionist Jewbastard Khazar Turks just love ...	1	0	0	0	0	0

Future Work

- Word embeddings to build a synonyms to alleviate class imbalance.
- Bidirectional Encoder Representations from Transformers - BERT




Questions?