Toxic Comment Classification

Team: WHY XL

Feng Wang, Lipin Yuan, Xing Jun , Yezhu Li, Zhonghe Han

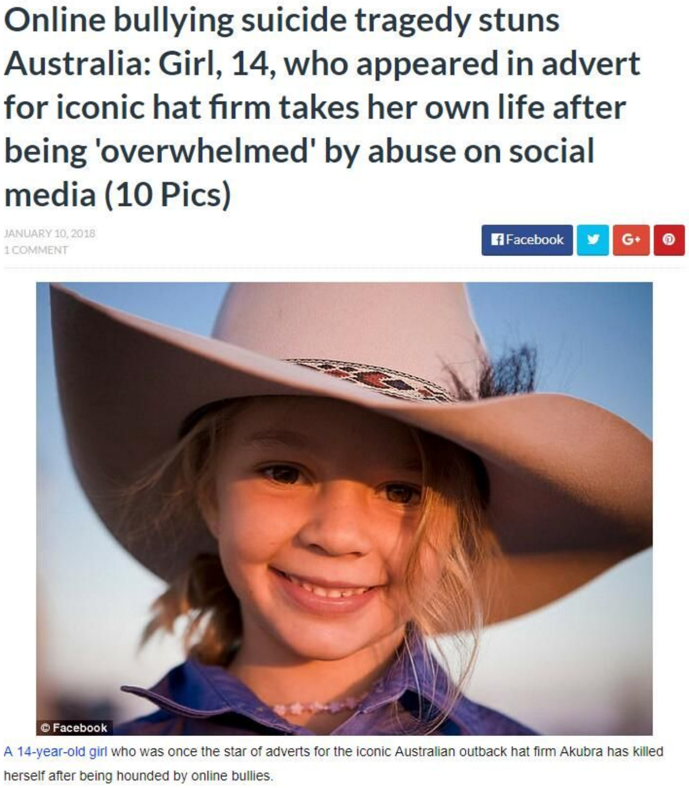
Agenda

- Background Introduction
- Data Preparation and Data Cleaning
- Data Exploration and Visualization
- Text Analytics: ML algorithm
- Conclusion & Shiny Demo

Why Detecting Toxic Comment is Important?

More than 43% of teens has experienced cyber bullying

Bully @bully
UR SO UGLY
6m
Bully
You're worthless
7s
Bully
2 seconds ago
NO ONE CARES!!
Bully said:
You're so PATHETIC
You're a LOSER.
You SUCK
fakebook
WORDS SCAR, RUMORS DESTROY, BULLIES KILL.
Bully
Ew. Don't talk to me.
bully
Can't believe you even show up to school. Nobody wants you there.
Bully
I HATE You
Nobody likes you.
Aneta Krajewska
Bully said:
Just give up now
1
Bully
You are STUPID
Bully
Ew, such an attention seeker.
Like Comments Share
Bully
You are so FAKE
You Tube
Bully
Stop trying to be cool, is not working.
1 second ago
bully
WEIRDO
3s



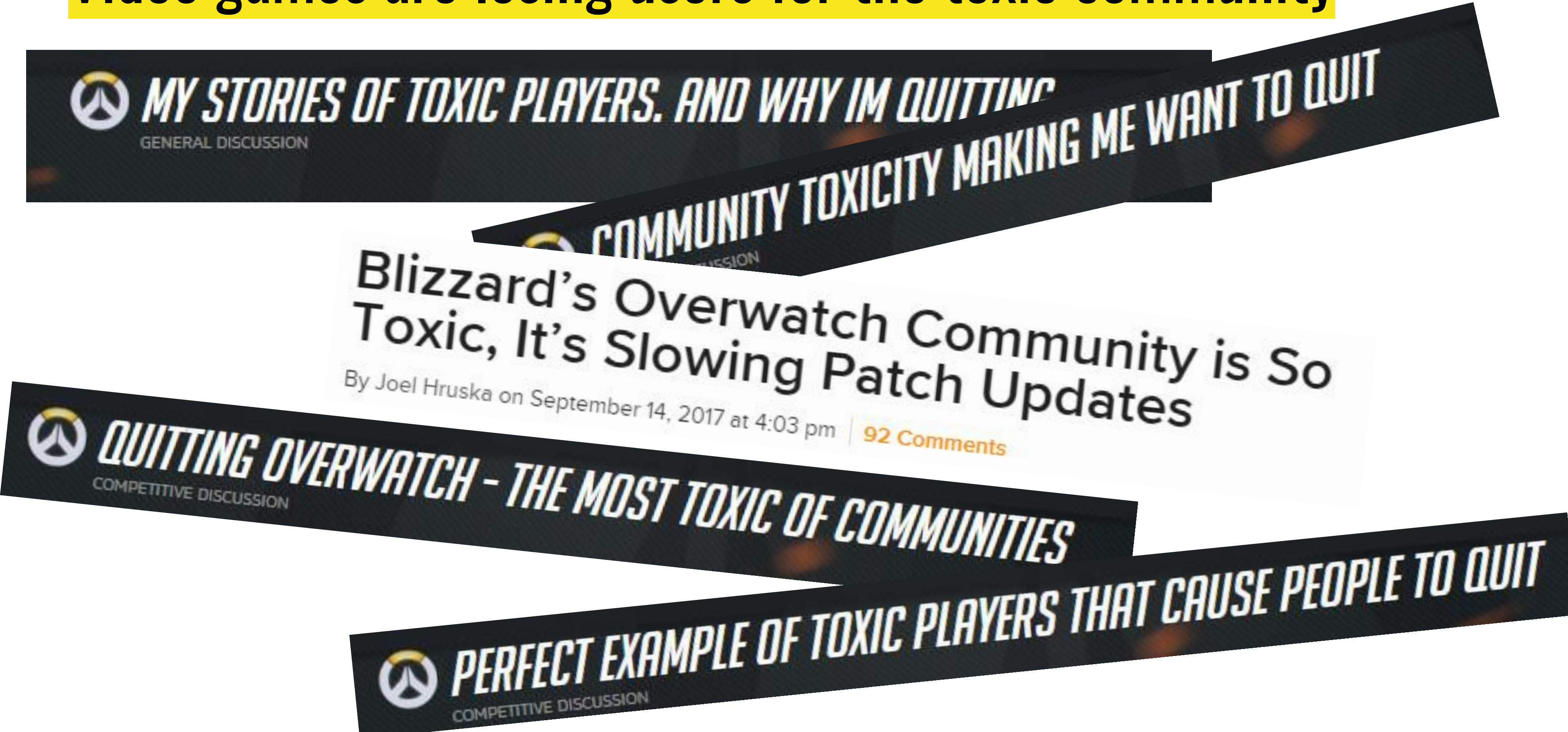
Online bullying suicide tragedy stuns Australia: Girl, 14, who appeared in advert for iconic hat firm takes her own life after being 'overwhelmed' by abuse on social media (10 Pics)

JANUARY 10, 2018
1 COMMENT

© Facebook

A 14-year-old girl who was once the star of adverts for the iconic Australian outback hat firm Akubra has killed herself after being hounded by online bullies.

Video games are losing users for the toxic community



MY STORIES OF TOXIC PLAYERS. AND WHY IM QUITTING

GENERAL DISCUSSION

COMMUNITY TOXICITY MAKING ME WANT TO QUIT

USSION

Blizzard's Overwatch Community is So Toxic, It's Slowing Patch Updates

By Joel Hruska on September 14, 2017 at 4:03 pm | 92 Comments

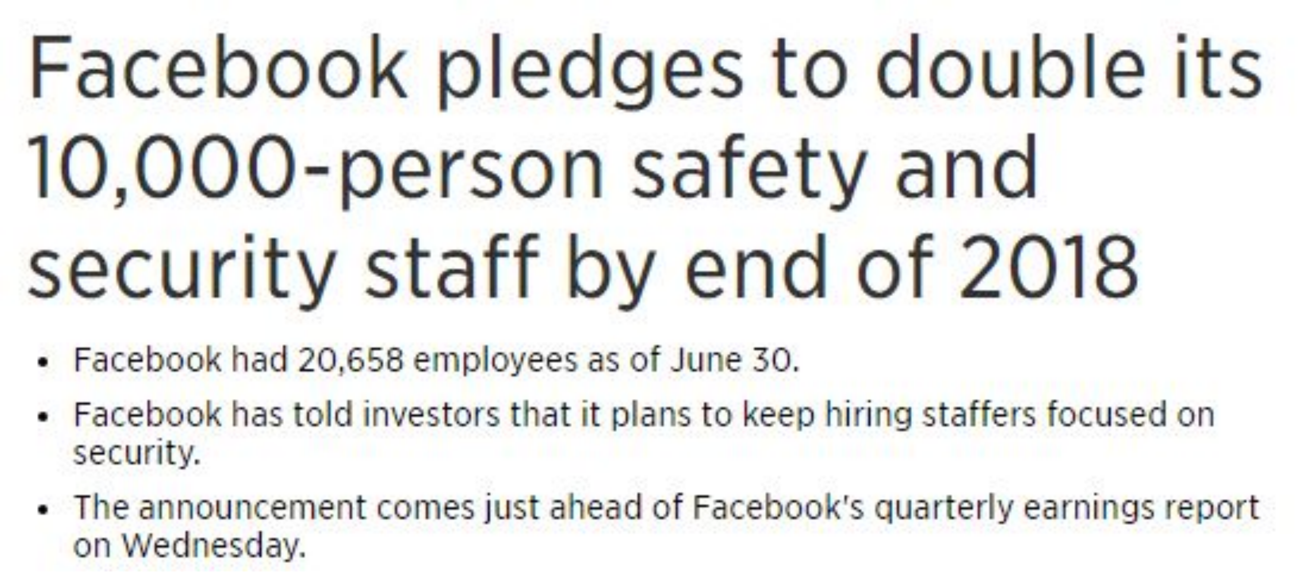
QUITTING OVERWATCH - THE MOST TOXIC OF COMMUNITIES

COMPETITIVE DISCUSSION

PERFECT EXAMPLE OF TOXIC PLAYERS THAT CAUSE PEOPLE TO QUIT

COMPETITIVE DISCUSSION

Manually check or report toxic users are inefficient and costful



Facebook pledges to double its 10,000-person safety and security staff by end of 2018

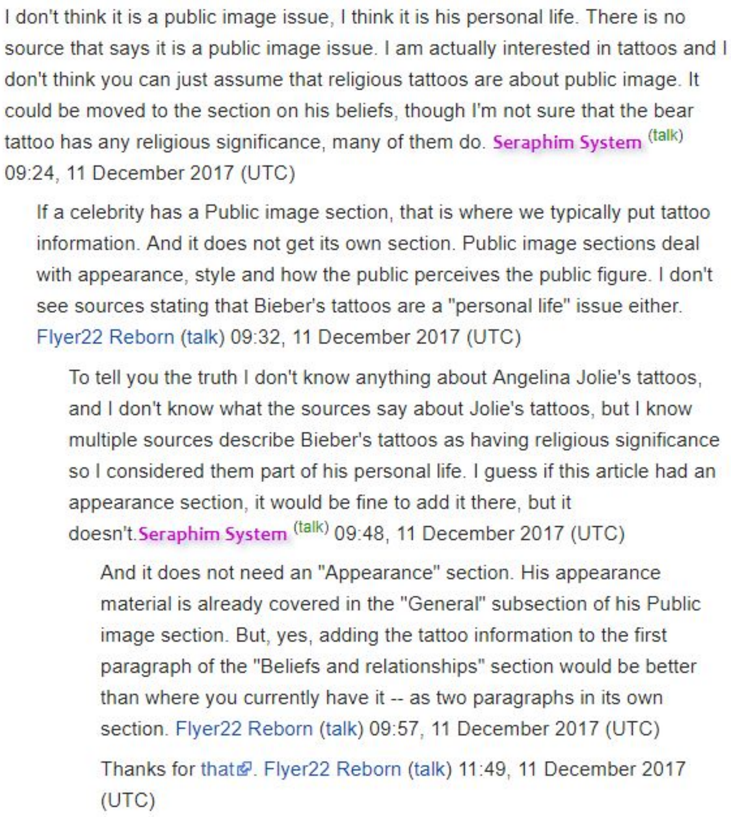
- Facebook had 20,658 employees as of June 30.
- Facebook has told investors that it plans to keep hiring staffers focused on security.
- The announcement comes just ahead of Facebook's quarterly earnings report on Wednesday.

Data Source: Wikipedia Talk Page

Talk:Justin Bieber

From Wikipedia, the free encyclopedia

Number of Comments: 159,571
Number of Labels: 6
% of Clean Comments:89.8%



I don't think it is a public image issue, I think it is his personal life. There is no source that says it is a public image issue. I am actually interested in tattoos and I don't think you can just assume that religious tattoos are about public image. It could be moved to the section on his beliefs, though I'm not sure that the bear tattoo has any religious significance, many of them do. Seraphim System (talk) 09:24, 11 December 2017 (UTC)

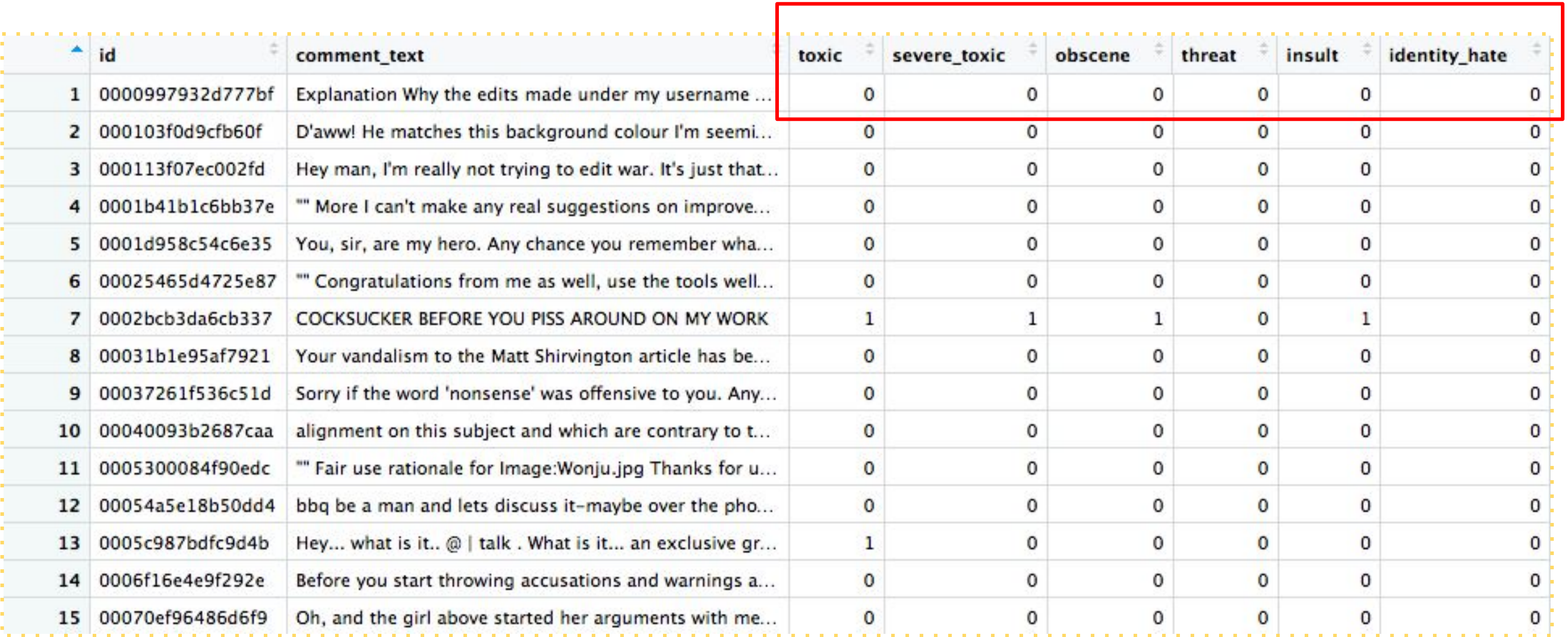
> If a celebrity has a Public image section, that is where we typically put tattoo information. And it does not get its own section. Public image sections deal with appearance, style and how the public perceives the public figure. I don't see sources stating that Bieber's tattoos are a "personal life" issue either. Flyer22 Reborn (talk) 09:32, 11 December 2017 (UTC)
>
>> To tell you the truth I don't know anything about Angelina Jolie's tattoos, and I don't know what the sources say about Jolie's tattoos, but I know multiple sources describe Bieber's tattoos as having religious significance so I considered them part of his personal life. I guess if this article had an appearance section, it would be fine to add it there, but it doesn't.Seraphim System (talk) 09:48, 11 December 2017 (UTC)
>>
>>> And it does not need an "Appearance" section. His appearance material is already covered in the "General" subsection of his Public image section. But, yes, adding the tattoo information to the first paragraph of the "Beliefs and relationships" section would be better than where you currently have it -- as two paragraphs in its own section. Flyer22 Reborn (talk) 09:57, 11 December 2017 (UTC)
>>>
>>> Thanks for that. Flyer22 Reborn (talk) 11:49, 11 December 2017 (UTC)

Data Preparation

A large number of Wikipedia - Talk Page comments which have been labeled by human raters for toxic behavior. The types of toxicity are:

- **toxic**
- **severe_toxic**
- **obscene**
- **threat**
- **insult**
- **identity_hate**

Dataset Screenshot



	id	comment_text	toxic	severe_toxic	obscene	threat	insult	identity_hate
1	0000997932d777bf	Explanation Why the edits made under my username ...	0	0	0	0	0	0
2	000103f0d9cfb60f	D'aww! He matches this background colour I'm seemi...	0	0	0	0	0	0
3	000113f07ec002fd	Hey man, I'm really not trying to edit war. It's just that...	0	0	0	0	0	0
4	0001b41b1c6bb37e	"" More I can't make any real suggestions on improve...	0	0	0	0	0	0
5	0001d958c54c6e35	You, sir, are my hero. Any chance you remember wha...	0	0	0	0	0	0
6	00025465d4725e87	"" Congratulations from me as well, use the tools well...	0	0	0	0	0	0
7	0002bcb3da6cb337	COCKSUCKER BEFORE YOU PISS AROUND ON MY WORK	1	1	1	0	1	0
8	00031b1e95af7921	Your vandalism to the Matt Shirvington article has be...	0	0	0	0	0	0
9	00037261f536c51d	Sorry if the word 'nonsense' was offensive to you. Any...	0	0	0	0	0	0
10	00040093b2687caa	alignment on this subject and which are contrary to t...	0	0	0	0	0	0
11	0005300084f90edc	"" Fair use rationale for Image:Wonju.jpg Thanks for u...	0	0	0	0	0	0
12	00054a5e18b50dd4	bbq be a man and lets discuss it-maybe over the pho...	0	0	0	0	0	0
13	0005c987bdfc9d4b	Hey... what is it.. @ | talk . What is it... an exclusive gr...	1	0	0	0	0	0
14	0006f16e4e9f292e	Before you start throwing accusations and warnings a...	0	0	0	0	0	0
15	00070ef96486d6f9	Oh, and the girl above started her arguments with me...	0	0	0	0	0	0

Goal

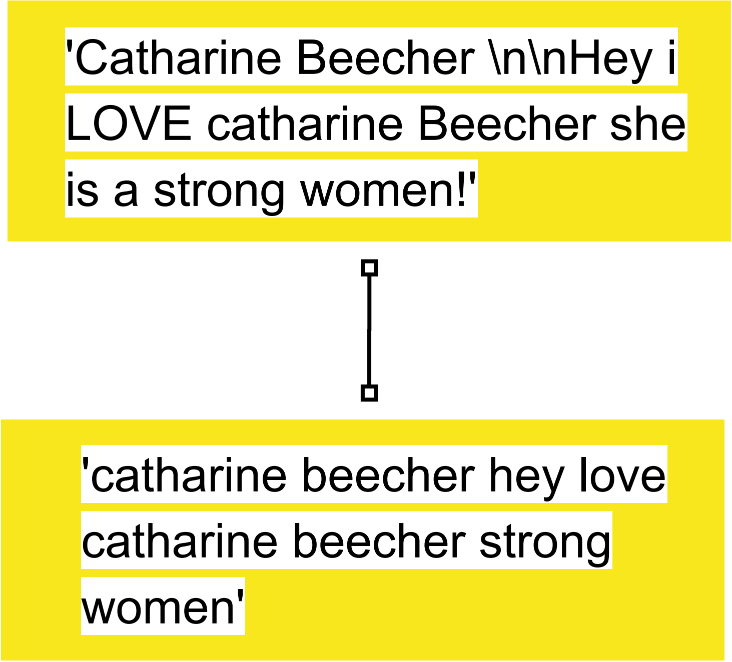
Building a multi-labeled model that is capable of detecting six types of of toxic comments like threats, obscenity, insults, and identity-based hate.

Creating Variables

Variable Type	**Variables**	**Reason**
Readability	# of sentences # of words Avg length of words % of Unique word # of letters Avg length of the words % of words in normal dictionary	Forms of text usually contain unconscious features of being toxic: e.g. In finance industry, a really long email usually means you are trying to cover something
Emotional Letters	% of Uppercase words	Typing in all caps is usually considered yelling
Emotional Punctuations	! ? ^	Those punctuations contain strong emotions which could lead to being toxic
Sentiment	Polarity Subjective	Sentiment scores contains the attitude of the comments

Data Cleaning

- Lowercase
- Remove Stopwords
- Remove Punctuation
- Remove Numbers
- Remove Non-alphanumeric Characters
- Remove Elements like ip, user, url



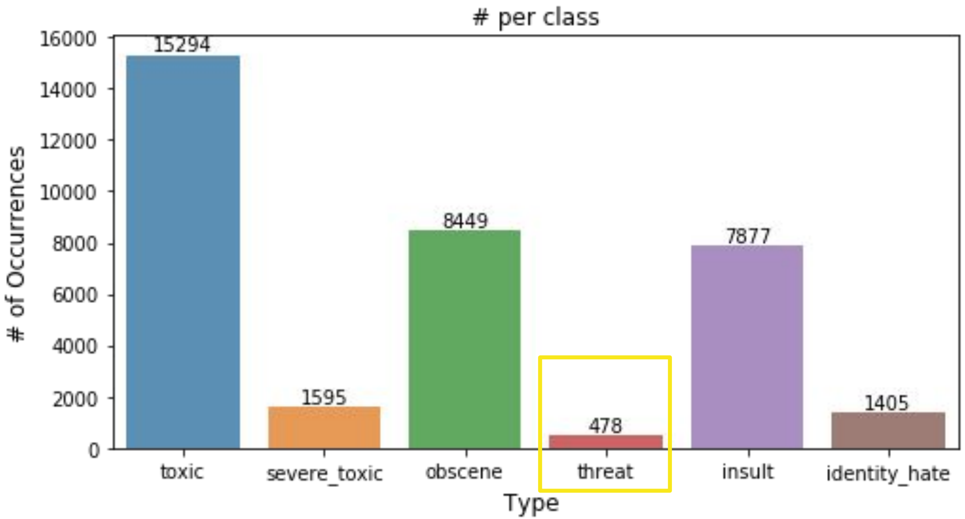
'Catharine Beecher \n\nHey i LOVE catharine Beecher she is a strong women!'

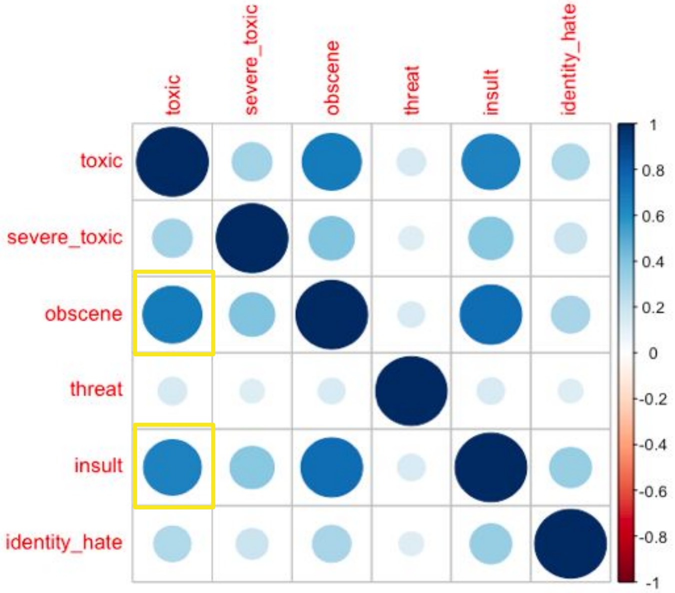
'catharine beecher hey love catharine beecher strong women'

Data Exploration & Visualization

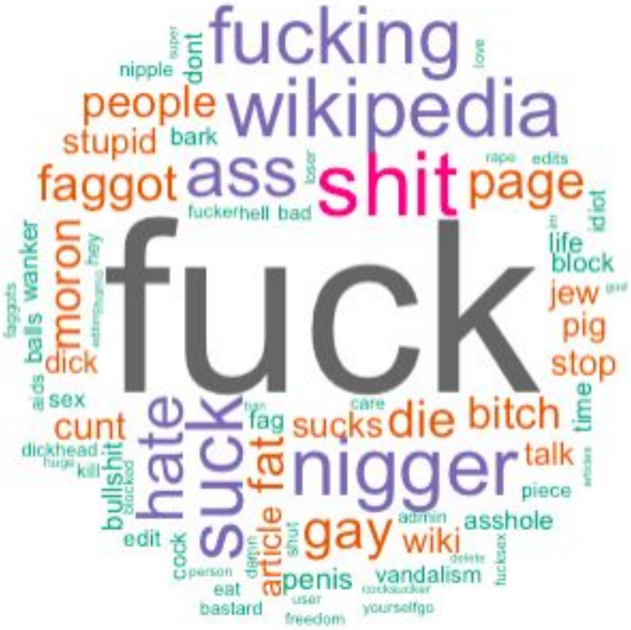
- Distribution and Correlation Matrix of Six Labels
- Word Cloud for Overall Comment Texts (Top 200)
- Word Clouds for Six Labels (Top 100)

Distribution & Correlation Between Six Labels

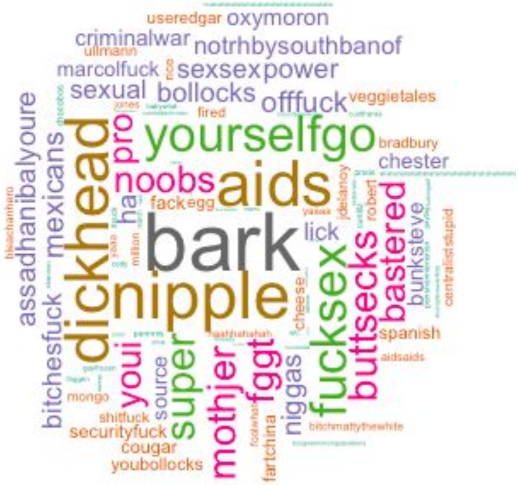




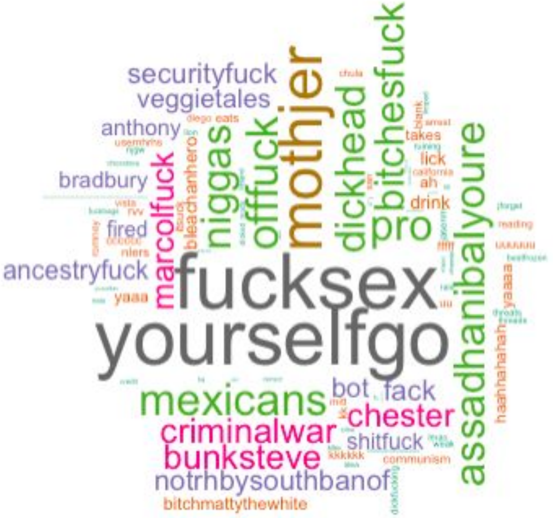
Word Cloud (Top 200) for Toxic Comments



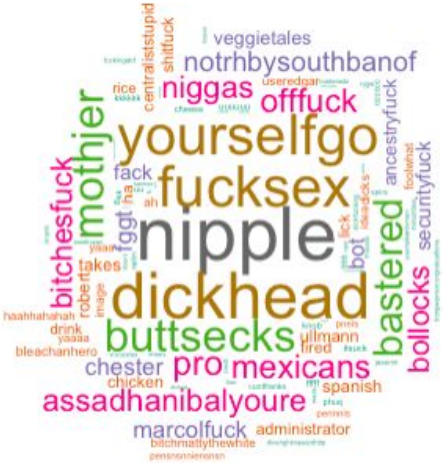
Word Cloud (Top 100) for Each Label (1/2)



Toxic

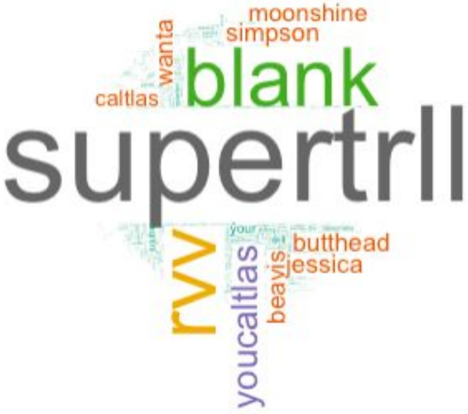


Severe Toxic

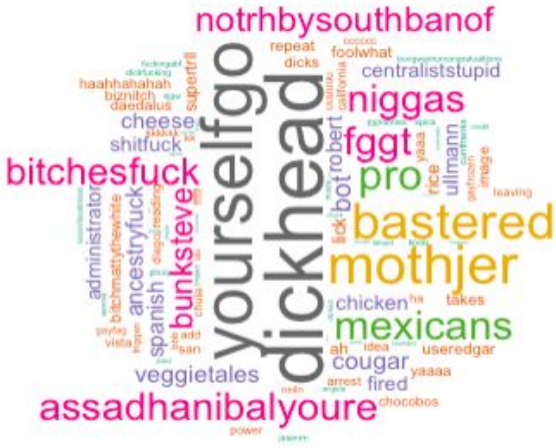


Obscene

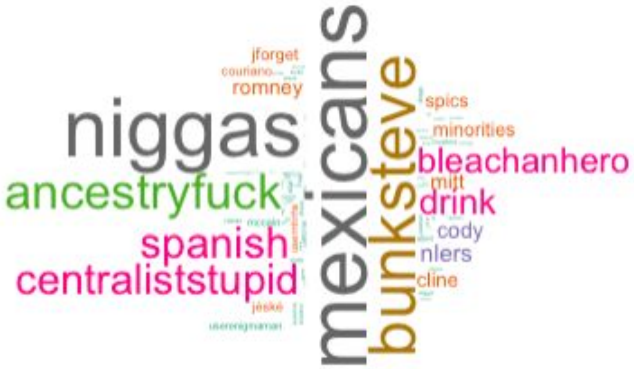
Word Cloud (Top 100) for Each Label (2/2)



Threat



Insult



Identity hate

Model Building

Dependent Variable:

Multi-labeled

Three Methods:

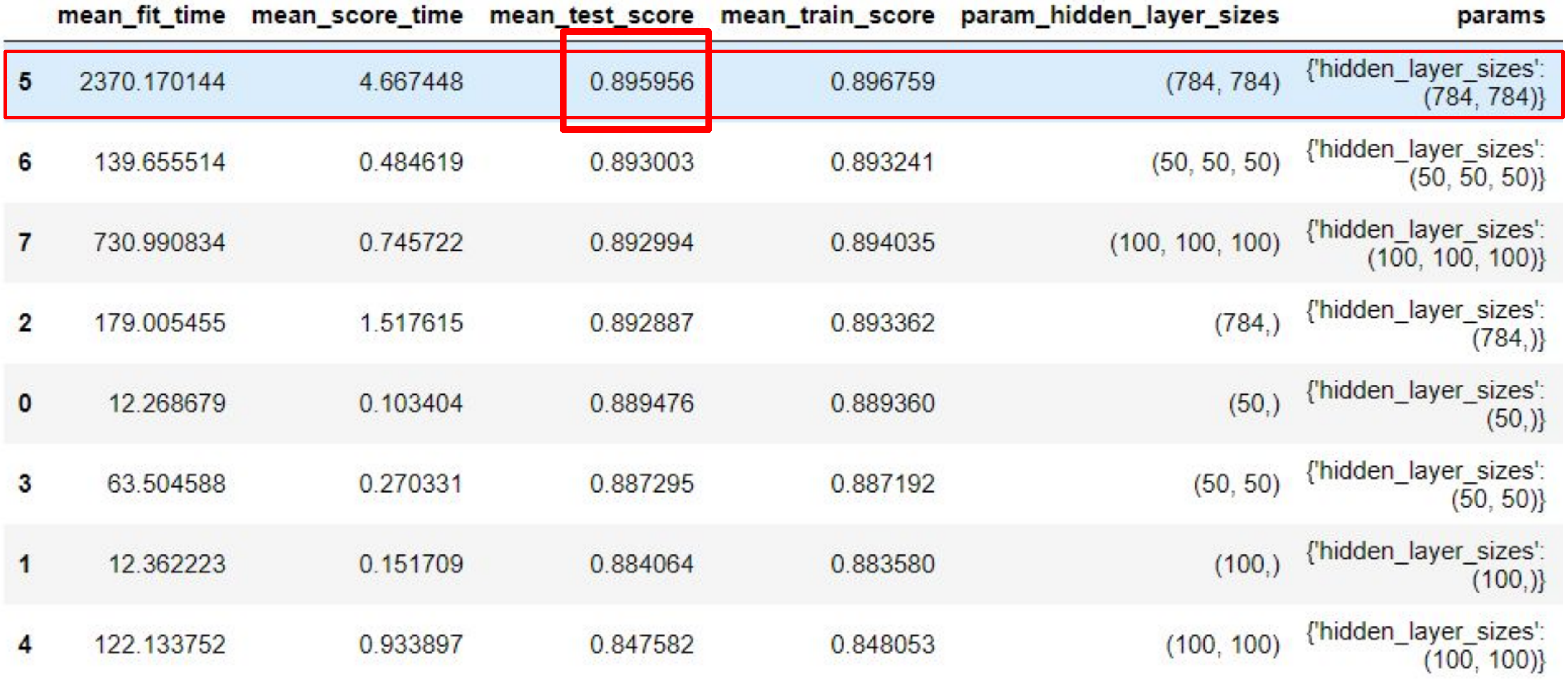
1. Only Variables
2. Term Document Matrix
3. Variables + TDM

We split the dataset into Train(70%) and Test(30%)

1. Generated Variables Only

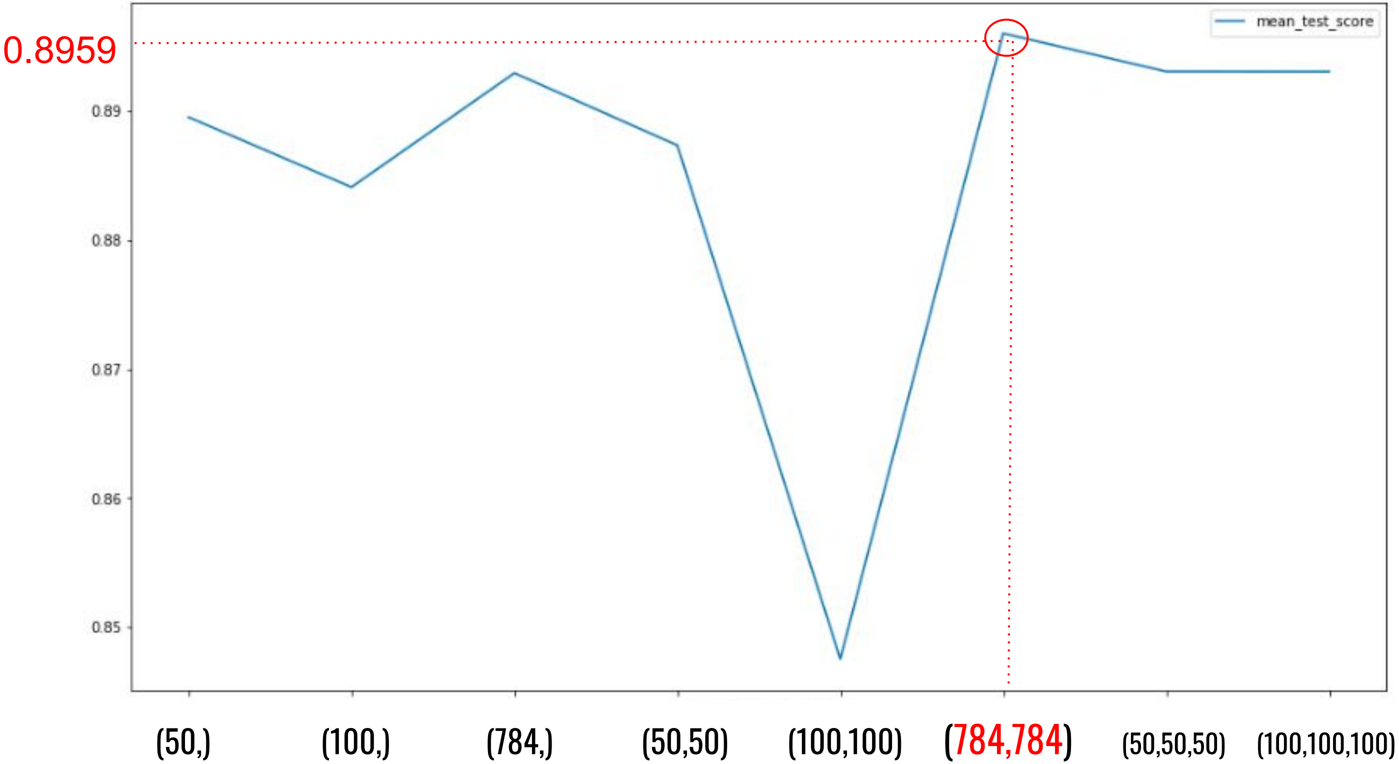
Model Name	Accuracy
Random Forest	0.8797
KNN	0.8787
MLP Classifier Neural Network ***	0.8941
~~Naive Bayes~~	0.1194
~~Decision Tree~~	0.8308

Further Explore



	mean_fit_time	mean_score_time	mean_test_score	mean_train_score	param_hidden_layer_sizes	params
5	2370.170144	4.667448	0.895956	0.896759	(784, 784)	{'hidden_layer_sizes': (784, 784)}
6	139.655514	0.484619	0.893003	0.893241	(50, 50, 50)	{'hidden_layer_sizes': (50, 50, 50)}
7	730.990834	0.745722	0.892994	0.894035	(100, 100, 100)	{'hidden_layer_sizes': (100, 100, 100)}
2	179.005455	1.517615	0.892887	0.893362	(784,)	{'hidden_layer_sizes': (784,)}
0	12.268679	0.103404	0.889476	0.889360	(50,)	{'hidden_layer_sizes': (50,)}
3	63.504588	0.270331	0.887295	0.887192	(50, 50)	{'hidden_layer_sizes': (50, 50)}
1	12.362223	0.151709	0.884064	0.883580	(100,)	{'hidden_layer_sizes': (100,)}
4	122.133752	0.933897	0.847582	0.848053	(100, 100)	{'hidden_layer_sizes': (100, 100)}

Mean Test Score & Hidden Layer Sizes



2. Text Analytics Results

Model Name	Accuracy
MLP	0.8700
KNN	0.8812
Random Forest ***	0.8973

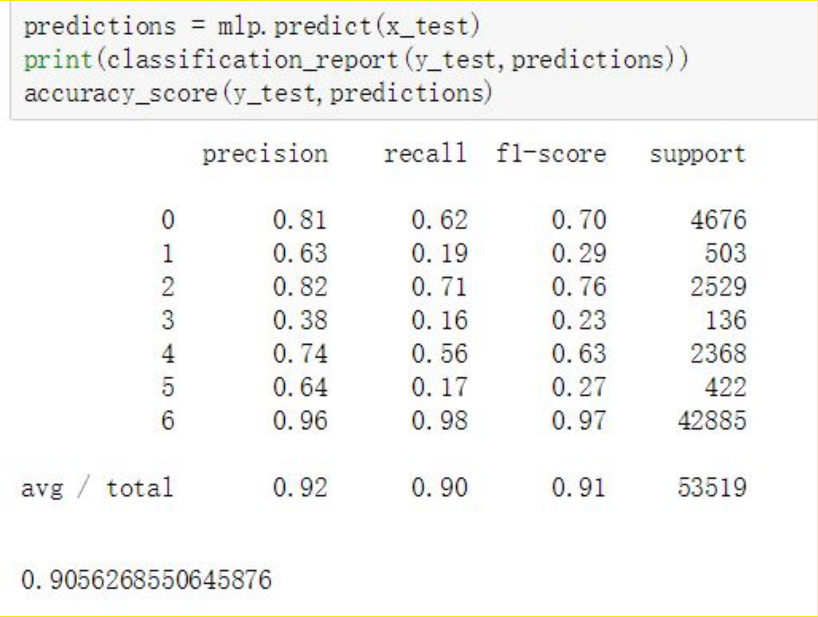
3. Combined Model

Model Name	Accuracy
MLP ***	0.9056
KNN	0.8903
Random Forest	0.9043

MLP Classifier Neural Network

Accuracy: 90.56%

Hidden_layer_sizes
=(30, 30, 30)



```
predictions = mlp.predict(x_test)
print(classification_report(y_test,predictions))
accuracy_score(y_test,predictions)
```

```
             precision    recall  f1-score   support

          0       0.81      0.62      0.70      4676
          1       0.63      0.19      0.29       503
          2       0.82      0.71      0.76      2529
          3       0.38      0.16      0.23       136
          4       0.74      0.56      0.63      2368
          5       0.64      0.17      0.27       422
          6       0.96      0.98      0.97     42885

avg / total       0.92      0.90      0.91     53519


0.9056268550645876
```

Keras Text Analytics

Model Name	Accuracy
RNN LSTM ***	0.9821

***LSTM: long short term memory

LSTM (Using Keras)

Order of words matter

Accuracy: 98.21%

Tokenizer:

max_features = 20000

Padding the words to 200:

maxlen = 200

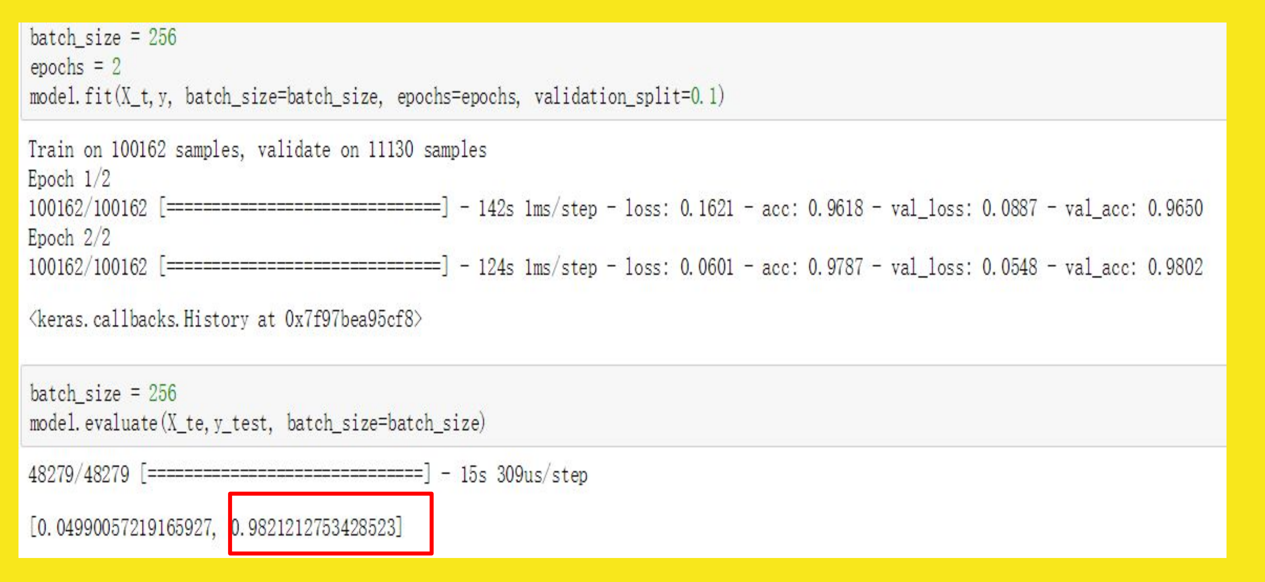
Hidden Layers:

Embedding for LSTM,

Global Max Pooling

Dense,

Dropout



```
batch_size = 256
epochs = 2
model.fit(X_t,y, batch_size=batch_size, epochs=epochs, validation_split=0.1)
```

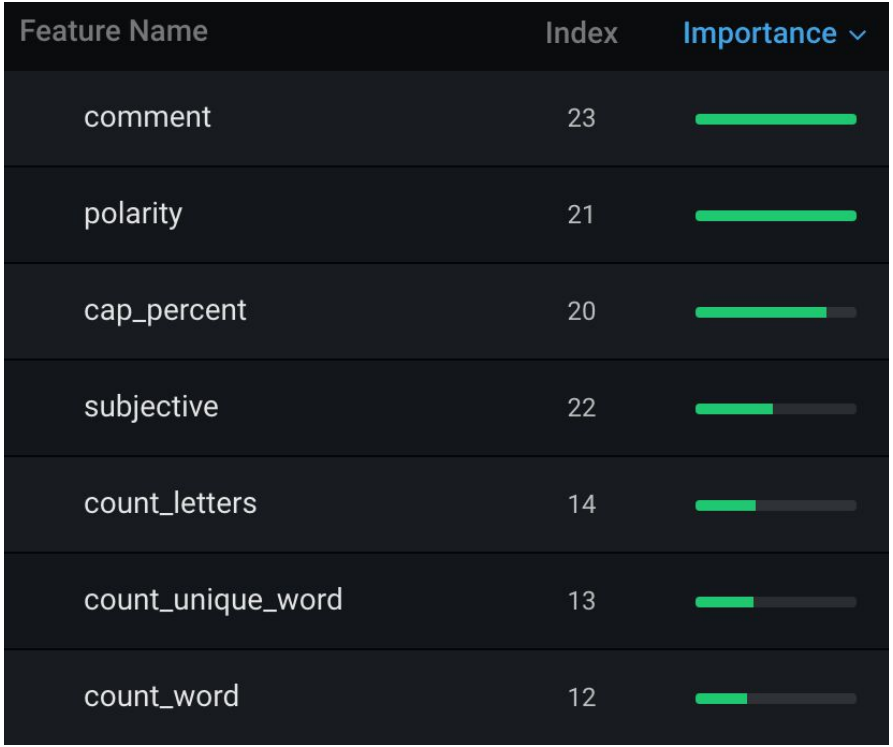
```
Train on 100162 samples, validate on 11130 samples
Epoch 1/2
100162/100162 [==============================] - 142s 1ms/step - loss: 0.1621 - acc: 0.9618 - val_loss: 0.0887 - val_acc: 0.9650
Epoch 2/2
100162/100162 [==============================] - 124s 1ms/step - loss: 0.0601 - acc: 0.9787 - val_loss: 0.0548 - val_acc: 0.9802

<keras.callbacks.History at 0x7f97bea95cf8>
```

```
batch_size = 256
model.evaluate(X_te,y_test, batch_size=batch_size)
```

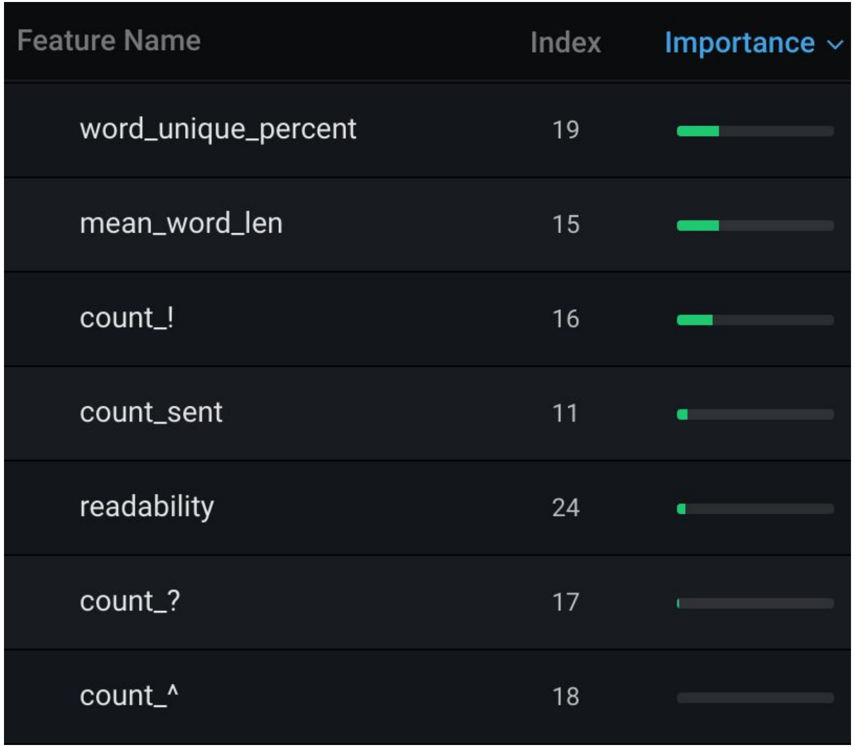
```
48279/48279 [==============================] - 15s 309us/step

[0.04990057219165927, 0.9821212753428523]
```

DataRobot - Binary Classification

Feature Importance

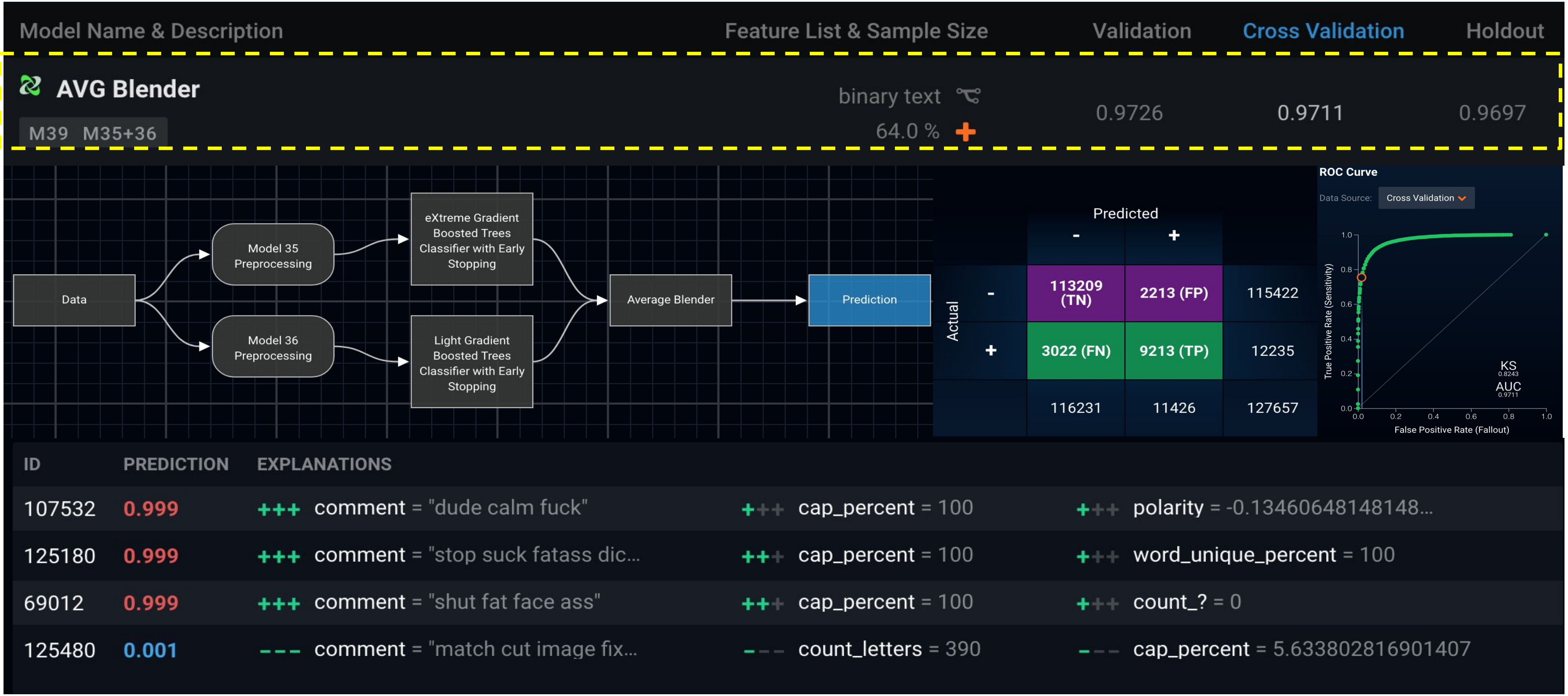


Feature Name	Index	Importance ⌄
comment	23	
polarity	21	
cap_percent	20	
subjective	22	
count_letters	14	
count_unique_word	13	
count_word	12	



Feature Name	Index	Importance ⌄
word_unique_percent	19	
mean_word_len	15	
count_!	16	
count_sent	11	
readability	24	
count_?	17	
count_^	18	

DataRobot Predictions



Model Name & Description	Feature List & Sample Size	Validation	Cross Validation	Holdout
AVG Blender M39 M35+36	binary text 64.0 %	0.9726	0.9711	0.9697

Data
Model 35 Preprocessing
Model 36 Preprocessing
eXtreme Gradient Boosted Trees Classifier with Early Stopping
Light Gradient Boosted Trees Classifier with Early Stopping
Average Blender
Prediction

		Predicted		
		-	+	
Actual	-	113209 (TN)	2213 (FP)	115422
	+	3022 (FN)	9213 (TP)	12235
		116231	11426	127657

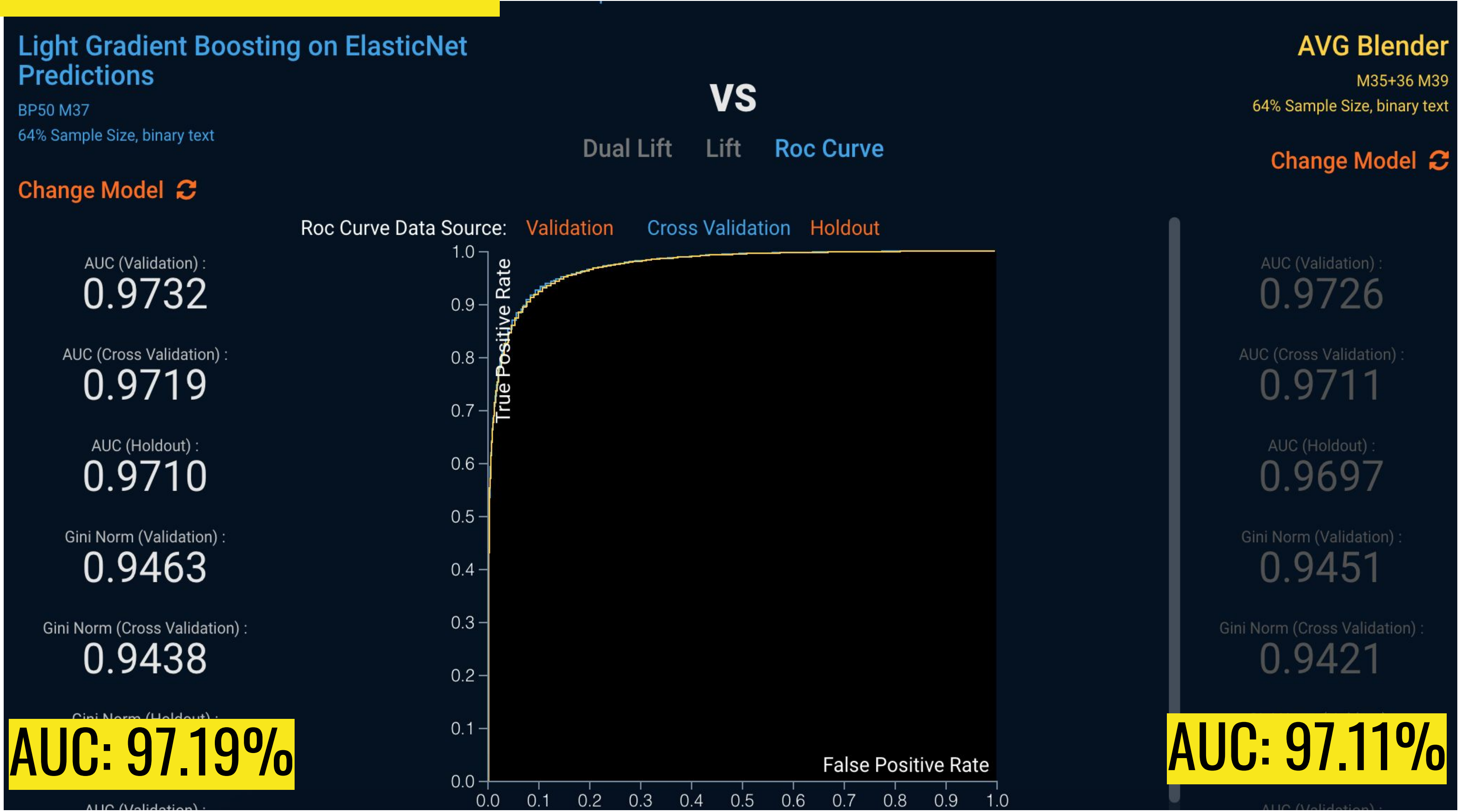
ROC Curve

Data Source: Cross Validation

1.0
0.8
0.6
0.4
0.2
0.0
True Positive Rate (Sensitivity)
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate (Fallout)
KS
0.8243
AUC
0.9711

ID	PREDICTION	EXPLANATIONS		
107532	0.999	+++ comment = "dude calm fuck"	+++ cap_percent = 100	+++ polarity = -0.13460648148...
125180	0.999	+++ comment = "stop suck fatass dic...	+++ cap_percent = 100	+++ word_unique_percent = 100
69012	0.999	+++ comment = "shut fat face ass"	+++ cap_percent = 100	+++ count_? = 0
125480	0.001	--- comment = "match cut image fix...	--- count_letters = 390	--- cap_percent = 5.633802816901407

DataRobot Predictions



Light Gradient Boosting on ElasticNet Predictions

BP50 M37

64% Sample Size, binary text

Change Model

VS

Dual Lift | Lift | Roc Curve

AVG Blender

M35+36 M39

64% Sample Size, binary text

Change Model

Roc Curve Data Source: Validation | Cross Validation | Holdout

	Light Gradient Boosting on ElasticNet Predictions	AVG Blender
AUC (Validation) :	0.9732	0.9726
AUC (Cross Validation) :	0.9719	0.9711
AUC (Holdout) :	0.9710	0.9697
Gini Norm (Validation) :	0.9463	0.9451
Gini Norm (Cross Validation) :	0.9438	0.9421

Gini Norm (Holdout) :

AUC (Validation) :

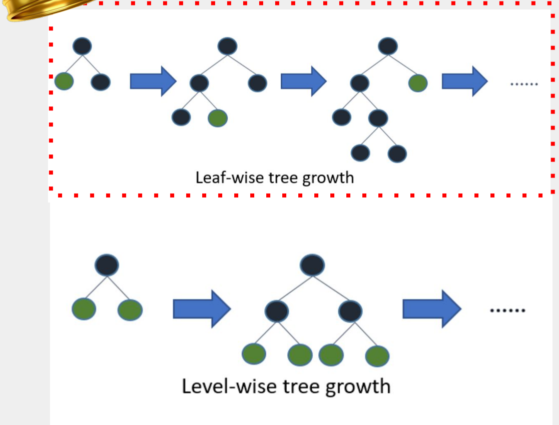
AUC: 97.19%

AUC: 97.11%

DataRobot Predictions

LightGBM

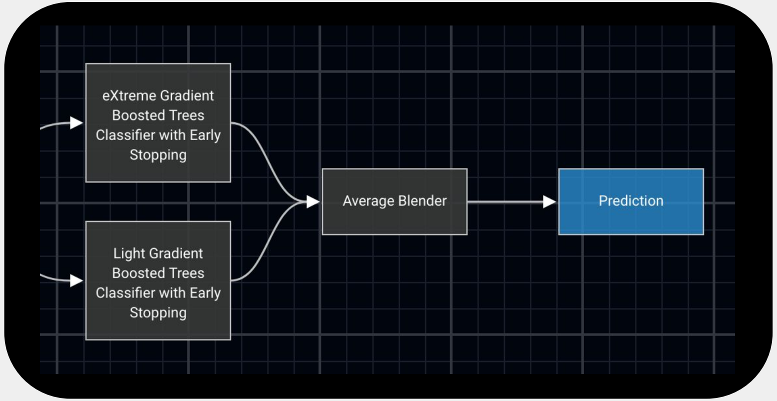
Best Model



- high speed
- handle the large size
- takes lower memory to run
- focuses on accuracy of results

AUC: 97.19%

AVG Blender

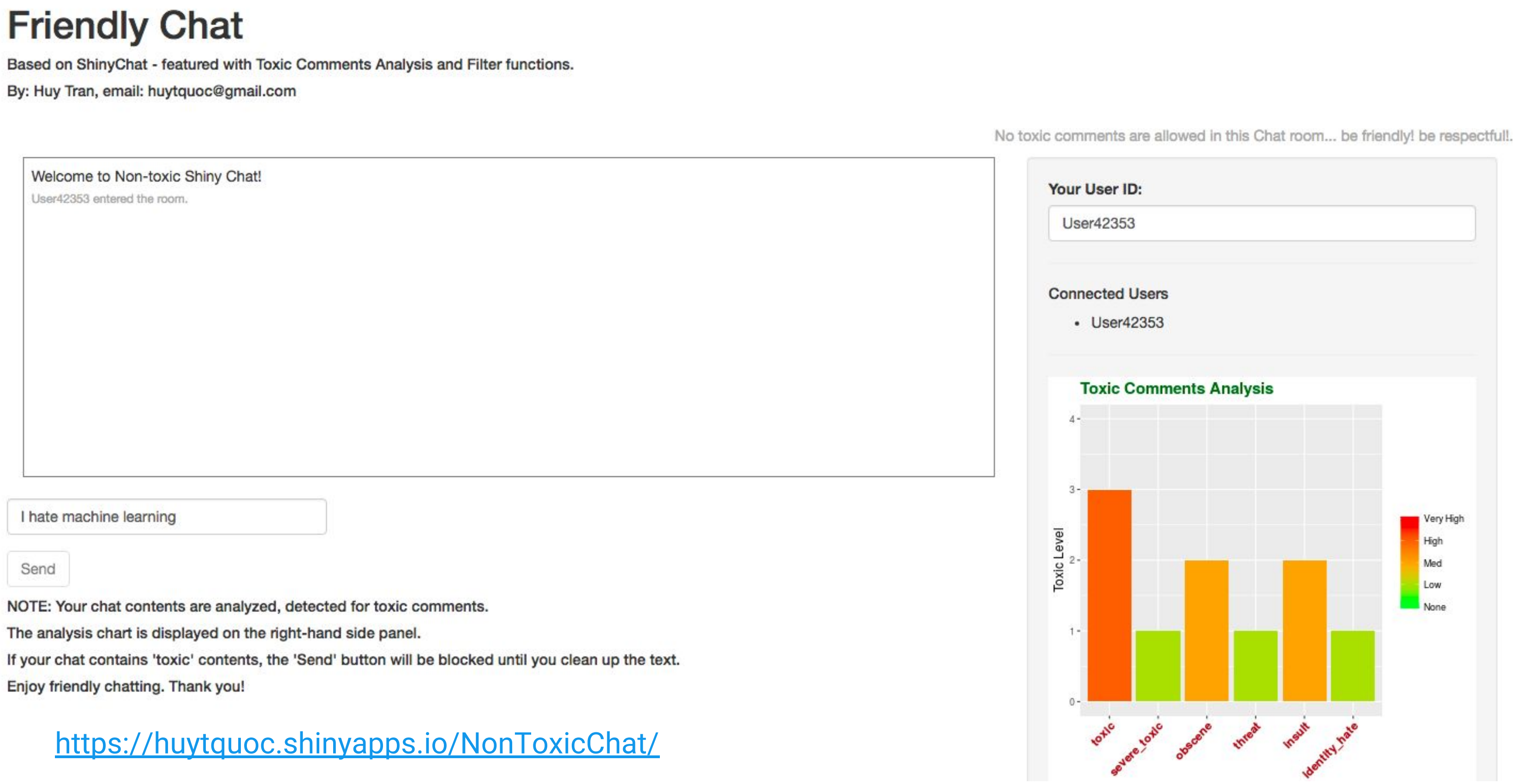

eXtreme Gradient Boosted Trees Classifier with Early Stopping
Light Gradient Boosted Trees Classifier with Early Stopping
Average Blender
Prediction

Averages the predictions of each input prediction

takes the predictions from several input models, and averages them together into a meta-model.

AUC: 97.11%

Business Application Demo



Friendly Chat

Based on ShinyChat - featured with Toxic Comments Analysis and Filter functions.

By: Huy Tran, email: huytquoc@gmail.com

No toxic comments are allowed in this Chat room... be friendly! be respectfull!.

Welcome to Non-toxic Shiny Chat!

User42353 entered the room.

I hate machine learning

Send

NOTE: Your chat contents are analyzed, detected for toxic comments.

The analysis chart is displayed on the right-hand side panel.

If your chat contains 'toxic' contents, the 'Send' button will be blocked until you clean up the text.

Enjoy friendly chatting. Thank you!

Your User ID:

User42353

Connected Users

- User42353

https://huytquoc.shinyapps.io/NonToxicChat/

Limitations

- **Manual Labels** - potentially involved subjective judgements
- Can only accurately detect toxic comments **in English**
- Cannot detect image or video **forms** of toxic comment
- Hard to make **Shiny** work

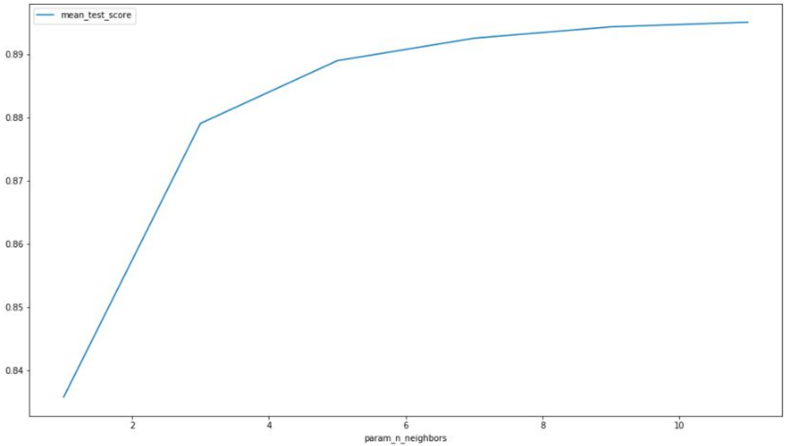
Thank you!

Questions?

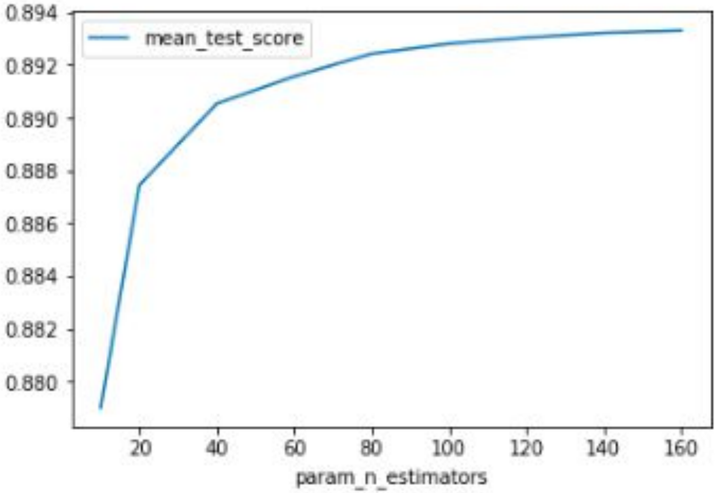
Appendix

Accuracy Plots (Variables Only)

KNN

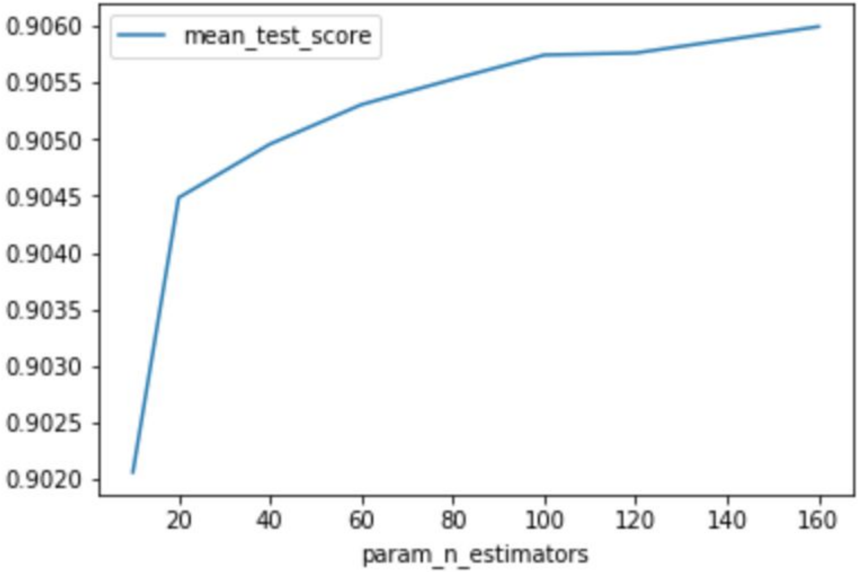


Random Forest

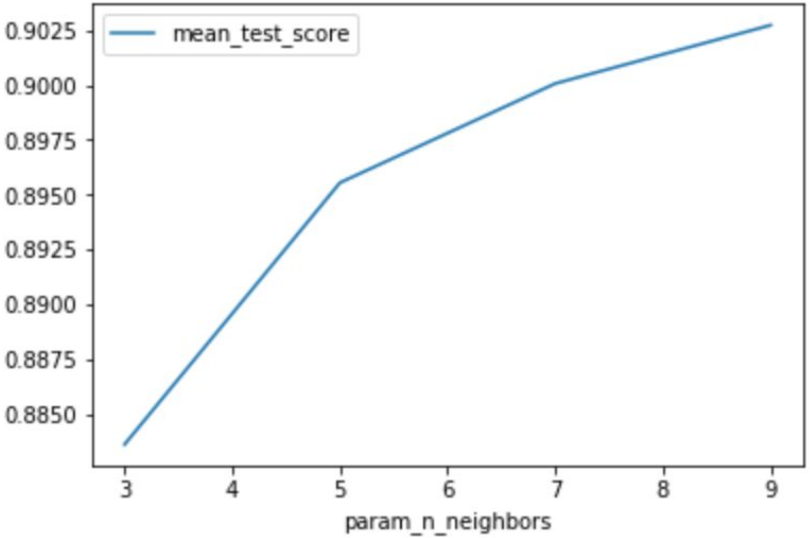


Accuracy Plots (Text Only)

Random Forest



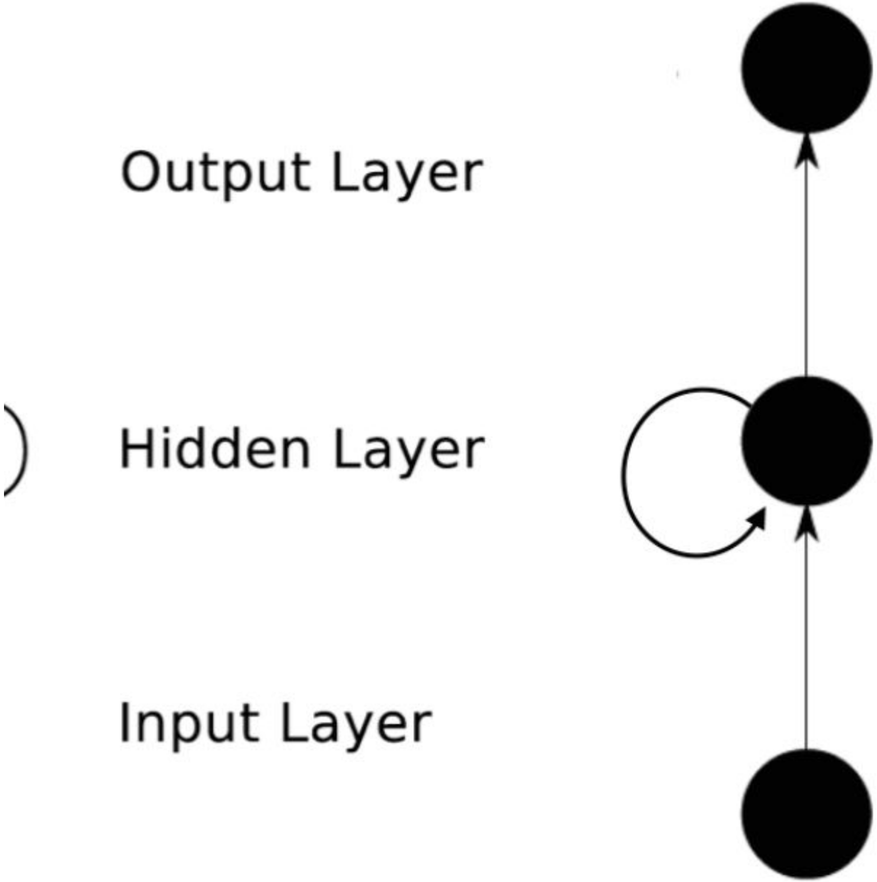
KNN



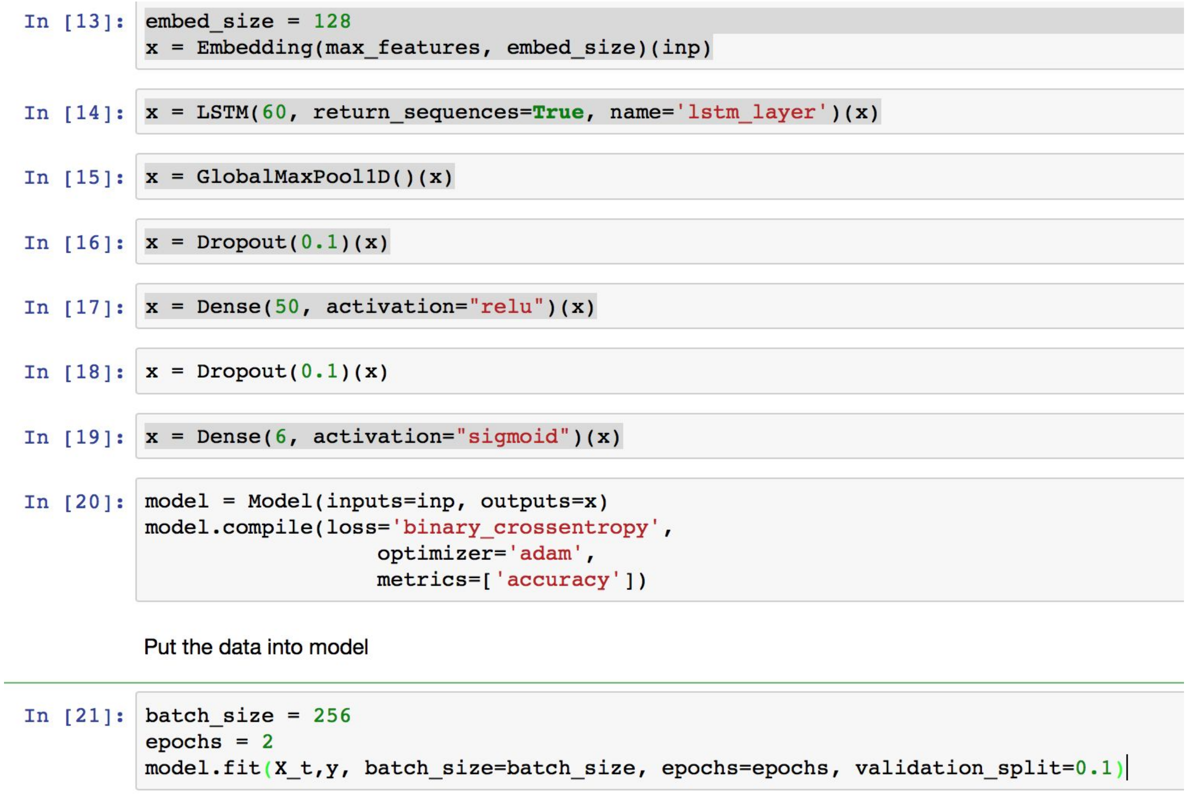
RNN

A **recurrent neural network** (**RNN**) is a class of artificial neural network where connections between nodes form a directed graph along a sequence. This allows it to exhibit dynamic temporal behavior for a time sequence. Unlike feedforward neural networks, RNNs can use their internal state (memory) to process sequences of inputs.

Long short-term memory. Long short-term memory (LSTM) units (or blocks) are a building unit for layers of a recurrent neural **network** (RNN).



Keras Model Code:



```
In [13]: embed_size = 128
         x = Embedding(max_features, embed_size)(inp)
```

```
In [14]: x = LSTM(60, return_sequences=True, name='lstm_layer')(x)
```

```
In [15]: x = GlobalMaxPool1D()(x)
```

```
In [16]: x = Dropout(0.1)(x)
```

```
In [17]: x = Dense(50, activation="relu")(x)
```

```
In [18]: x = Dropout(0.1)(x)
```

```
In [19]: x = Dense(6, activation="sigmoid")(x)
```

```
In [20]: model = Model(inputs=inp, outputs=x)
         model.compile(loss='binary_crossentropy',
                           optimizer='adam',
                           metrics=['accuracy'])
```

Put the data into model

```
In [21]: batch_size = 256
         epochs = 2
         model.fit(X_t,y, batch_size=batch_size, epochs=epochs, validation_split=0.1)
```

References:

1. **Multiclass and multilabel algorithms**

http://scikit-learn.org/stable/modules/multiclass.html

2. **Toxic Comments Classification, and 'Non-toxic' Chat Application**

https://nycdatascience.com/blog/student-works/toxic-comments-classification-and-non-toxic-chat-application/

3. **Quick draw image recognition**

https://github.com/kradolfer/quickdraw-image-recognition