

# Hate speech in news/media



Rene Angeles  
Data Incubator Interview  
Nov / 2019

# Intro

The information/data era has its charm but also **challenges**



Lots of effort in:

- Fake news detection / factualness
- Toxicity detection on comments (obscenities, insults, treats, hate-speech) on social media

**Motivation: un-hack democracies / prevent violence.**

# Our work: quantify speech features

NLP + deep learn the speech of news/media

After a little work:

( Torch (LSTM, GRU, Conv), Sklearn, Seaborn, Torchtext, Spicy (Stopwords, Stemming) Embeddings, Glove, Fasttext, CUDA, SageMaker, API, Lambda, hyperparam, Recall emphasis, Kaggle benchmark 97.7% (Leader 98.8 auc\_roc), 4 data sets (200 K) )

The problem/project definition is focusing on measuring the hate speech in news, and I already have the proof of concept results

- a) deliberate attack
- b) directed towards a specific group of people
- c) motivated by aspects of the group's identity.

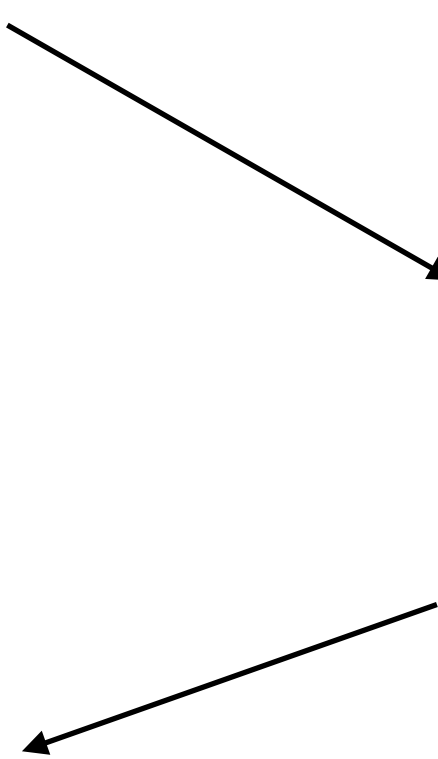
# Project first application

Hate speech data sets:

- 150K Wikipedia comments  
Jigsaw/Google toxic comments.
- 10K White supremacist forum  
arXiv:1705.00648

Fake/True news:

12.8K Counts/Metadata on  
True, mostly true, barely true,  
half true, fake, pants on fire  
arXiv:1705.00648



```
graph LR; A[Hate speech data sets] --> C[Torch (LSTM, GRU, Conv), Torchtext, Spicy (Stopwords, Stem) Embed (Glove, Fasttext) hyperpar, Recall Emphasis]; B[Fake/True news] --> C;
```

Torch (LSTM, GRU, Conv), Torchtext, Spicy (Stopwords, Stem) Embed (Glove, Fasttext) hyperpar, Recall Emphasis

# Pay back time!

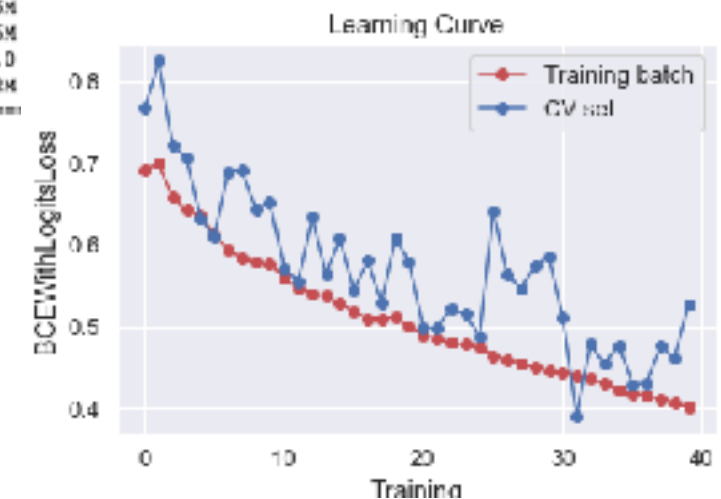
Hate Speech Dataset from a White Supremacy Forum  
<https://arxiv.org/pdf/1809.04444.pdf>

Data	ACC	RECALL	PRECISION	F1	ROC_AUC
Test/ Hate Speech	0.839	0.890	0.803	0.844	0.921
Test/ Jigsaw	0.858	0.742	0.044	0.084	0.879

Layer	Kernel shape	Output shape	Params	Mult-Adds
0_embedding	[100, 20002]	[256, 191, 100]	2.0002M	2.0002M
1_conv.conv2d_0	[1, 100, 2, 100]	[256, 100, 130, 1]	20.1k	3.8M
2_conv.conv2d_1	[1, 100, 3, 100]	[256, 100, 139, 1]	30.1k	5.67M
3_conv.conv2d_2	[1, 100, 4, 100]	[256, 100, 138, 1]	40.1k	7.52M
4_conv.conv2d_3	[1, 100, 5, 100]	[256, 100, 137, 1]	50.1k	9.35M
5_conv.conv2d_4	[1, 100, 6, 100]	[256, 100, 136, 1]	60.1k	11.16M
6_dropout	-	[256, 500]	-	-
7_fc	[500, 6]	[256, 6]	3.005k	3.0k

Totals	
Total params	2.233796M
Trainable params	2.233796M
Non-trainable params	0.0
Mult-Adds	39.5032M

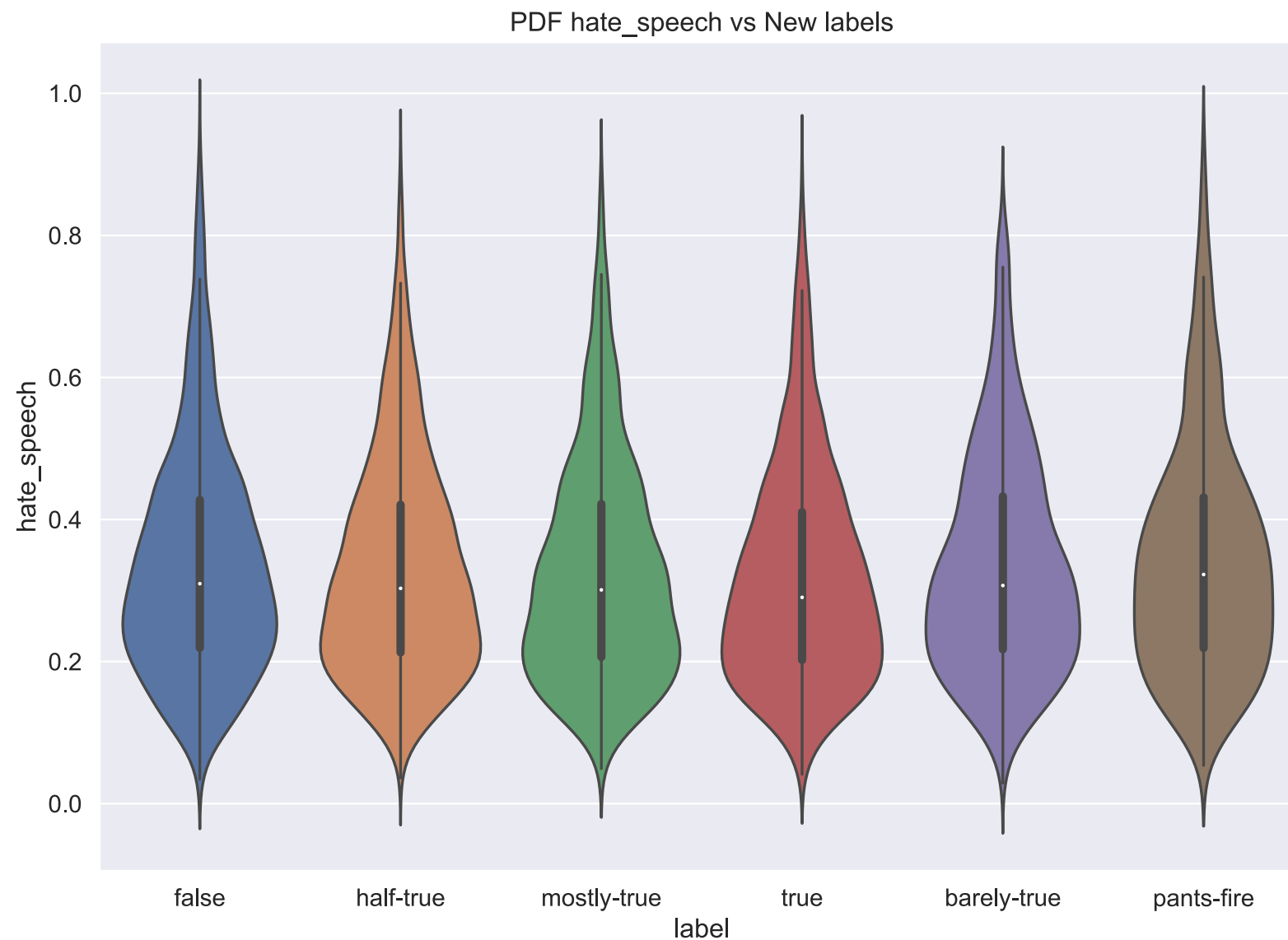


```
In [142]: quote = """When Mexico sends its people, they're
not sending their best. They're not sending you.
They're not sending you. They're sending people
that have lots of problems, and they're bringing
those problems with us. They're bringing drugs.
They're bringing crime. They're rapists. And some,
I assume, are good people."""
```

```
In [143]: custom_speech(model, quote)
```

```
Out[143]: 'Hate speech detected'
```

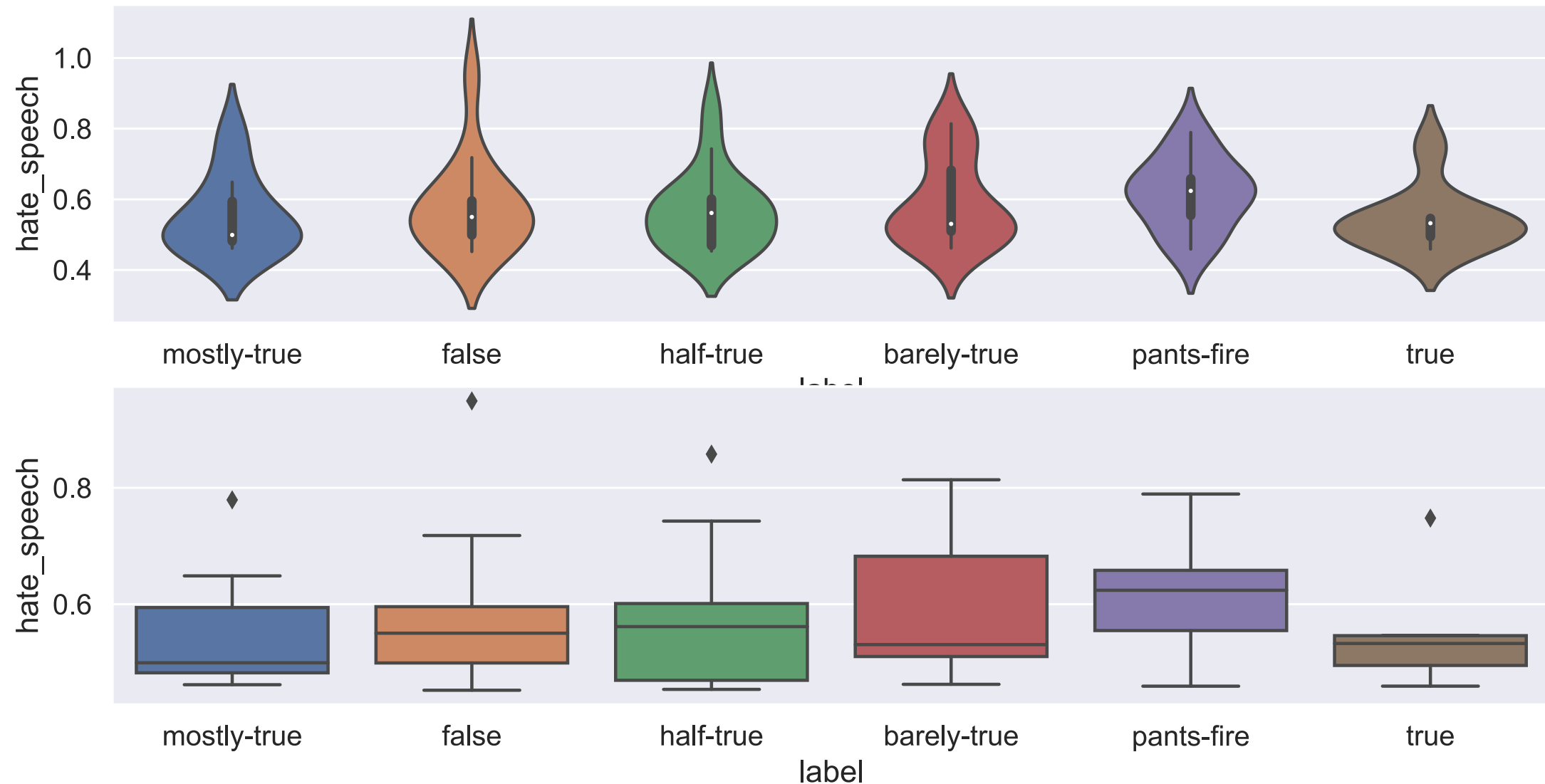
# Easy... most news



Facebook is manly after the tails

# Hate speech in fake news

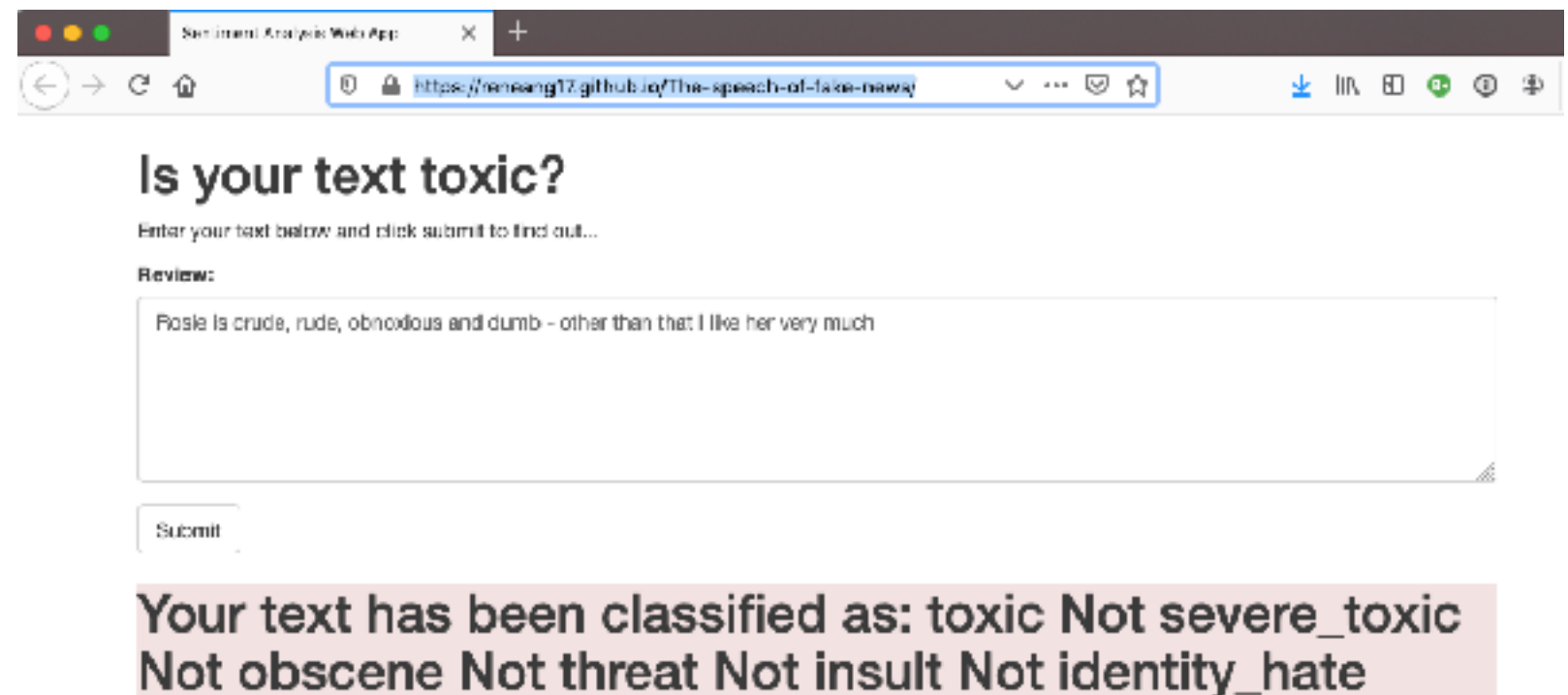
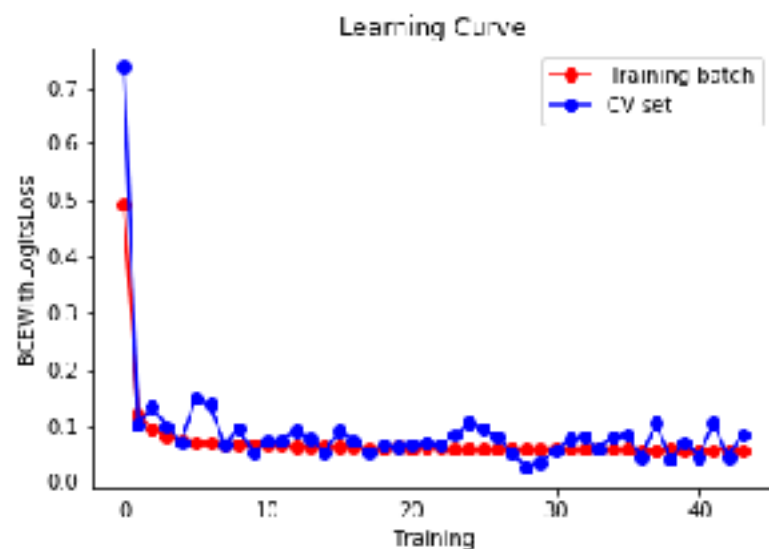
PDF of hate\_speech > 0.45 vs Fake News labels



“Liar, Liar Pants on Fire”: A New Benchmark Dataset for Fake News Detection

# Deployment

App that announces results. Determine whether a comment is toxic / obscene/ threat / identity-hate



[https://reneang17.github.io/  
The-speech-of-fake-news/](https://reneang17.github.io/The-speech-of-fake-news/)

Side prod: Utils to use  
TorchText with  
SageMaker



# References / Data sets

- “Liar, Liar Pants on Fire”: A New Benchmark Dataset for Fake News Detection
- Jigsaw/Google toxic comment classification challenge
- Hate Speech Dataset from a White Supremacy Forum
- Getting real about fake news
- Bag of tricks for efficient text classification