Do Men and Women Tweet Differently?

Short answer:

Sometimes they do,
and sometimes they don’t.

(And even if the latter case means
we can’t usually guess gender from tweets,
we’re at least good at picking out
the ones where we can.)

Specifically,

among tweeters with gendered first names,

there are 13.4% of tweets

where we can say

with probability > .75

whether the tweet came from someone

with a male or female name

Sneak preview of model results

Test accuracy is high for high-probability predictions

PREDICTION	COUNT	ACCURACY
PROB(Female) > 75%	**690**	**81%**
PROB(Female) between 50% and 75%	**4379**	**61%**
PROB(Male) between 50% and 75%	**4666**	**60%**
PROB(Male) > 75%	**715**	**83%**

THE STATIC MODEL

DATA

- Random Tweets from 2019-05-21 thru 2019-06-01
- Only English tweets with gender easily identified from first name
- Modal gender (male) downsampled to balance sample by gender
- Train/Valid/Test split by time and user ID

Model 1

Universal Sentence Encoder (Large) (“USE-L”) from TensorFlow Hub

Transformer-based 512-dimensional sentence embedding model

Training Model 1

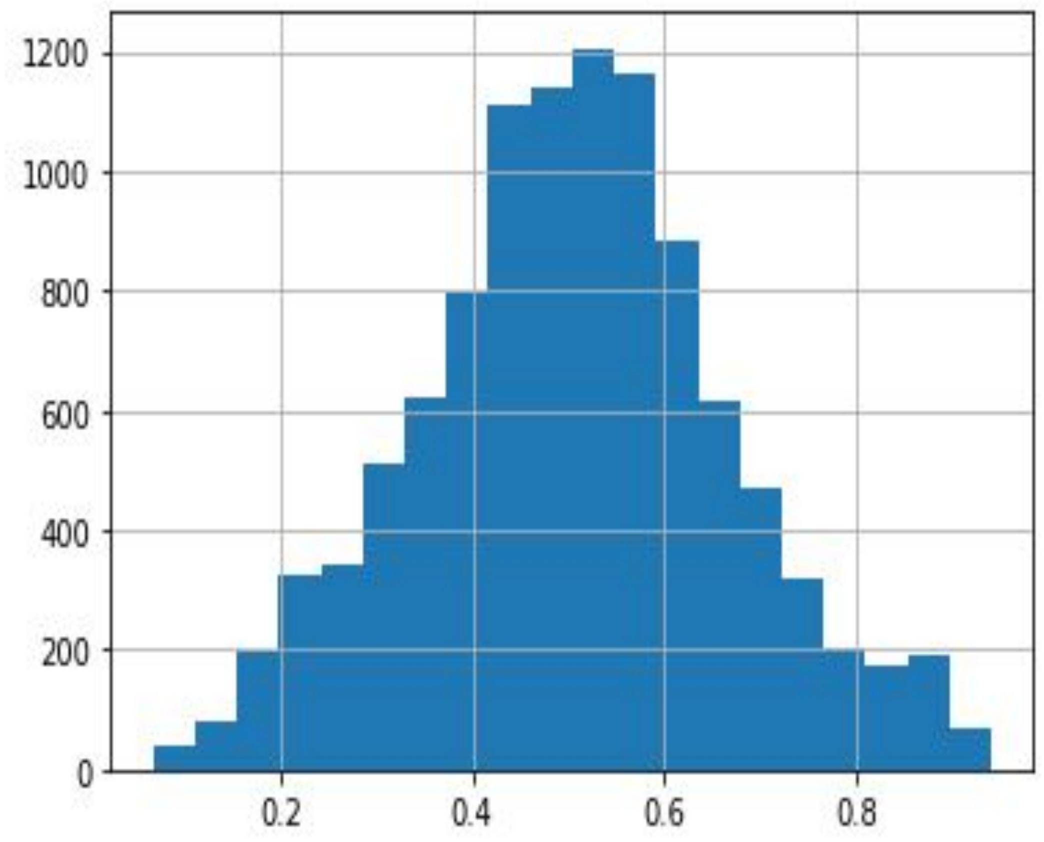
- Add 512-dimensional dense layer, 80% dropout, and sigmoid prediction
- Train with original embeddings, then fine-tune embeddings
- Over 200 million parameters
- Can only train full model for one epoch before it overfits

Model 1 Test Results

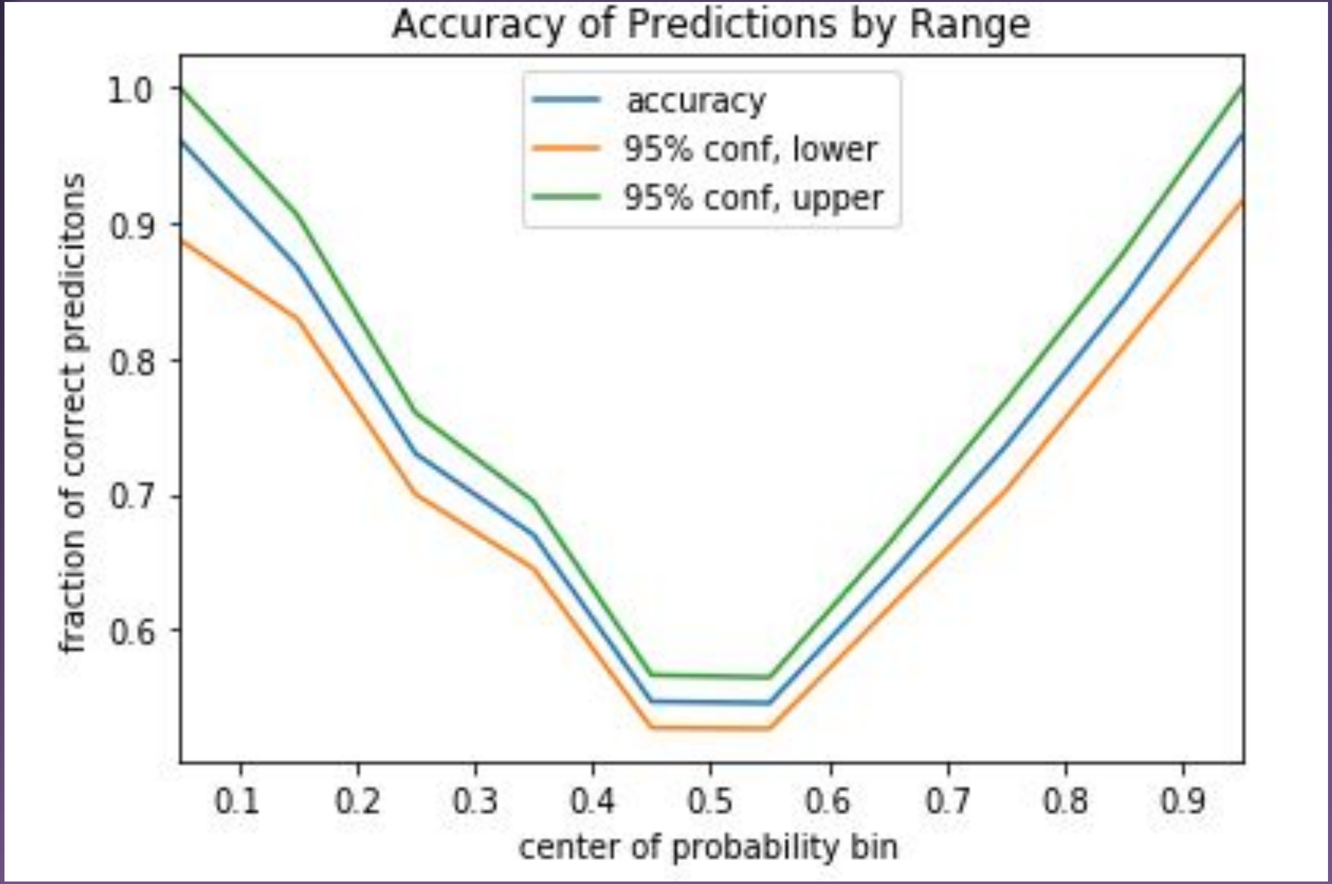
- Overall Accuracy 62.8%
- 0.683 ROC AUC

Overall Confusion Matrix		Predicted	
		Female	Male
Actual	Female	3205	2020
	Male	1864	3361

Distribution of Test Set Predictions (probability of being male)



Generalization is very good: actual label frequencies in test set correspond tightly to predicted probabilities



Model 2

“Home-grown” LSTM model

Produces 240-dimensional tweet representation at top hidden layer

Model 2 Characteristics

- Bottom embedding layer initialized with Glove vectors
- Two LSTM branches, 2 levels deep, alternating forward-then-backward and backward-then-forward
- Extra parallel dense layers with residual connections around them
- Multiple types of dropout, with 75% dropout at prediction layer

Training Model 2

- Train with original (Glove) embeddings, then fine-tune embeddings
- About 6.3 million total parameters
- Trains for 35 epochs initially and 29 for fine tuning

Model 2 Test Results

- Overall Accuracy 60.7%
- 0.655 ROC AUC

Overall Confusion Matrix		Predicted	
		Female	Male
Actual	Female	3287	1938
	Male	2166	3059

Overall,

Model 1 performs better than Model 2

(perhaps to be expected, since it’s a richer model)

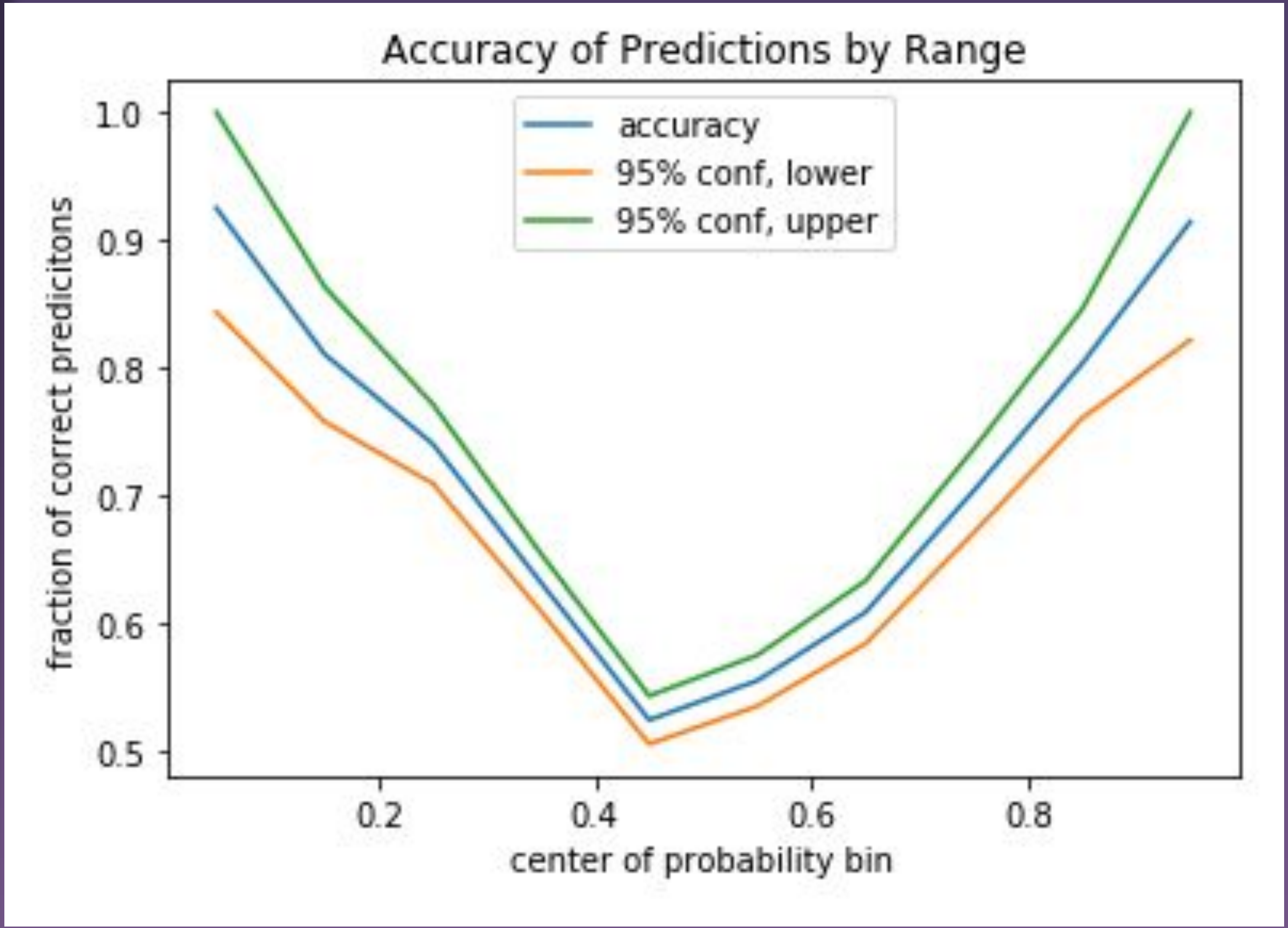
but

Model 2 likely encodes different information

and thus may be useful

to form a more complete picture

(Like Model 1) Model 2 also generalizes well



Model 3

Simple Pooled Embedding Model

Produces 400-dimensional tweet representation
meant to capture **word-level** differences in male-female diction

Model 3 Characteristics

- Bottom embedding layer initialized with Glove vectors
- Parallel max-pooling and average-pooling layers, to capture both highly gender-specific and typical gender-associated diction
- Only parameters are embedding weights and prediction coefficients
- Moderate (10% to 30%) dropout

Training Model 3

- Train with original (Glove) embeddings, then fine-tune embeddings
- In this case, **the fine tuning is the point**
- About 6 million embedding weights and 400 prediction coefficients
- Trains for 12 epochs initially and 25 for fine tuning

Model 3 Test Results

- Overall Accuracy 59.8%
- 0.647 ROC AUC

<table>
<tr><td colspan="2" rowspan="2">Overall Confusion Matrix</td><td colspan="2">Predicted</td></tr>
<tr><td>Female</td><td>Male</td></tr>
<tr><td rowspan="2">Actual</td><td>Female</td><td>3156</td><td>2069</td></tr>
<tr><td>Male</td><td>2130</td><td>3095</td></tr>
</table>

By itself

Model 3 performs worst of all 3

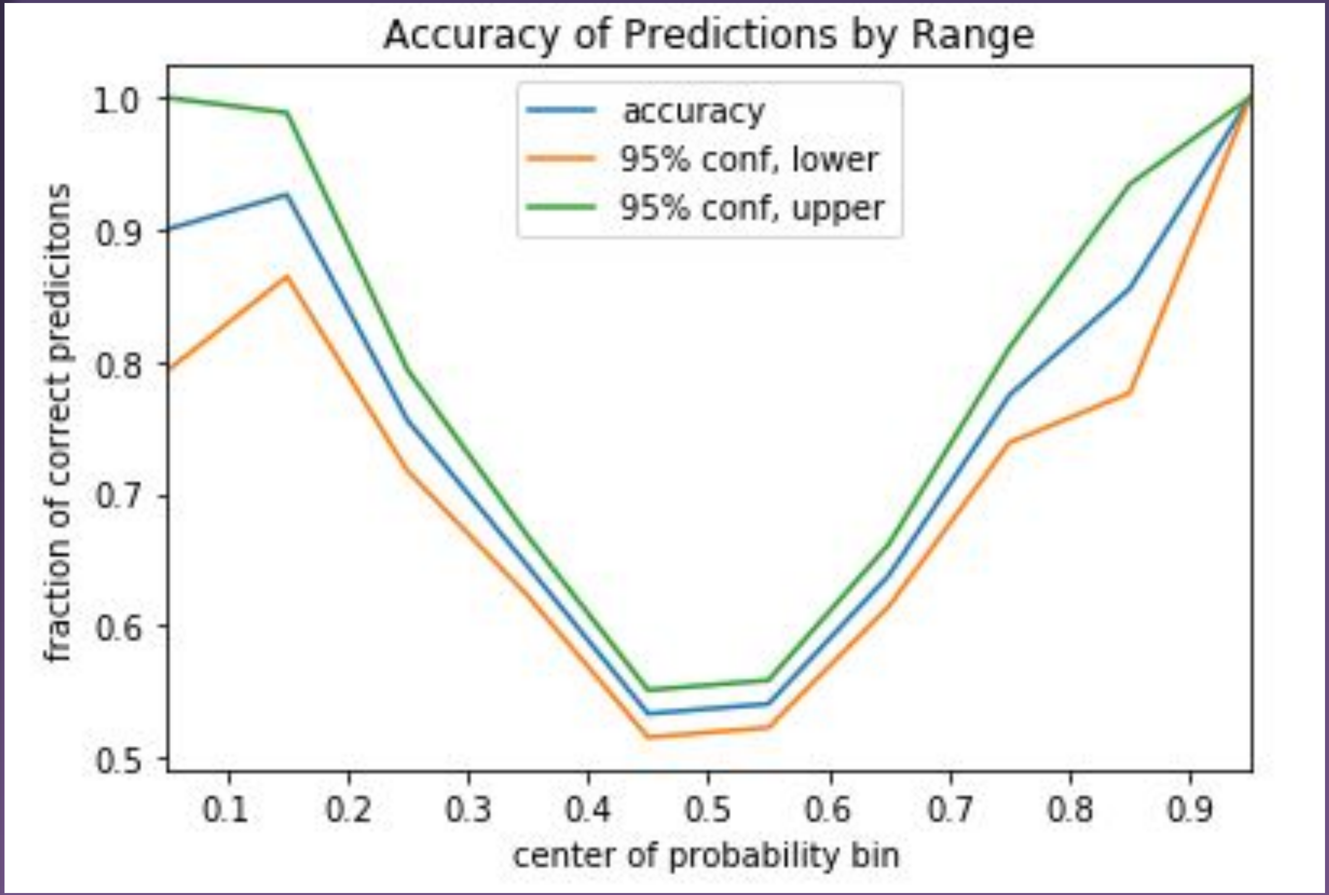
but this is definitely to be expected,
since it's intentionally a less rich model
designed specifically to encode different information
(word-level rather than tweet-level).

The hope is that

information from Model 3
will improve overall predictive capability

Model 3 also generalizes well…

but some of its predictions are "too good" (It understates the strength of its strongest predictions.)



Combined Model

PCA-Quadratic-Logistic Model

using all three sets of tweet embeddings

PCA-Quad-LR Model Characteristics

- Input is a set of embeddings and/or activations from hidden layer(s) of one or more underlying models
- Principal components generated from inputs (dimension determined by validation)
- Quadratic features generated from principal components
- L2-regularized logistic regression applied to quadratic features

PCA-Quad-LR Model Progression

1. Try with just Model 1 (USE-L) embeddings

2. Add activations from Model 2 (LSTM) and re-optimize (including a scaling factor to control relative importance of underlying models)

3. Add activations from Model 3 (Pooled) and re-optimize (including an additional scaling factor)

Results

- PCA-Quad-LR model slightly underperforms Model 1 on test set (ROCAUC=0.675 with USE-L embeddings alone as inputs)
- But it is easier to use for online learning and potentially easier to interpret
- Adding Model 2 activations produces a slight improvement (ROCAUC=0.677)
- Adding Model 3 activations does not affect test set performance

ONLINE LEARNING

DATA

- Random Tweets from 2019-05-21 thru 2019-07-06
- Selection and downsampling as with static data
- Split into batches by time (with batch size as an optimizable hyperparameter)
- Note that first few batches overlap with static Train/Valid/Test data

FITTING PROCEDURE

1. Fit PCA-Quad-LR model on first batch using stochastic gradient descent for a specified number of steps

2. Predict next batch and save predictions

3. Update fitted coefficients (using same number of SGD steps) on batch just predicted

4. Repeat 2 and 3 until end of data

HYPERPARAMETER OPTIMIZATION

1. Designate first 60% of data as “burn-in”, next 20% as “validation”, and last 20% as “test”
2. Fit to all data using described procedure
3. Repeat and choose hyperparameters based on accuracy of predictions for data designated as “validation”
4. Evaluate chosen model on “test” data

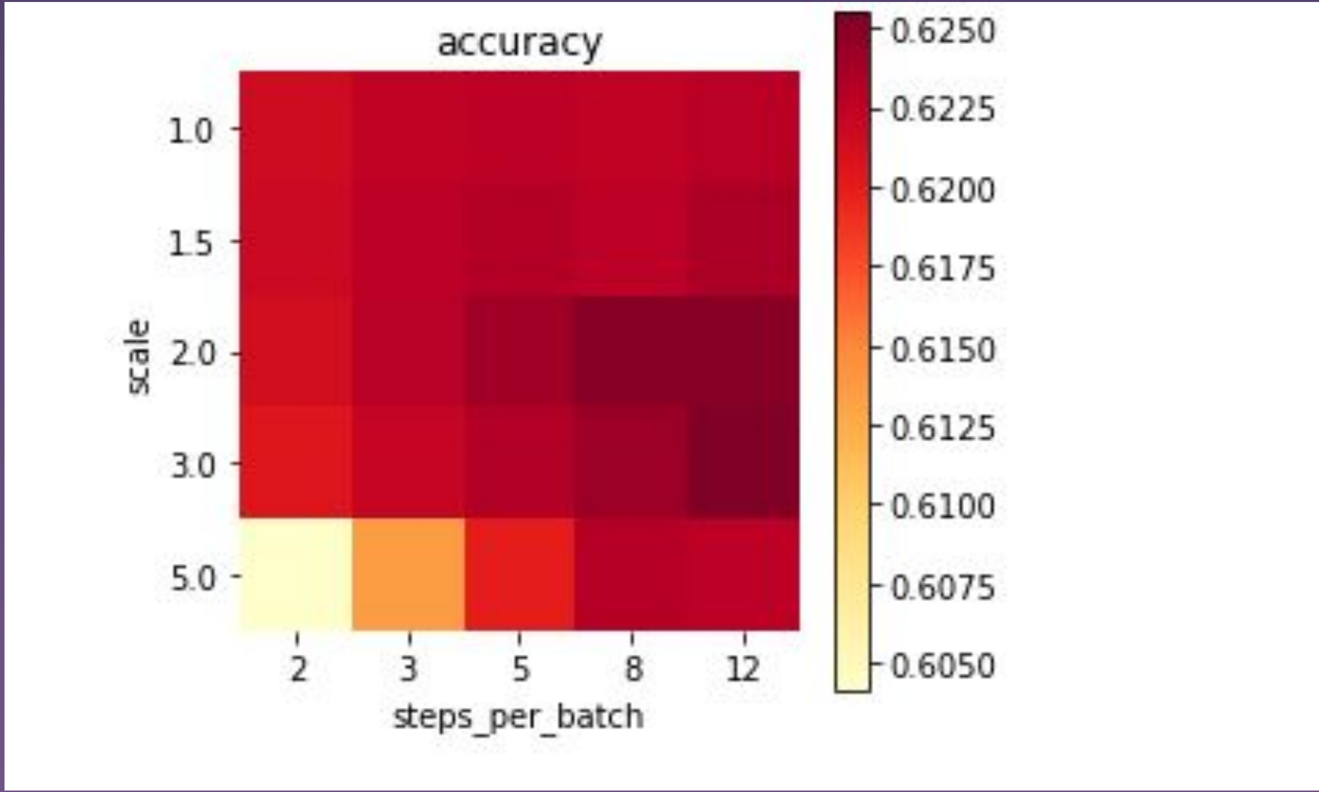
HYPERPARAMETER OPTIMIZATION

There are 4 hyperparameters:

- Scale factor applied to all inputs
- Batch size
- “Alpha” (SGD parameter that determines regularization and learning rate)
- Number of steps per batch

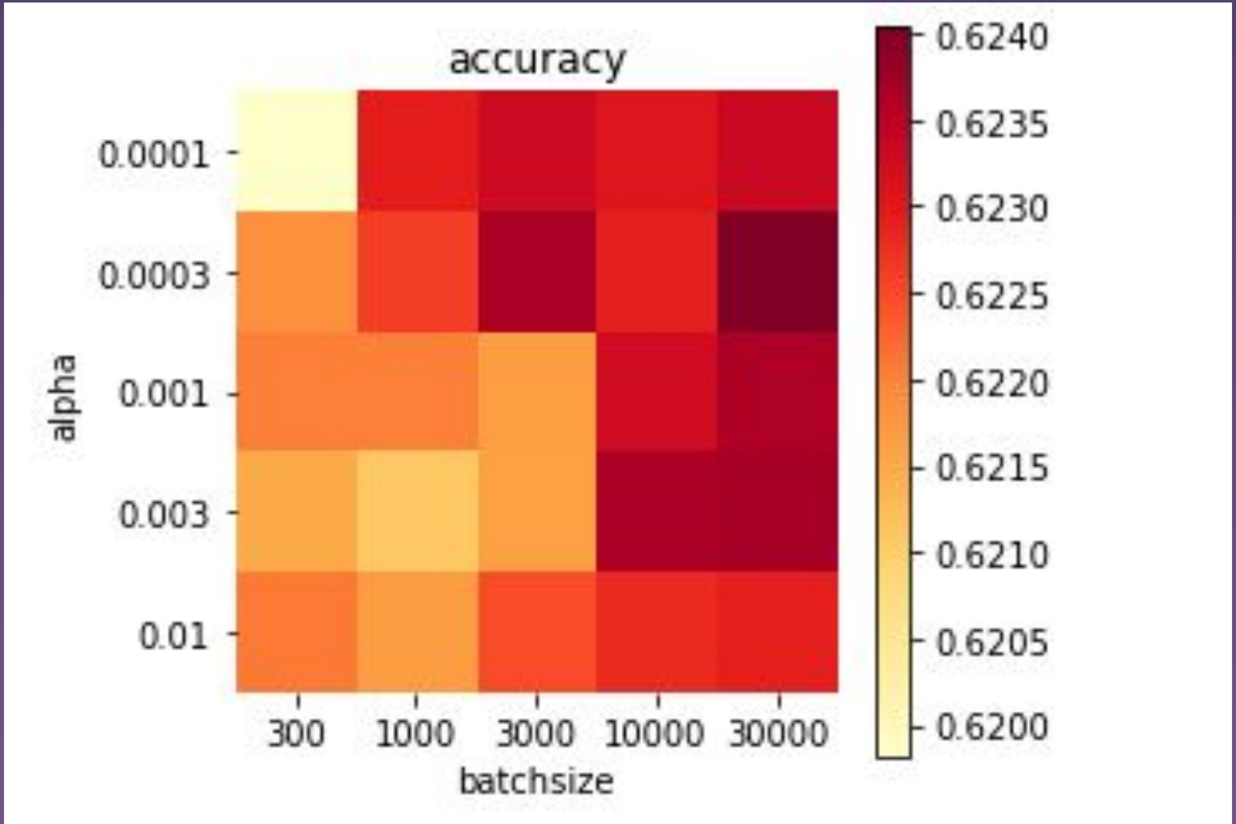
Example from Optimization:

Validation set accuracy for values of scale and steps-per-batch with batch size and alpha held constant



Example from Optimization:

Validation set accuracy for values of batch size and alpha with scale and steps-per-batch held constant



FINAL ONLINE LEARNING MODEL

Hyperparameters:

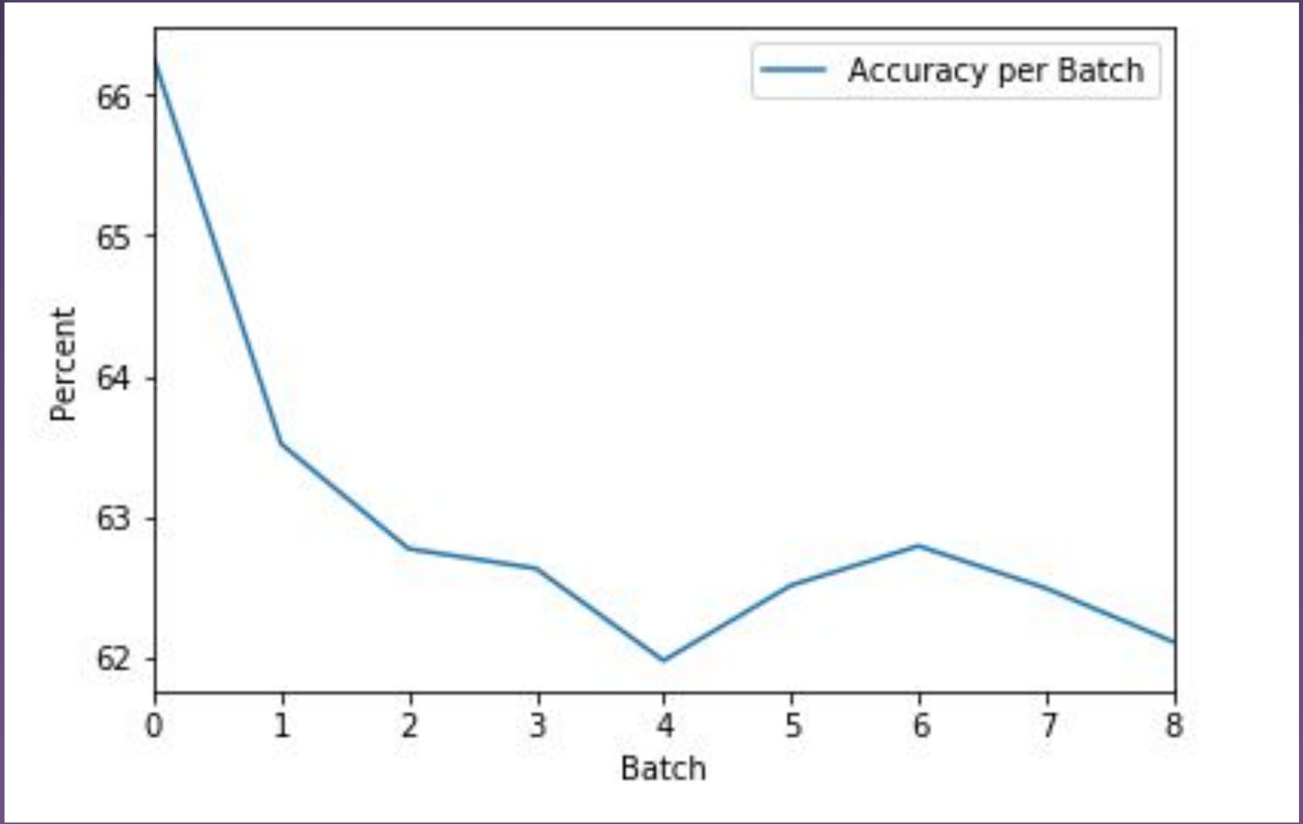
- Scale = 3.5
- Batch size = 20000
- Steps-per-batch = 20
- Alpha = 0.0004

Test set performance:

- Accuracy = 62.3%
- ROC AUC = 0.676

Large chosen batch size (20000) and steps-per-batch (20) imply there is not much “online” aspect to this online learning: it is largely just re-fitting the whole model to a moving window

Performance of Online Learning Model Over Time



Salient Points:

- High accuracy on first 2 batches because they are where the embeddings were trained and optimized
- After that, accuracy is fairly stable, with no improving or deteriorating trend
- Good that there is no improving trend, because it means there are enough data for an optimal fit
- Good that there is no deteriorating trend, because it means the underlying embeddings will not have to be re-fit often

Recommendations:

- Keep an eye on performance (extending the previous chart into the future as new data come online), and re-fit the underlying models when it starts to decline
- Evaluate cost-effectiveness of alternative approaches using just Model 1: Either do online learning with the deep learning model alone or just keep refitting it
- If budget permits, consider building a multi-branch deep learning model that includes all three underlying models

ADDENDUM: ATTEMPTS AT INTERPRETATION

Tweets with extreme prediction scores

Most “Masculine”

Now we have football instead. Proof that the old days were mostly terrible.

8.8 tho, according to the match stats.

At the same that Betway and West Ham strike a record deal.

getting a bit tempestuous on the pitch STK v CAR end of 3rd...

Almora's diving catch in front of him to end top 2nd (robbing Dietrich of a hit) was a 3-Star grab per Statcast. Co…

Most “Feminine”

Seeing so happy and playing with makeup makes my heart happy... now I want to play with my makeup this week 💙💙

So good I had to share! Check out all the items I'm loving on from @alwaysmorefinds …

Check out what I just added to my closet on Poshmark: Social Butterfly BFYHC. via

I cannot get over how grown many of my babies are!!! Ayodele, Itzel, Caitlin, Sheena,…

I bought an apron that has little strawberries and I’ve never been so excited!

Broadly, extreme predictions are consistent with stereotypes:

- Men tweet about sports
- Women tweet about clothes, makeup, shopping, and emotions

Tweets scoring high on principal components that predict male or female

Most “masculine” component: “Argumentativeness”

- Sure, but apparently you dont see the other issues that brings up. Such as the measu…
- Except without a compelling reason to do so.'
- On what "grounds"? Another NOTHINGBURGER!
- that's such bs! and comes from a spineless complicit person who supports the
- And it shouldn’t be avoided for political expediency either.
- What a BS headline. Either you know that and are being dishonest or your incredibly stupid. Or both.\nProbably both.
- Very simplistic. But. Ok.' ' No shit.' '   No shit!
- Bit tricky when he supports it 🤷🏼‍♀️

Most “feminine” component: “Negative Interpersonal Emotions”

- Like how can you lie to my face alllll the time . You must have no self respect for the things your doing
- I hate that kind of individual. The make life so complicated
- But I have that little part of me being like, “do they think I’m a fuckin weirdo”\nThough I know that’s not totally the case.
- Hate people like this who say ‘if you actually watched’.\n\nIf you had any idea who you are talking to, you would kno…
- I wish I can tell ppl to stop loving me and just go on with their life , is that selfish?🤔😔
- Really frustrates me how someone can be horrible to an animal x
- I never give up on people that i love, but if i do, know that you really messed up.
- No, that’s immaturity. I might not express I think he’s cute again but you don’t automatically stop liking someone… '

Miscellaneous Thought:

This project was about a way to target marketing efforts, but as it turns out, marketing targets may be more the cause than the effect. References to Poshmark, for example, occur overwhelmingly in tweets with near-100% probability of being from women.

Anyhow...

I GOTTA GO...

Thanks to SlidesCarnival for the template.