Teammate Names: John Doe

Comparing Models and Datasets for Classifying Misinformation

Ashir Rao’s Big Data ML Capstone project

Classifying Misinformation

Misinformation, promoted by bad actors, is a huge problem in society.

There has been a lot of work already in using ML to solve this problem.

I wanted to see what type of data and what type of models are best for building the best model to solve this problem.

Roadmap

The Data:

Dataset 1: LIAR

- From ACM
- More features
- Shorter text

Dataset 2: ISOT

- UVictoria
- Less features
- Longer Text

LIAR

entry point for launching an IPython kernel.

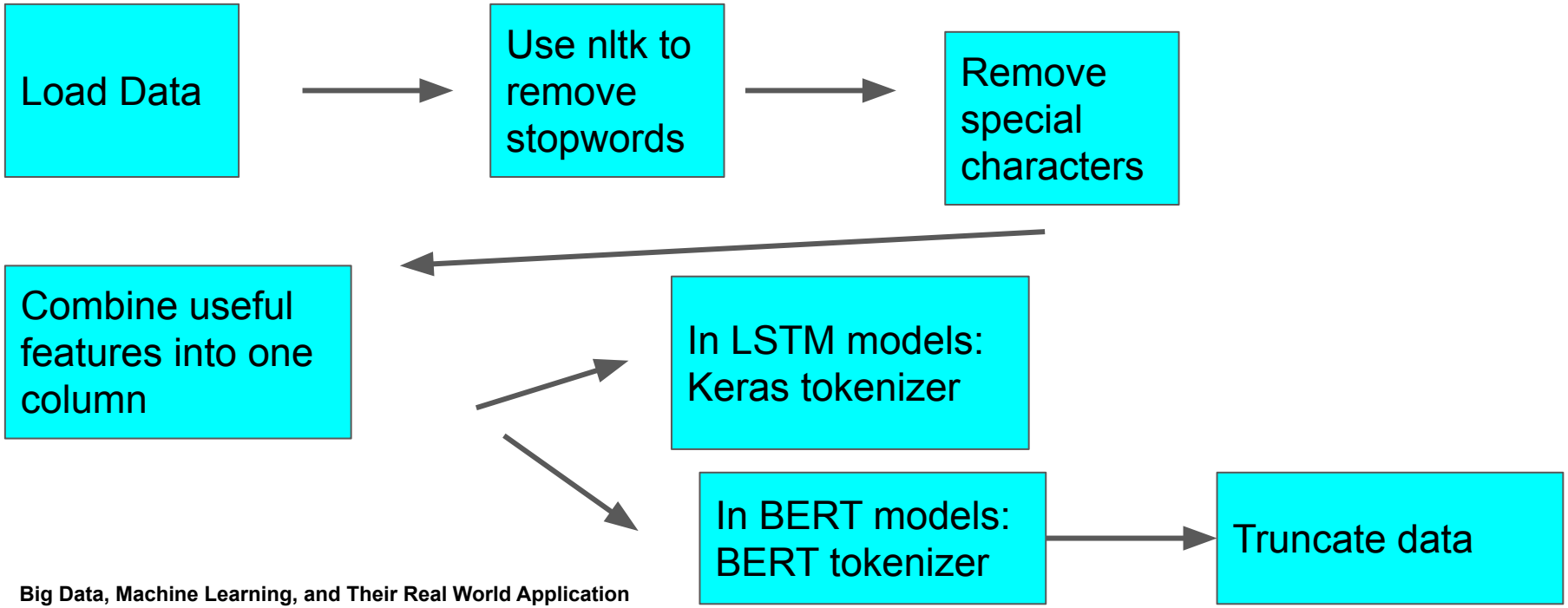
	file	valid	text	topic	person	position	state	party	a	b	c	d	e	context
1	7891.json	0	campaign finance congress taxes earl blumenau...	campaign finance congress taxes	earl blumenauer	u representative	oregon	democrat	0.0	1.0	1.0	1.0	0.0	u ways means hearing
2	8169.json	1	poverty jim francesconi member state board h...	poverty	jim francesconi	member state board higher education	oregon	none	0.0	1.0	1.0	1.0	0.0	opinion article
3	929.json	1	economy stimulus barack obama president ill...	economy stimulus	barack obama	president	illinois	democrat	70.0	71.0	160.0	163.0	9.0	interview cbs news
4	9416.json	0	guns jim rubens small business owner new ha...	guns	jim rubens	small business owner	new hampshire	republican	1.0	1.0	0.0	1.0	0.0	interview gun shop hudson n h
5	6861.json	1	education state budget andy berke lawyer sta...	education state budget	andy berke	lawyer state senator	tennessee	democrat	0.0	0.0	0.0	0.0	0.0	letter state senate education committee chairw...
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
4440	10405.json	0	energy job accomplishments bobby scott u con...	energy job accomplishments	bobby scott	u congressman	virginia	democrat	1.0	2.0	1.0	5.0	0.0	news release
4441	5479.json	0	health care message machine 2012 voting	health care message machine 2012 voting	marcy	representative ohio's ninth congressional	ohio	democrat	0.0	2.0	4.0	4.0	0.0	campaign

ISOT

	title	text	subject	date	class
0	As U.S. budget fight looms, Republicans flip t...	us budget fight looms republicans flip fiscal ...	politicsNews	December 31, 2017	1
1	U.S. military to accept transgender recruits o...	us military accept transgender recruits monday...	politicsNews	December 29, 2017	1
2	Senior U.S. Republican senator: 'Let Mr. Muell...	senior us republican senator 'let mr mueller j...	politicsNews	December 31, 2017	1
3	FBI Russia probe helped by Australian diplomat...	fbi russia probe helped australian diplomat ti...	politicsNews	December 30, 2017	1
4	Trump wants Postal Service to charge 'much mor...	trump wants postal service charge 'much more' ...	politicsNews	December 29, 2017	1
...	...	...	...	...	...
23476	McPain: John McCain Furious That Iran Treated ...	mcpain john mccain furious iran treated us sai...	Middle-east	January 16, 2016	0
23477	JUSTICE? Yahoo Settles E-mail Privacy Class-ac...	justice yahoo settles email privacy classactio...	Middle-east	January 16, 2016	0
23478	Sunnistan: US and Allied ‘Safe Zone’ Plan to T...	sunnistan us allied safe zone’ plan take terri...	Middle-east	January 15, 2016	0
23479	How to Blow $700 Million: Al Jazeera America F...	blow 700 million al jazeera america finally ca...	Middle-east	January 14, 2016	0
23480	10 U.S. Navy Sailors Held by Iranian Military ...	10 us navy sailors held iranian military – sig...	Middle-east	January 12, 2016	0

44898 rows × 5 columns

Preprocessing - similar in both datasets



4 Models: dataset + model

Liar + LSTM	Liar + BERT
ISOT + LSTM	ISOT + BERT

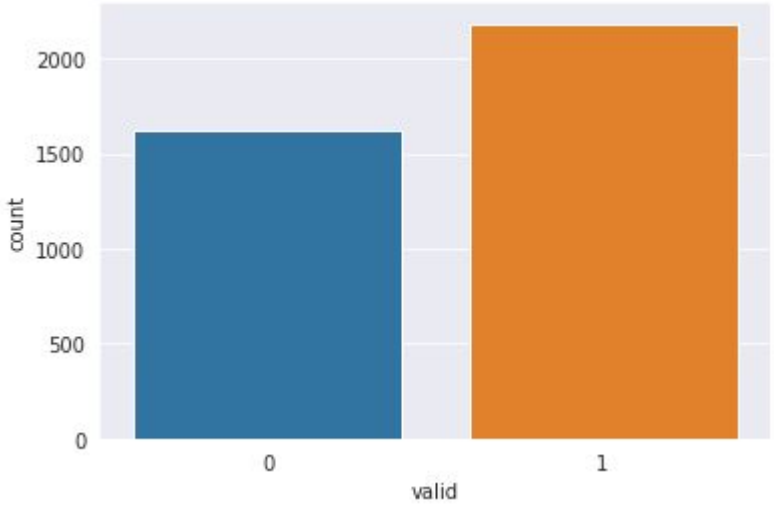
Accuracy using validation-split

```
- val_loss: 0.7032 - val_accuracy: 0.5315
- val_loss: 0.8704 - val_accuracy: 0.5734
- val_loss: 0.8963 - val_accuracy: 0.5664
- val_loss: 1.3574 - val_accuracy: 0.6119
- val_loss: 1.5509 - val_accuracy: 0.5734
- val_loss: 1.7973 - val_accuracy: 0.5699
- val_loss: 2.0666 - val_accuracy: 0.5524
- val_loss: 2.3114 - val_accuracy: 0.5839
- val_loss: 2.2232 - val_accuracy: 0.5874
- val_loss: 2.4238 - val_accuracy: 0.5909
```

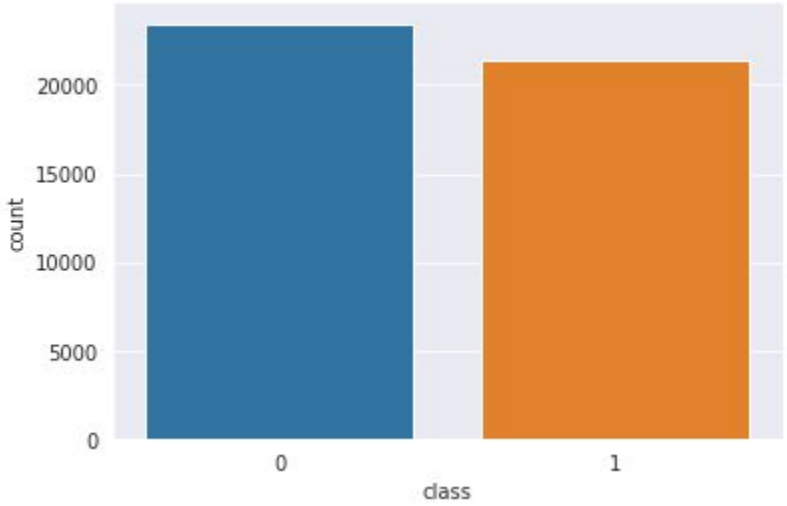
←gives this while fitting

Quick Data Viz:

LIAR (after labeling classes)



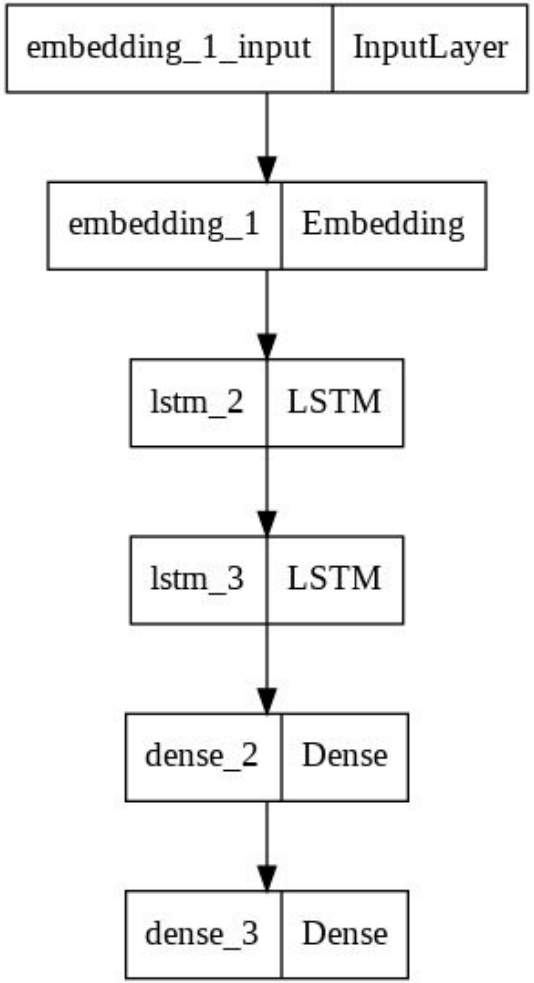
ISOT



Model Construction - LSTM

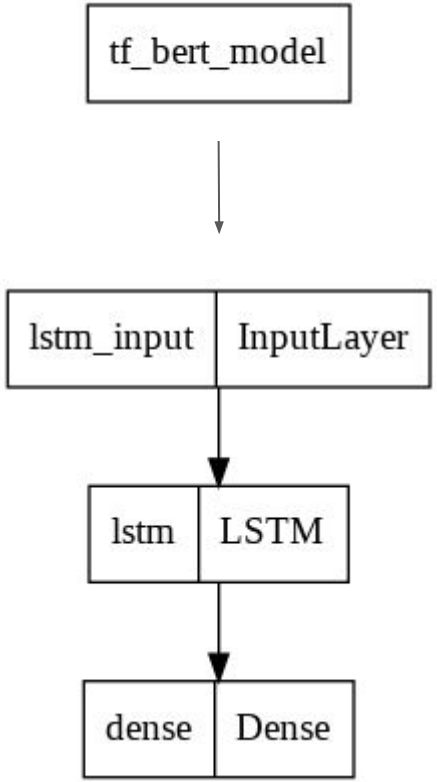
```
model = tf.keras.models.Sequential()

model.add(Embedding(10000, output_dim = embed_size, input_length=1000))
model.add(LSTM(units = 128, activation = 'relu', recurrent_activation = 'tanh', return_sequences = True, recurrent_dropo
model.add(LSTM(units = 64,activation = 'relu', recurrent_activation = 'tanh', recurrent_dropout = 0, dropout = 0.1))
model.add(Dense(units = 32, activation = 'relu'))
model.add(Dense(1, activation = 'sigmoid'))
model.compile(optimizer = keras.optimizers.Adam(learning_rate = 0.01), loss = tf.keras.losses.binary_crossentropy, metri
```



Model Construction - BERT

```
import numpy as np
output = [np.array(model(input_ids[i])['last_hidden_state']) for i in range(2000)]
output = np.asarray(output)
output.shape
```



Performance

1. ISOT + LSTM - 96.49% val acc, 96.31% test acc
   (Baseline model - 99.2%)
2. LIAR + LSTM - 60.44% val acc
3. LIAR + BERT - 46.85% val acc
4. ISOT + BERT - 0% val acc

Progression of tuning

ISOT + LSTM

69% -> 80.43 -> 96.49%
By not using sigmoid at the beginning -> Less epochs

LIAR + LSTM

54.45% -> 60.44%
By not using sigmoid at the beginning

Limitations

- Used the same LSTM model on LIAR and ISOT datasets
  - A different model may have worked better on LIAR
- Not enough RAM to train BERT model on all 44,000 articles in ISOT
  - May have led to BERT underperforming
  - Had to truncate the data to a fraction of what it originally was
- BERT output was thrown directly into an LSTM model
- LIAR dataset had several blanks
- LIAR dataset was harder to discern -- it would be tough for even a human

Conclusions

- Best model was ISOT + LSTM
- ISOT performed 160.3% better than LIAR with same model
- It is necessary to use other approaches for shorter data, regardless of whether it has more features
- Performance of BERT inconclusive

Personal Conclusions

- ML takes time, to learn and to train
- Not everything is in your control when making a model
  - Almost like Wizardry (oooh)
- Cleaning data is a lot, if not most, of the job
- Lots of experimentation is needed to get a good model
- Colab Pro exists for a reason

…And that’s okay!