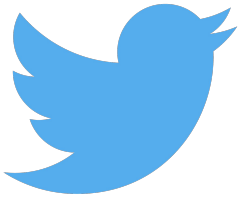
Sentiment Analysis
Of Coronavirus Tweets

By: Masoud Faramarzi, Maryam Kafi Kang, and Lily Sisouvong



Presentation Overview

Section 01:

NLP, Dataset, Preprocessing and Transfer learning

Section 02:

Language and Classifier Models, Hyperparameters Optimization

Section 03:

Results, Conclusion and Recommendations

Project Objective

Train a model that determines if a tweet has a negative, neutral or positive sentiment regarding the Covid-10 vaccination.

Positive

Neutral

Negative

Natural Language Processing(NLP)

Natural language processing strives to build machines that understand and respond to text or voice data and respond with text or speech of their own in much the same way humans do.

Universal Language Model Fine-Tuning

“Transfer Learning”

- Pretrained model
- Trained on Wikipedia articles
- Is able to predict a given number of words

- Adjusts the prior model using tweets from dataset
- Is able to predict tweet specific words

- Uses the encoding from the language model stage
- Be able to classify sentiments

Datasets

- General Tweet Dataset (with sentiments) - 40,000 tweets
  - Used in both language model and classifier training
- Vaccination Tweet Dataset (without sentiment) - 212,982 tweets
  - Used in language model training and later in prediction phase
  - Pfizer/BioNTech, Sinopharm, Sinovac, Moderna, Oxford/AstraZeneca, Covexin, Sputnik V

Preprocessing the data

Handles

@tiffanylue i know i was listenin to bad habit earlier and i started freakin at his part =[
Layin n bed with a headache ughhhh...waitin on your call...
Funeral ceremony...gloomy friday...
wants to hang out with friends SOON!
@dannycastillo We want to trade with someone who has Houston tickets, but no one will.
Re-pinging @ghostridah14: why didn't you go to prom? BC my bf didn't like my friends
I should be sleep, but im not! thinking about an old friend who I want. but he's married now. damn, & he wants me 2! scandalous!
Hmmm. http://www.djhero.com/ is down

not included in the example

Emojis

Blanks

Hashtags

URLS

Fastai Preprocessing Steps

Tokenization

- Splitting a large sample of text into words

```
["This", "is", "a", "text"]
```

- Added functionality to create special tokens “xxbos” and “xxmaj”

xxbos xxmaj pfizers experimental covid-19

Numericalization

- Create a vocabulary
- Convert tokens to integers using index

```
['xxunk', 'xxpad', 'xxbos', 'xxfld',
'xxmaj', 'xxup', 'xxrep', xxwrep', 'the',
'.', ',', 'a', 'and', 'of', 'to', 'is',
'it', 'in', i', 'this']
```

Fine tune the Language Model

- Pre-trained **RNN** model, **AWD_LSTM,** which is built into **fastai**
- Fine Tune the model with our dataset
- Goal
  - Extracting the numerical features from data
  - Predict next words of text

The classification Model

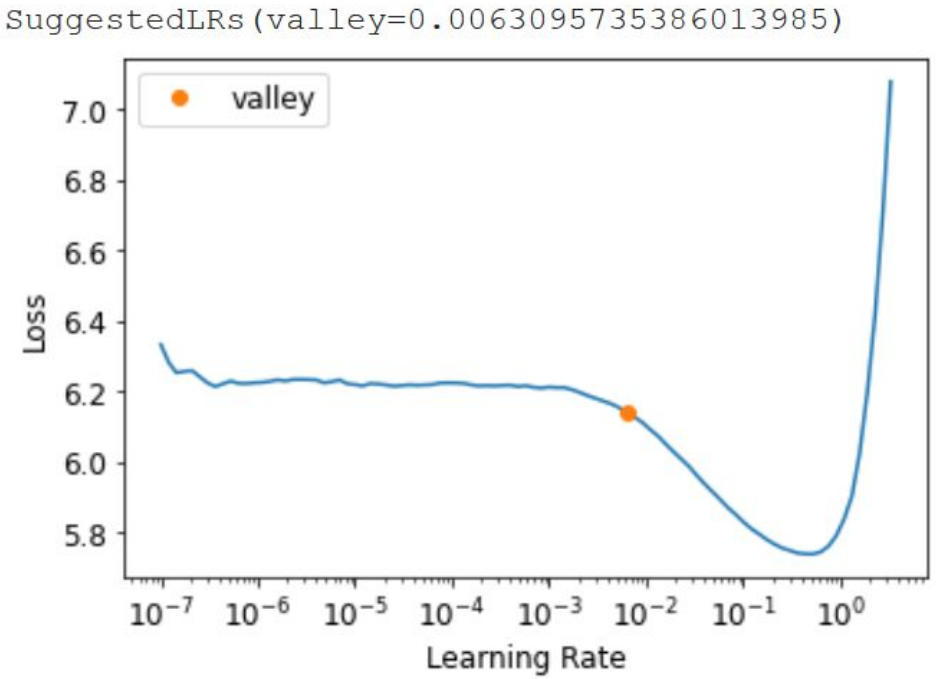
- **RNN** model, **AWD_LSTM,** which is built into **fastai**
- Using the fine tuned Language model(except the last layer) to update the weights of classifier
- Goal
  - Predict the sentiment of each sentence

Probabilistic Sentiment Classification of unseen tweets

- A Positive sentence
  - learn.predict("While the world has been on the wrong side of history this year, hopefully, the biggest vaccination effort we've have")
  - Output : ('positive', tensor(2), tensor([0.0472, 0.2188, 0.7341]))
    - the first sentence is negative with the probability of 4%
    - the first sentence is neutral with the probability of 21%
    - the first sentence is positive with the probability of 73%

- A Negative sentence
  - learn.predict(" I have just been in a bad mood for 40 years! ")
  - Output : ('negative', tensor(0), tensor([0.8558, 0.1184, 0.0258]))
    - the first sentence is negative with the probability of 85%
    - the first sentence is neutral with the probability of 11%
    - the first sentence is positive with the probability of 2%

Optimum value for learning rate in Language/ classifier Models



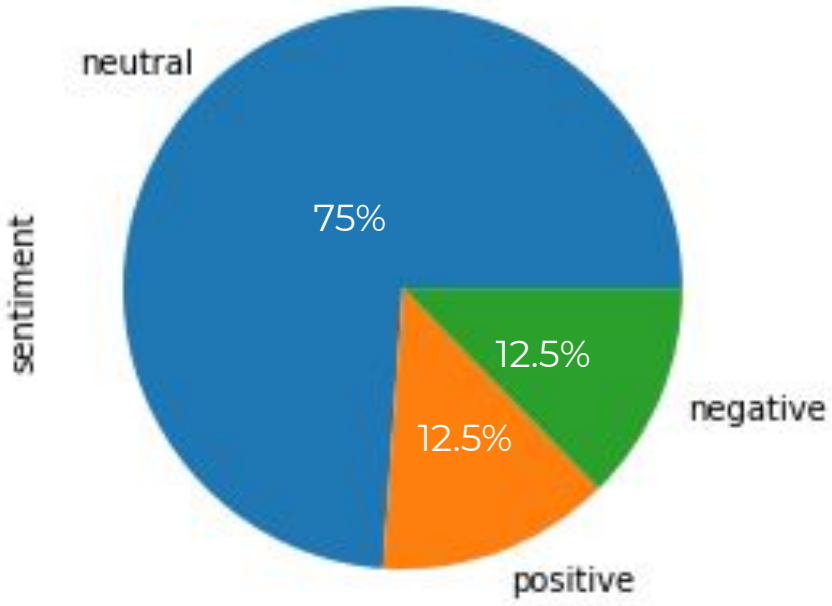
Hyperparameters for Language Model

Accuracy	Learning Rate	Dropout Multiplier
41%	0.0025	0.30
44%	0.1	0.30
44.5%	0.03	0.30
45.6%	0.3	0.50
46.3%	0.3	0.15

Hyperparameters for Classifier

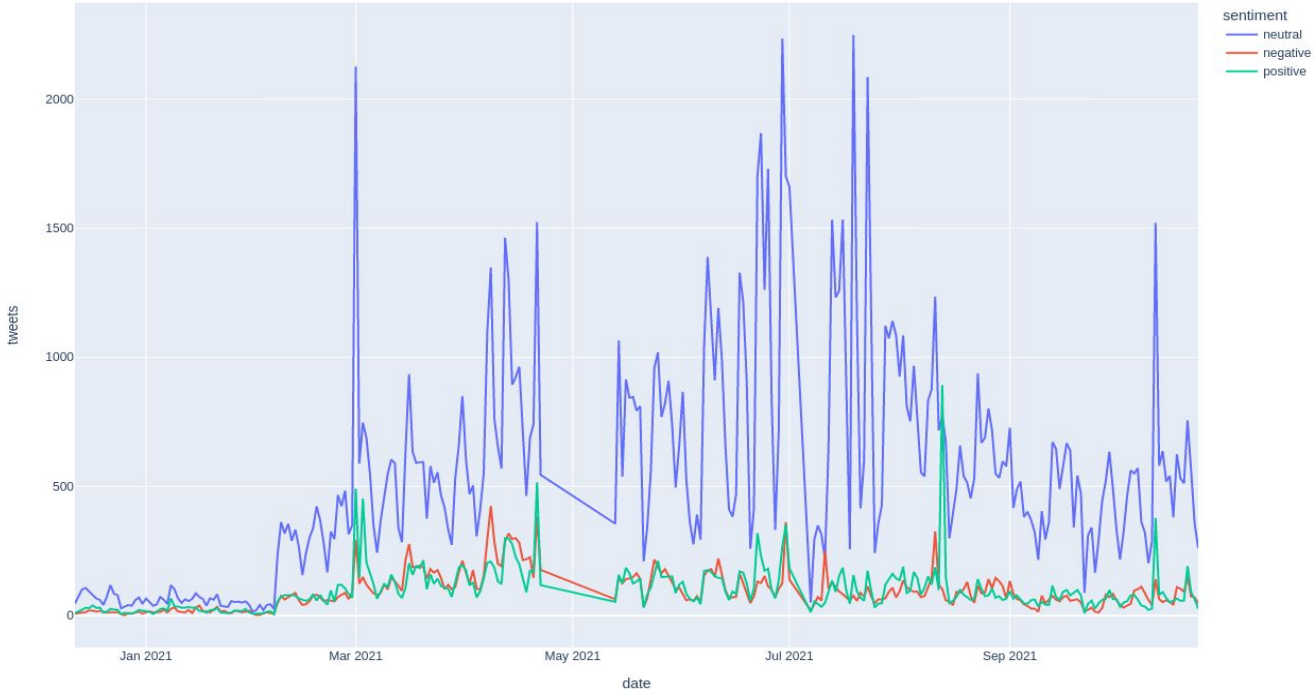
Accuracy	Learning Rate	Dropout Multiplier
73.70	0.006	0.20
74.72	0.006	0.60
75.00	0.001	0.30
76.10	0.006	0.30
76.20	0.03	0.3

Predicting the sentiment of vaccination tweets

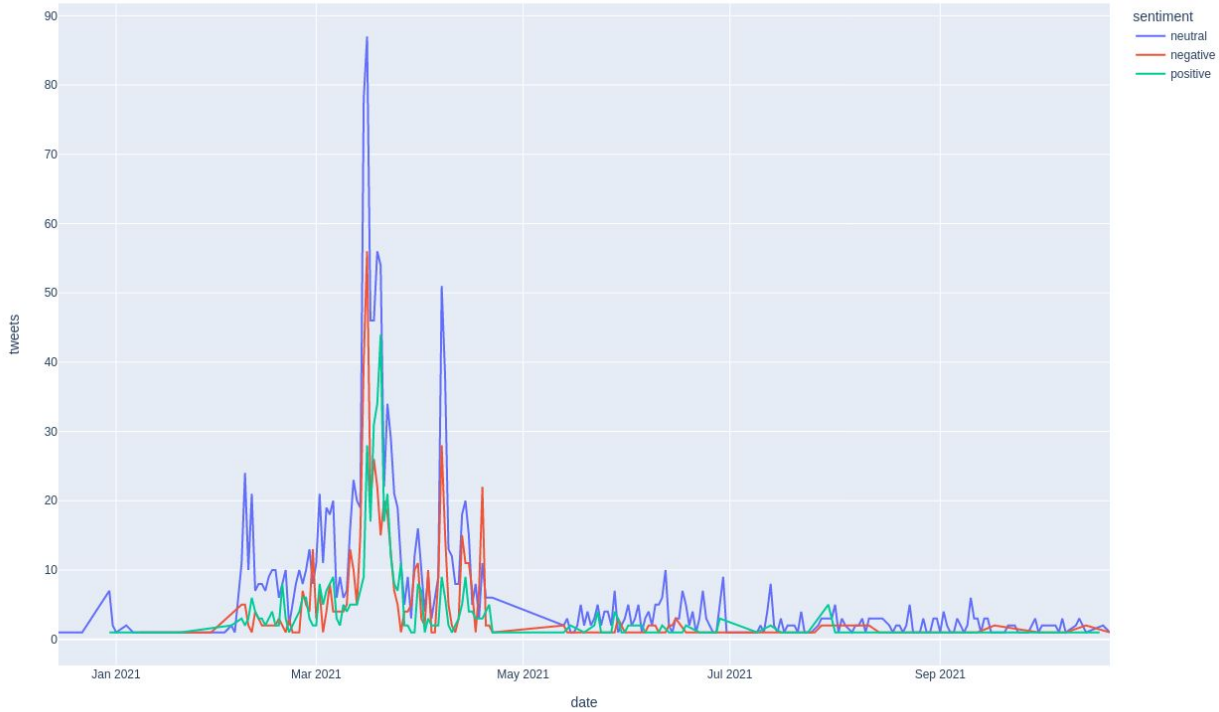


Visualizations

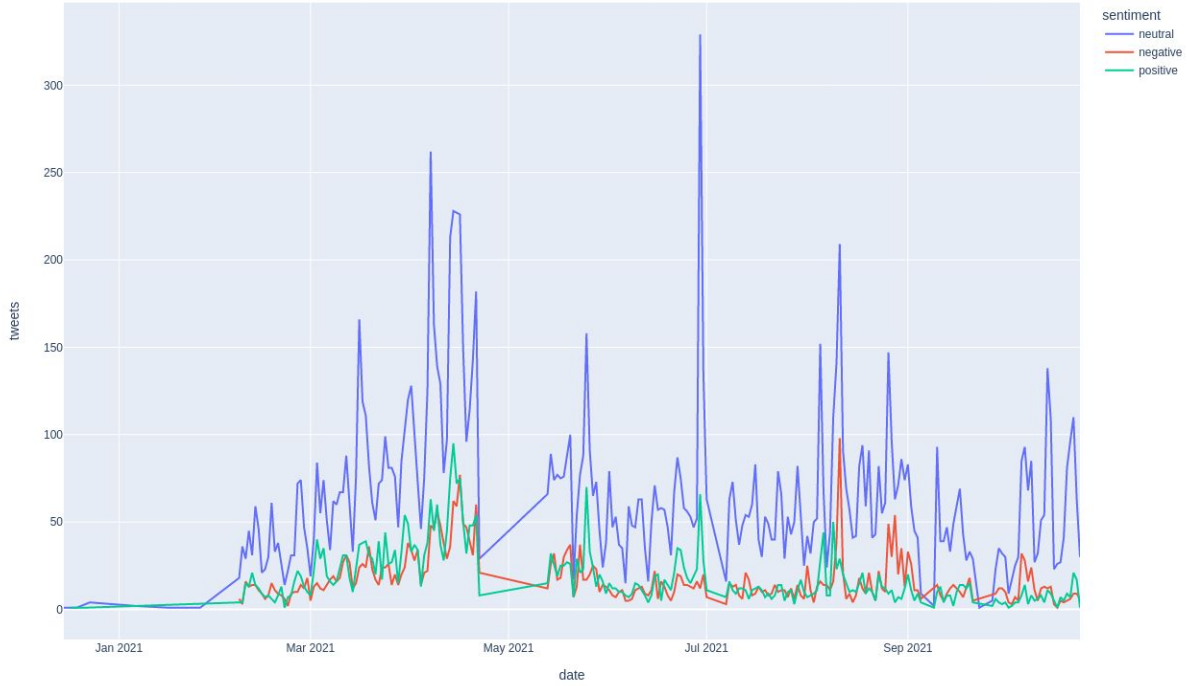
Timeline showing sentiments of Tweets About Covid-19 Vaccines



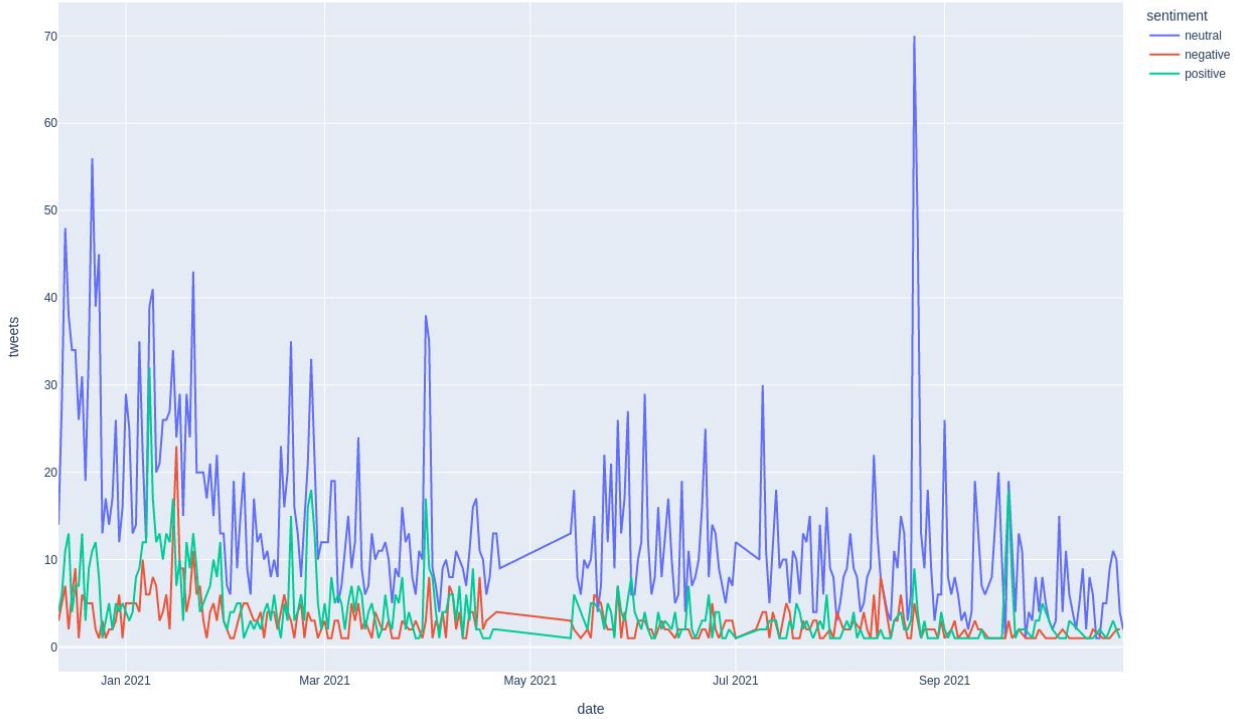
Timeline showing sentiments of Tweets About Oxford/AstraZeneca Vaccines



Timeline showing sentiments of Tweets About Moderna Vaccines



Timeline showing sentiments of Tweets About PFizer Vaccines



Conclusion

Optimum Hyperparameters

Function	Learning Rate	Dropout Multiplier	Model Algorithm
Language model	0.03	0.15	AWD-LSTM
Classifier	0.03	0.3	AWD-LSTM

Accuracies

Language Model

- Highest accuracy: 45.6%

Classifier

- Highest accuracy: 76.2%

Recommendations

Trying multiple training algorithms other than default AWD-LSTM (e.g. a config of AWD-LSTM with different number of layers and neurons)

Training classifier using labeled vaccination dataset

Thank you!

Any questions?

References

- fastai_NLP, retrieved Dec 19 2021 from:https://github.com/fastai/fastbook/blob/master/10_nlp.ipynb
- Fastai learning rate, retrieved Dec 19 2021 from: https://fastai1.fast.ai/callbacks.lr_finder.html
- fastai, retrieved Dec 19 2021 from: https://docs.fast.ai/
- GitHub,fastai, retrieved Dec 19 2021 from: https://github.com/fastai/fastbook/blob/master/10_nlp.ipynb