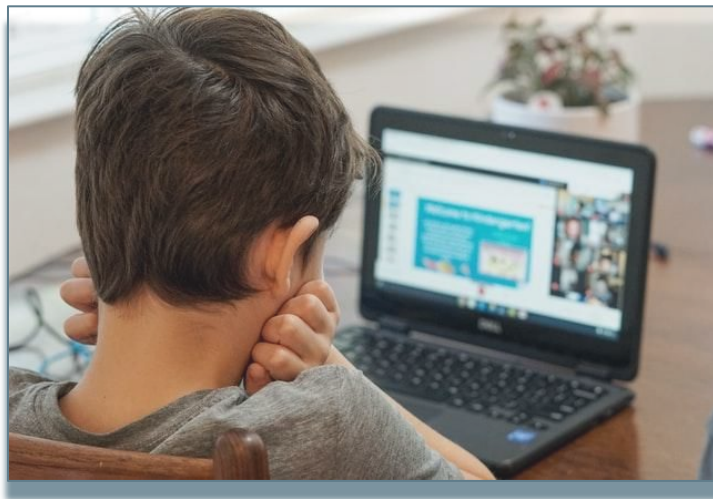


# Capstone Project

## Leah Pope

### *Learning During COVID-19 Sentiment Classifier*

81720-ONL-DS-FT  
01/26/2021



# Hello!

## I'm Leah Pope

<https://github.com/lspope>

[leah@metisconsultingllc.com](mailto:leah@metisconsultingllc.com)

<https://leahspope7.medium.com>

<https://www.linkedin.com/in/leahspope>





“

**This project explores the question...**

*"What is the public sentiment in the United States on K-12 learning during the COVID-19 pandemic?"*

# Methodology

- ◎ 30,599 Tweets from Users across the US
  - Collected via Twitter Stream & Queries
  - Kaggle [Tweets about Distance Learning](#)
  - Filtered out non-US Tweets
- ◎ Label sentiment - Positive, Negative, Neutral
  - Human & Sentiment Tools
  - Use sentiment where Tools agree



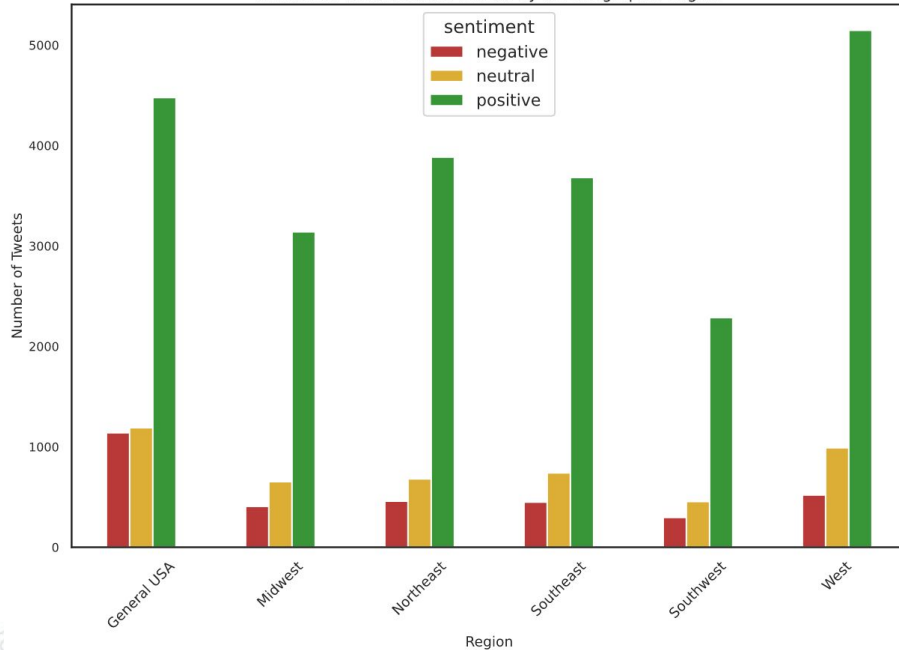
# Methodology (cont.)

- ◎ Discover trends using EDA
  - Topic Modeling using Unsupervised Machine Learning to organize Pos/Neg/Neutral Tweets into topics
- ◎ Create multiclass (Pos/Neg/Neutral) classifier
  - Natural Language Processing
  - Supervised Machine Learning



# Exploratory Data Analysis

Sentiment Breakdown of Tweets by US Geographic Region



- Vast majority of Tweets have Positive sentiment
- True for every Region and every State in every Region
- Suggests that Twitter being used to communicate positive statements on K-12 Ed during COVID
- Explore the Topic Models for more context on Tweet content

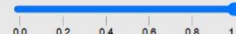
# Topic Modeling

## US Topics

```
In [4]: # model topics in all the Tweets
usa_tm_tuple = perform_topic_modeling(us_tweets, , 30)
pyLDAvis.sklearn.prepare(usa_tm_tuple[0], usa_tm_tuple[1], usa_tm_tuple[2])
```

Out[4]: Selected Topic:  Previous Topic Next Topic Clear Topic

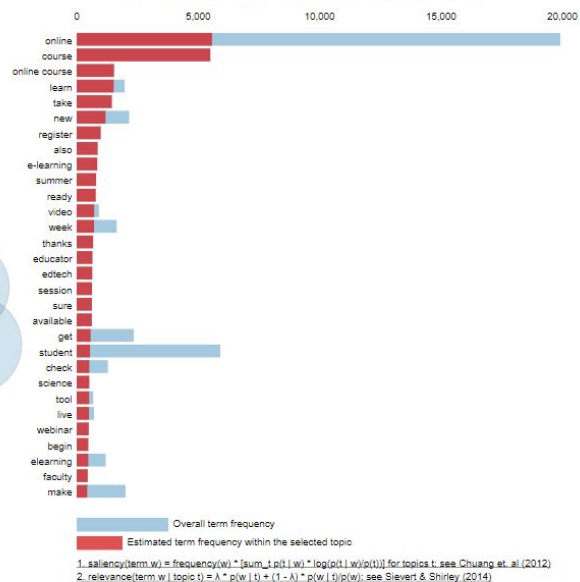
Slide to adjust relevance metric:<sup>(2)</sup>  
 $\lambda = 1$



Intertopic Distance Map (via multidimensional scaling)



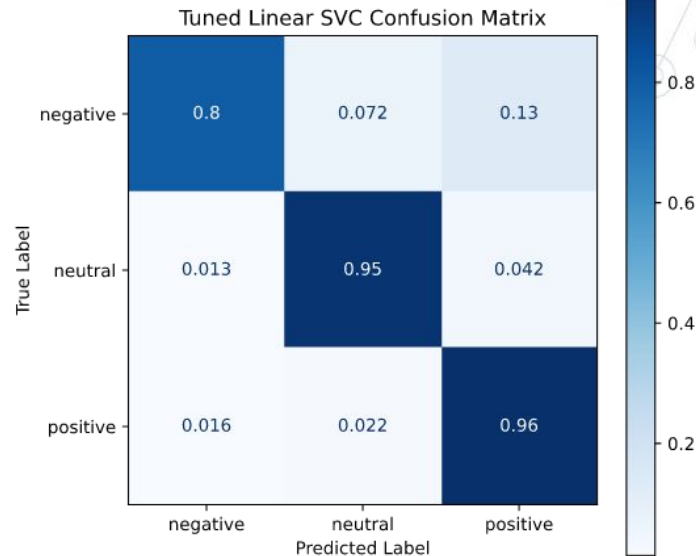
Top-30 Most Relevant Terms for Topic 4 (12.7% of tokens)



Explore Topics for yourself [here!](#)





# Classifier

- ◎ Good performance achieved!
  - Weighted F1 Score **94%**
  - Balances Precision & Recall
- ◎ *Confusion Matrix* shows correct (and incorrect) predictions
  - **Negative:** correctly predicted **80%**
  - **Neutral:** correctly predicted **95%**
  - **Positive:** correctly predicted **96%**





# Classifier Predictions

@jac___13 Sorry Jackie! I've kept our kids fully remote, even opting out of the hybrid model. I just can't trust other outside of my bubble right now. At least you guys in K-12 are in the earlier phases to get vaccinated. My wife in higher-Ed is not. I truly feel for educators right now.	
Parents! 🗣️ @DaytonLive365 is offering a free "virtual field trip" for local K-12 students from Feb. 1 to March 14. Register here for access to stream Grammy-nominated duo, Black Violin: <a href="https://t.co/XR29EKm2TT">https://t.co/XR29EKm2TT</a> <a href="https://t.co/mCb6W2vUZp">https://t.co/mCb6W2vUZp</a>	
Screw the open schools COVID movement for damaging the term open schools. An Open School is a K-12 self-directed democratic school. JCOS in Denver Colorado is the last Open School in USA. Here's an EP we released w/ a former student & teacher. <a href="https://t.co/3hJ2sGwMLD">https://t.co/3hJ2sGwMLD</a> <a href="https://t.co/jblfpO80UP">https://t.co/jblfpO80UP</a>	
There are two parts to the bill; the second would require districts to offer a 5 day per week, in-person option for all students, but only *after* the completion of vaccination phase 1b. K-12 teachers and staff are included in SDMAC's 1b plan.	

# Recommendations

## For Education-focused Stakeholders:

EDA shows that Twitter is being used in the US to communicate primarily **Positive** statements on K-12 Ed during COVID

- ◎ Recommend Stakeholders explore the Topic Modeling analysis of these Positive Tweets for additional insight into content

EDA shows that West and Southwest Regions have a 'vocal' state, far exceeding others in # of Tweets. 'Vocal' states *are* the most populous in their Region **but** is it reasonable to have such a large lead?

- ◎ Recommend further Region/State population analysis to check for over-representation



# Recommendations

## For Data Scientists building upon this work:

The collected data has a class imbalance issue with 73.9% of the data labeled as Positive, 10.7% Negative, and 15.4% Neutral

- ⦿ Recommend hyperparameter tuning uses the 'balanced' class\_weight for LinearSVC or any other classification algorithm you may choose
- ⦿ If class\_weight is not an option, consider using Random Over Sampling to address the imbalance



# Future Work

## Dashboard Apps

- ◎ Customizable Topic Modeling
  - User selects specific Regions and/or States and desired # of topics and top words
- ◎ Classify streaming Tweets off Live Twitter
  - User selects Location & refines query



## Improve Classifier Performance

- ◎ Collect more data & improve query terms

# Thanks!

Questions?

[leah@metisconsultingllc.com](mailto:leah@metisconsultingllc.com)

Credits:

Additional Data: Kaggle [Tweets about Distance Learning](#)

Presentation template: [SlidesCarnival](#)

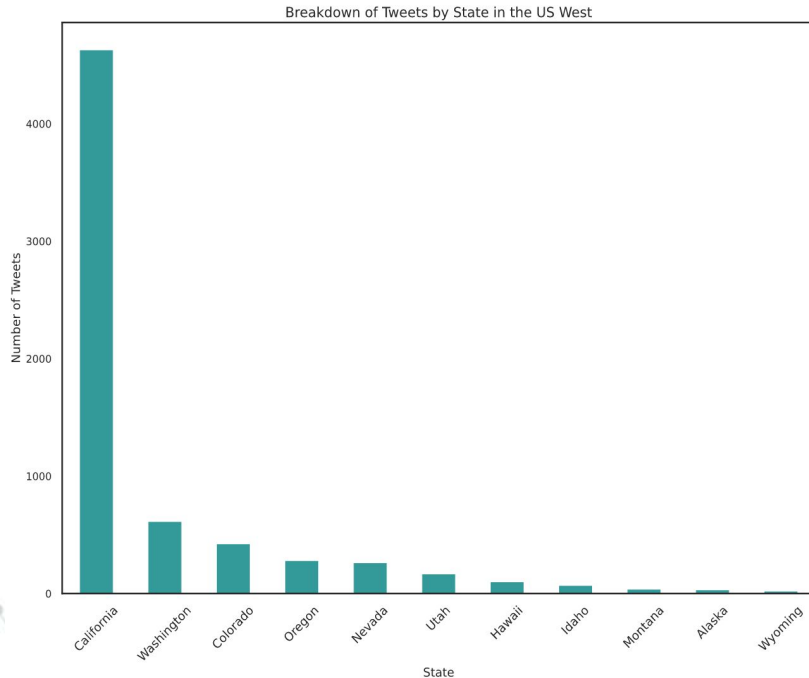
Photographs: [Unsplash](#)

# APPENDIX





# Exploratory Data Analysis



- 2 Regions have a 'vocal' State (Tweet count far exceeds others)
- West:** 69.6% of Tweets are from California, 9.3% from Washington
- Southwest:** (*not shown*) - 71.50% of Tweets are from Texas, 18.40% from Arizona