Datathon 2021

Team Member(s): Minh Hai Nguyen

Introduction

Hi, I’m Hai! (Minh Hai Nguyen)

Nationalities: Czech & Vietnamese

University: Queen Mary, University of London

Major: MSc. Machine Learning for Visual Data Analytics

Background: BSc. Computer Science

LinkedIn: https://www.linkedin.com/in/nminhhai/

Topic: The New Normal under COVID

NLP, sentiment analysis, twitter web scraping, data visualisation with streamlit

Why is this important?

COVID is a recent event which had a global impact

Twitter is a massive social media platform and using NLP, one can gain valuable insight

NLP is powerful in social sciences where one can analyse human’s attributes such as behavior, emotion or preference, hence the application is wide (understanding customers, replicating humans’ responses through chatbots etc.)

Method

NLP: Sentiment Classifier - Pre-processing

Amazon app review dataset obtained from: http://jmcauley.ucsd.edu/data/amazon/

Feature selection - only ratings and reviews themselves were used

Reviews with 3 star and above were considered as POSITIVE

Subsampled the dataset such that the classes were balanced and the sample size was small enough for the computational resources available

Removed punctuations and stop words

Dataset size after subsampling: ~50,000

NLP: Sentiment Classifier - Building a model

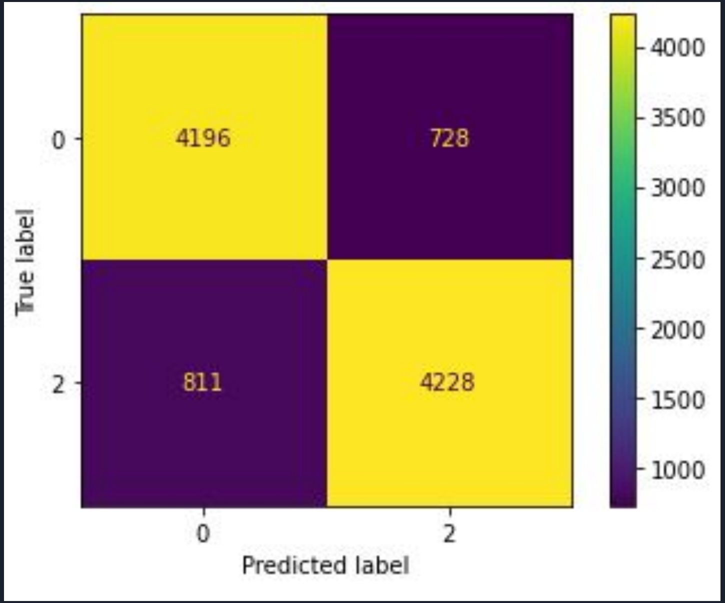
Train/Test split: 80/20 (stratified splitting)

Transformed to TFIDVectors

Algorithm: Logistic Regression

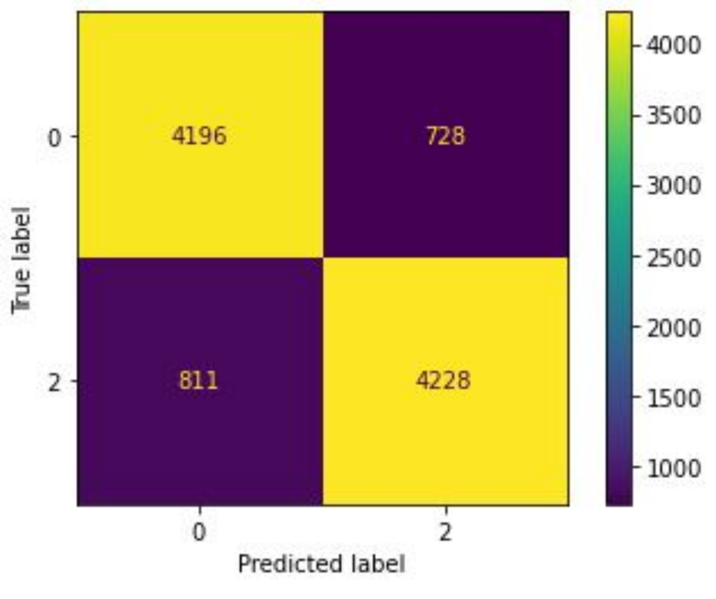
Accuracy: 84.5% on unseen test case

Confusion Matrix:

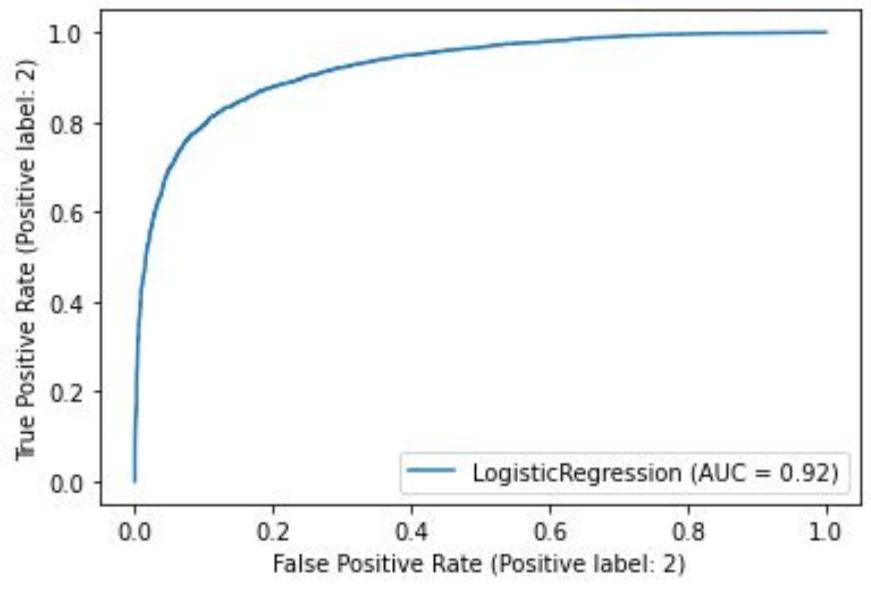


NLP: Sentiment Classifier - Model Performance

Confusion Matrix

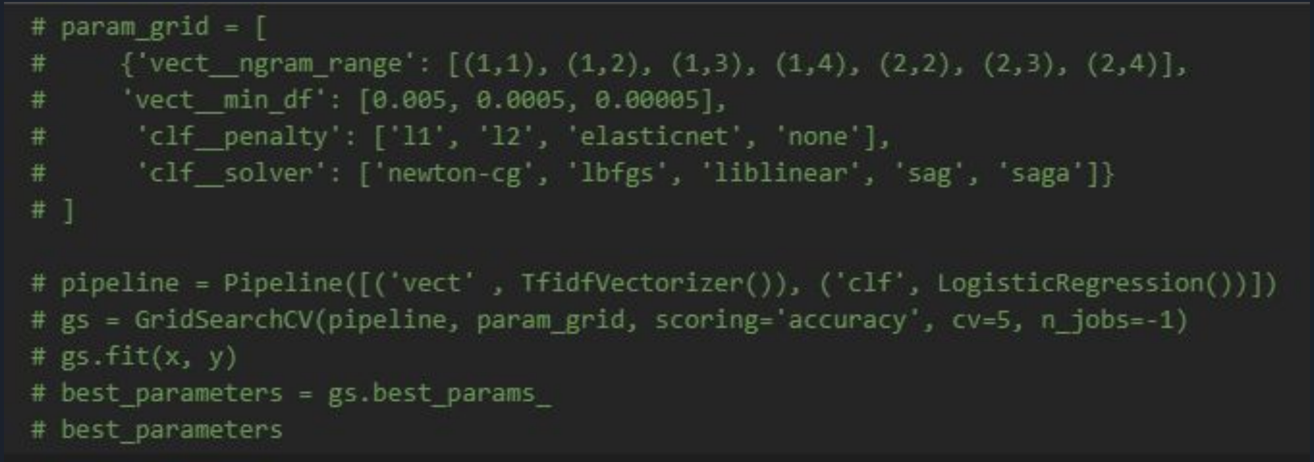


ROC Curve



NLP: Sentiment Classifier - Improvement

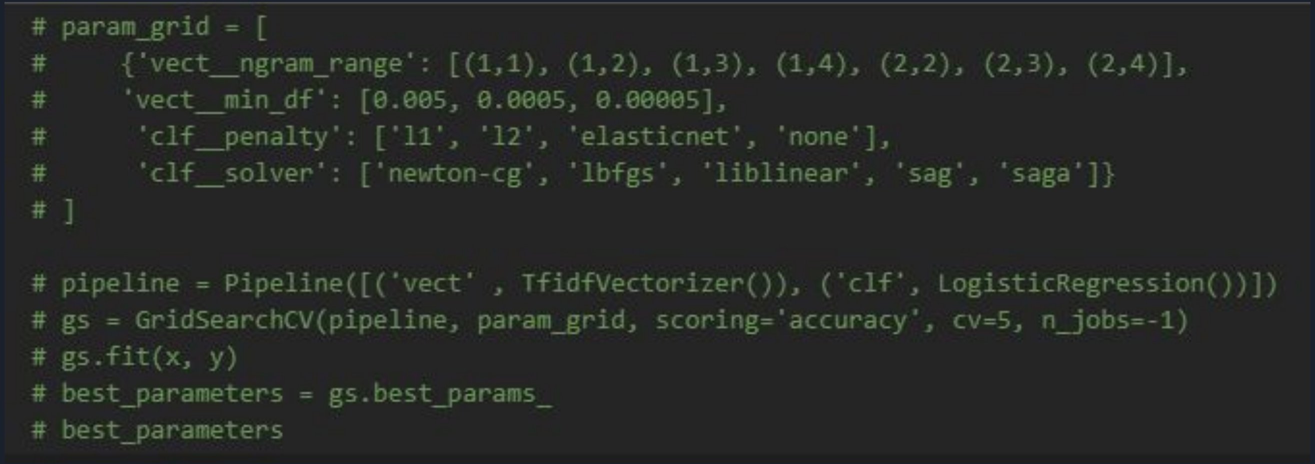
One can apply GridSearch or RandomSearch for tuning the vectorizer and the model (finding the optimal n-gram level for classification)



```
# param_grid = [
#     {'vect__ngram_range': [(1,1), (1,2), (1,3), (1,4), (2,2), (2,3), (2,4)],
#     'vect__min_df': [0.005, 0.0005, 0.00005],
#      'clf__penalty': ['l1', 'l2', 'elasticnet', 'none'],
#      'clf__solver': ['newton-cg', 'lbfgs', 'liblinear', 'sag', 'saga']}
# ]

# pipeline = Pipeline([('vect' , TfidfVectorizer()), ('clf', LogisticRegression())])
# gs = GridSearchCV(pipeline, param_grid, scoring='accuracy', cv=5, n_jobs=-1)
# gs.fit(x, y)
# best_parameters = gs.best_params_
# best_parameters
```

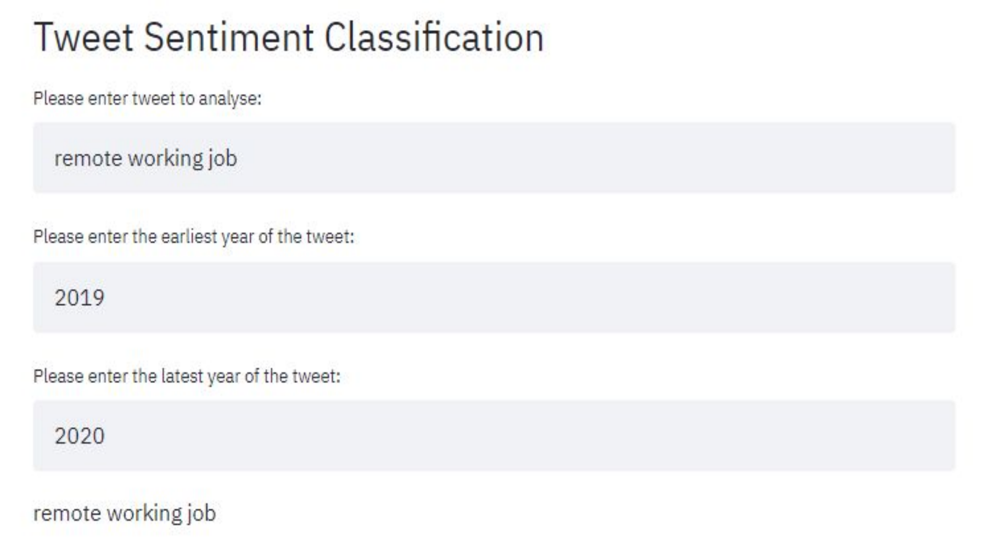
NLP: Sentiment Classifier - Improvement

One can apply GridSearch or RandomSearch for tuning the vectorizer and the model (finding the optimal n-gram level for classification)



```
# param_grid = [
#     {'vect__ngram_range': [(1,1), (1,2), (1,3), (1,4), (2,2), (2,3), (2,4)],
#     'vect__min_df': [0.005, 0.0005, 0.00005],
#      'clf__penalty': ['l1', 'l2', 'elasticnet', 'none'],
#      'clf__solver': ['newton-cg', 'lbfgs', 'liblinear', 'sag', 'saga']}
# ]

# pipeline = Pipeline([('vect' , TfidfVectorizer()), ('clf', LogisticRegression())])
# gs = GridSearchCV(pipeline, param_grid, scoring='accuracy', cv=5, n_jobs=-1)
# gs.fit(x, y)
# best_parameters = gs.best_params_
# best_parameters
```

NLP: Sentiment Classifier - Post-processing (App) Dynamic predictions based on user input + data visualisation



Tweet Sentiment Classification

Please enter tweet to analyse:

remote working job

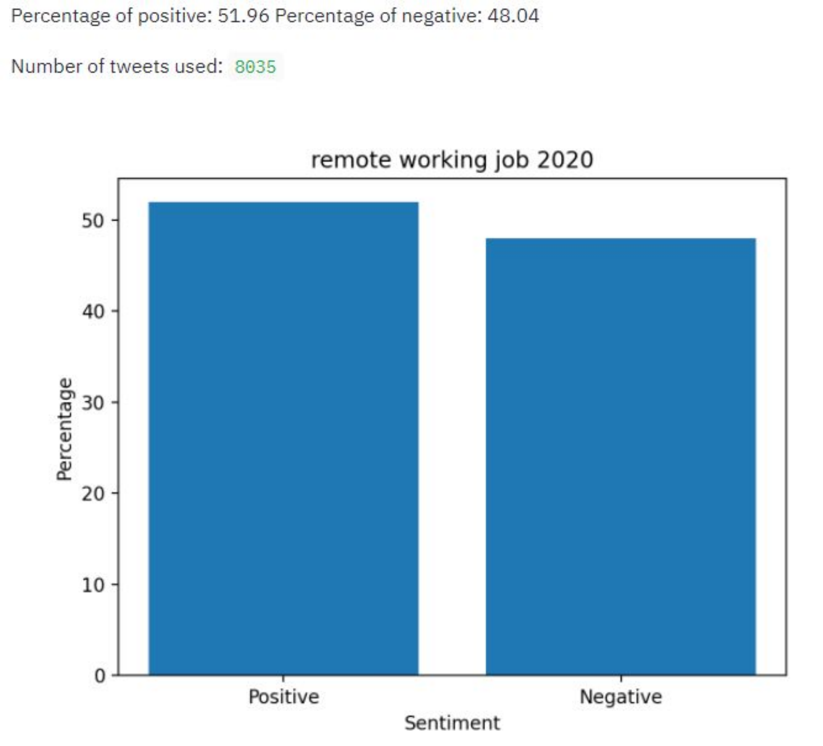
Please enter the earliest year of the tweet:

2019

Please enter the latest year of the tweet:

2020

remote working job



Percentage of positive: 51.96 Percentage of negative: 48.04

Number of tweets used: 8035

remote working job 2020

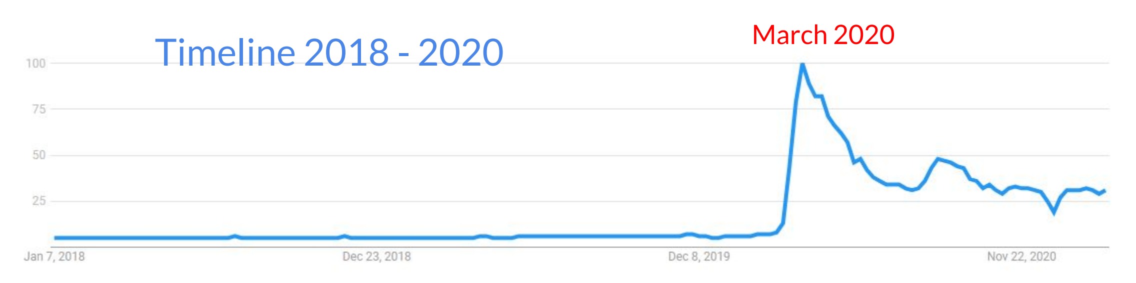
Results

Interviews and meetings held online - The new norm

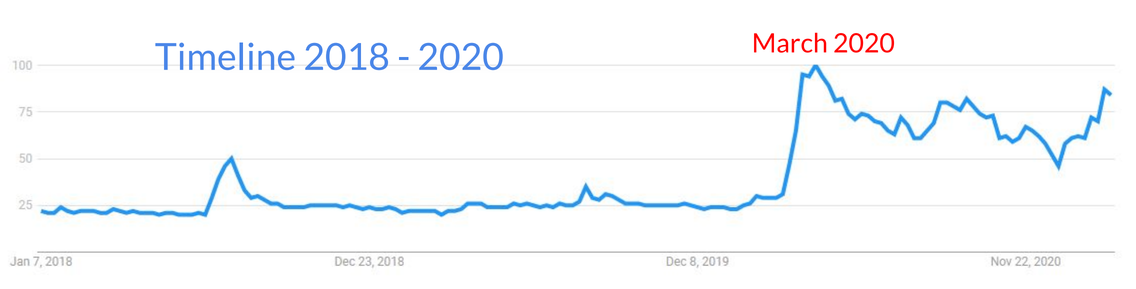
Do people like online meetings and interviews?

Zoom's Increase in Relevancy - Google Trends

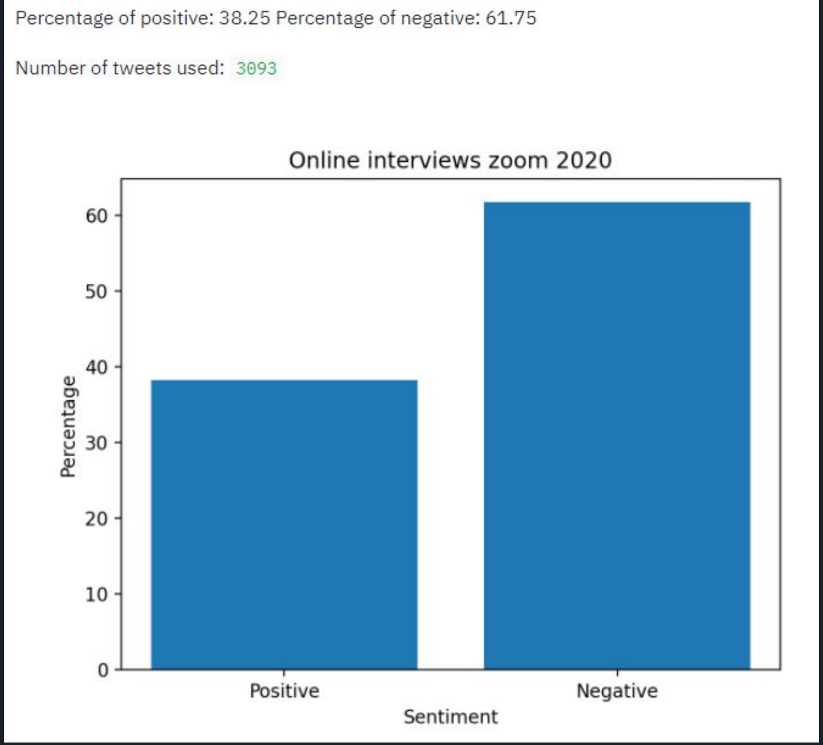
Google Search
of the word "Zoom"



YouTube Search
of the word "Zoom"



Tweets containing ‘online, interviews, zoom’
~38% Positive tweets, ~62% Negative tweets

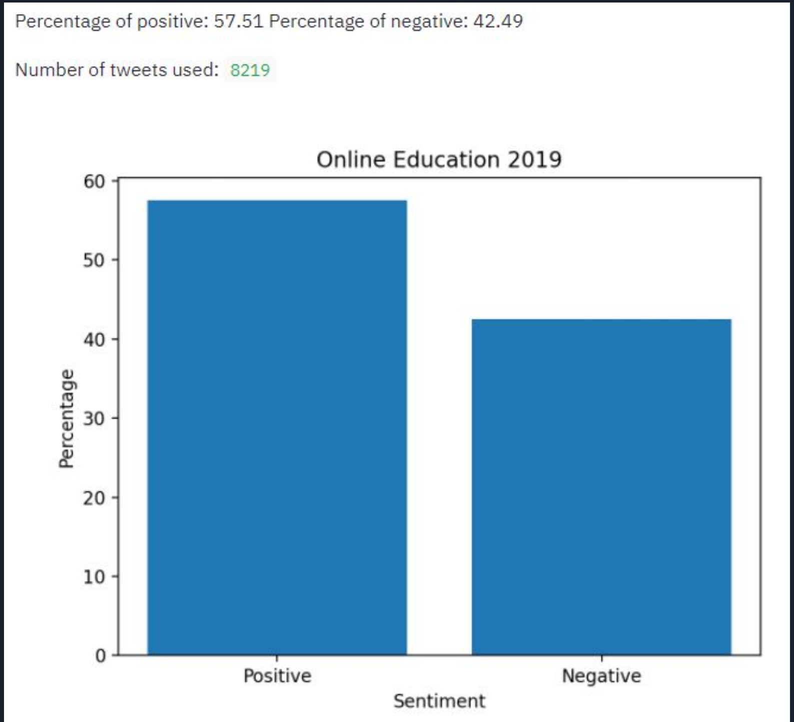


```
Percentage of positive: 38.25 Percentage of negative: 61.75

Number of tweets used:  3093
```

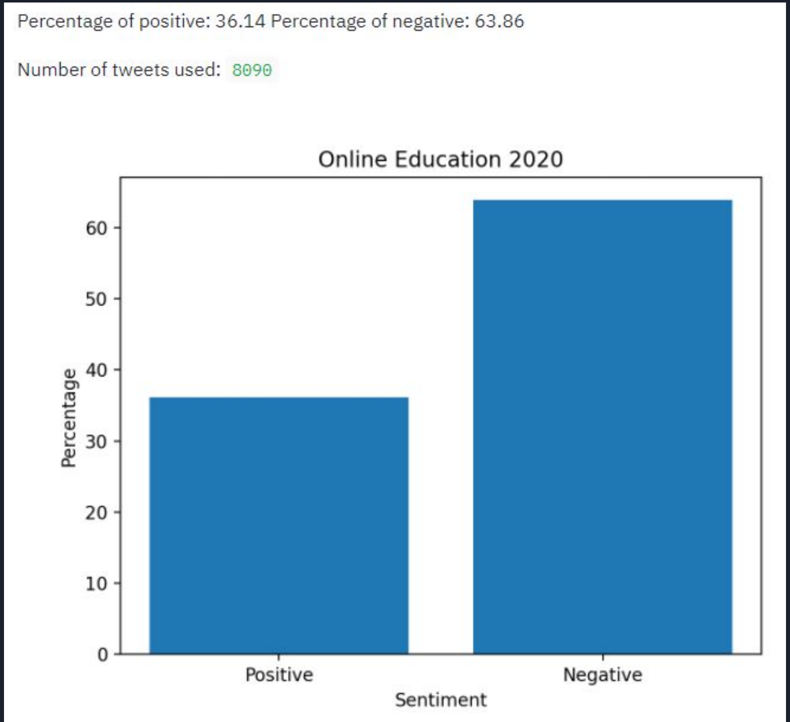
Online Education - The New Norm

How did University Students view online education?

Tweet containing ‘online, education’
~20% decrease in positive tweets in 2020-2021


Percentage of positive: 57.51 Percentage of negative: 42.49

Number of tweets used: 8219


Percentage of positive: 36.14 Percentage of negative: 63.86

Number of tweets used: 8090

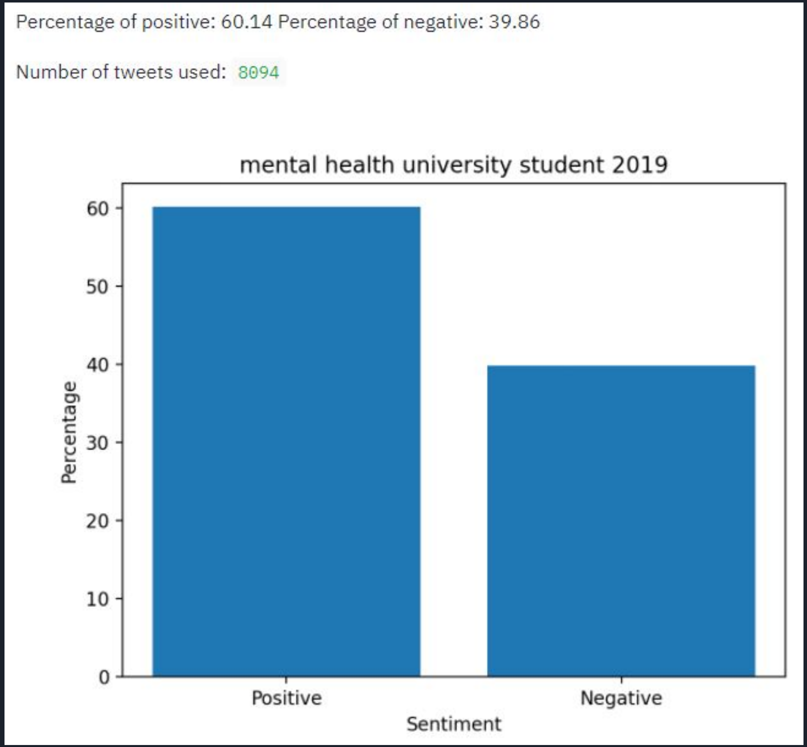
University Students outlook on Online Education ~20% decrease in positive tweets in 2020-2021

Interpretation: Prior to 2020, online education often referred to Coursera, edx, Udemy which are appreciated by many learners. In 2020, online university education became the norm, and sentiment towards ‘online education’ is no longer viewed in the positive light

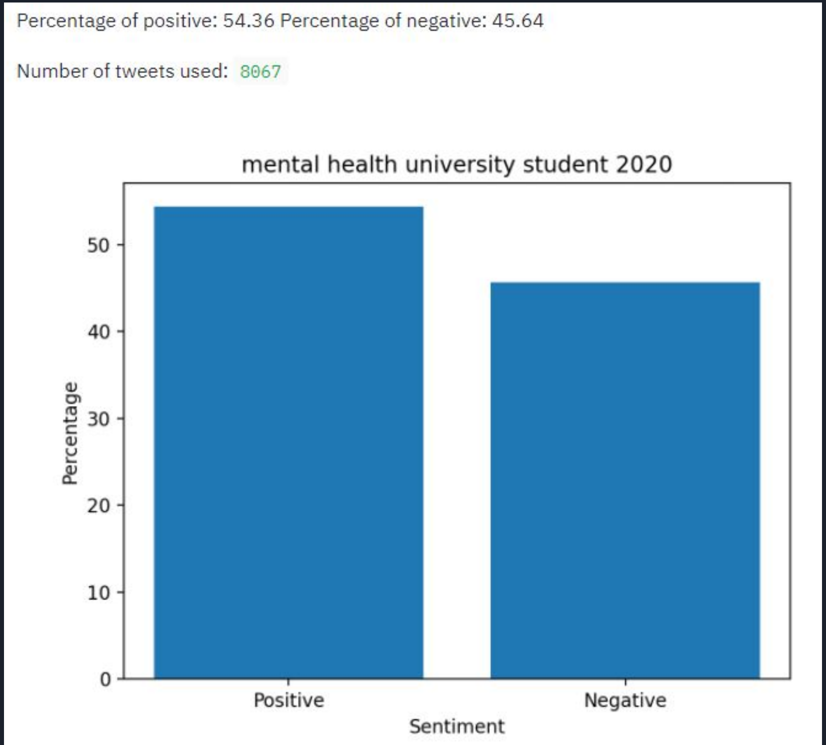
Mental Health of University Students

Is there any change in sentiment regards to tweets about mental health?

Tweets containing ‘mental, health, university, student’ ~6% decrease in positive tweets in 2020

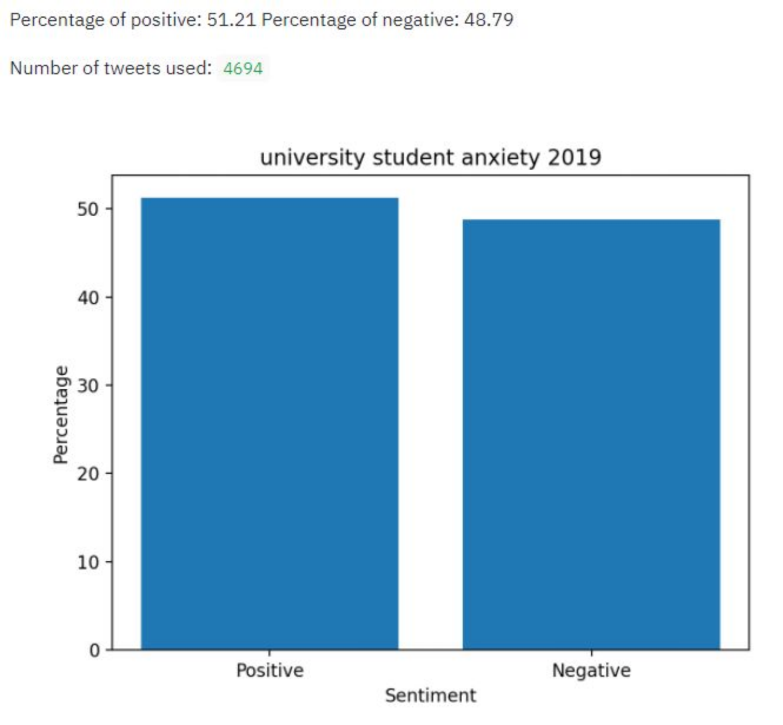


```
Percentage of positive: 60.14 Percentage of negative: 39.86

Number of tweets used:  8094
```



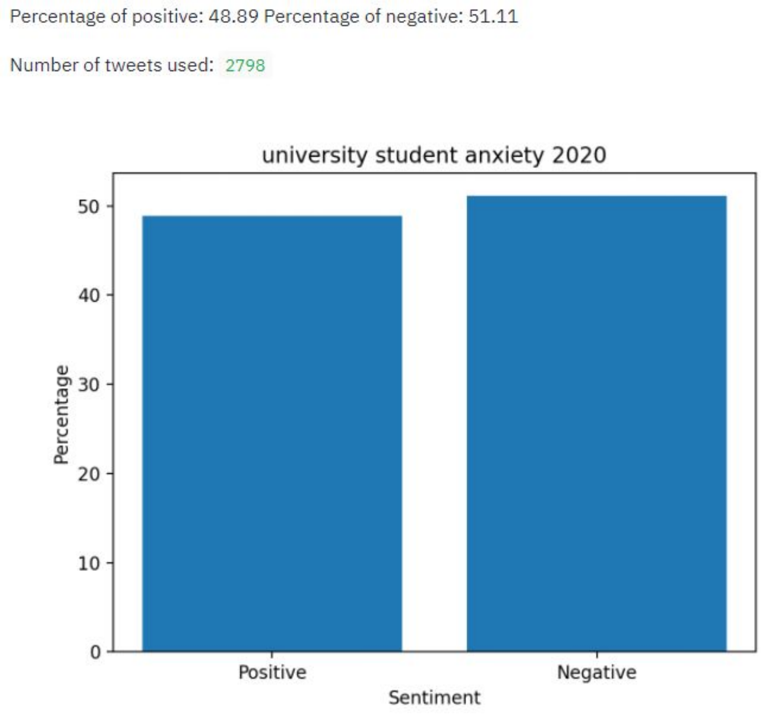
```
Percentage of positive: 54.36 Percentage of negative: 45.64

Number of tweets used:  8067
```

Tweets containing ‘university, student, anxiety’
~2% decrease in positive tweets in 2020
~50% decrease in number of tweets



Percentage of positive: 51.21 Percentage of negative: 48.79

Number of tweets used: 4694



Percentage of positive: 48.89 Percentage of negative: 51.11

Number of tweets used: 2798

Tweet containing ‘university, student, anxiety’
~2% decrease in positive tweets in 2020
~50% decrease in number of tweets

Interpretation: 2019 had higher percentage of positive tweets; the tweets could have been related to mental support and overcoming anxiety. However, with 50% decrease in number of tweets related to anxiety, perhaps people did feel more anxious and therefore no longer posted actively positive tweets regarding anxiety.

Interpretations of reduction in positive tweets regarding mental health by University Students

Lack of social networking

Limited in-class interaction (although engagement is increased in chat!)

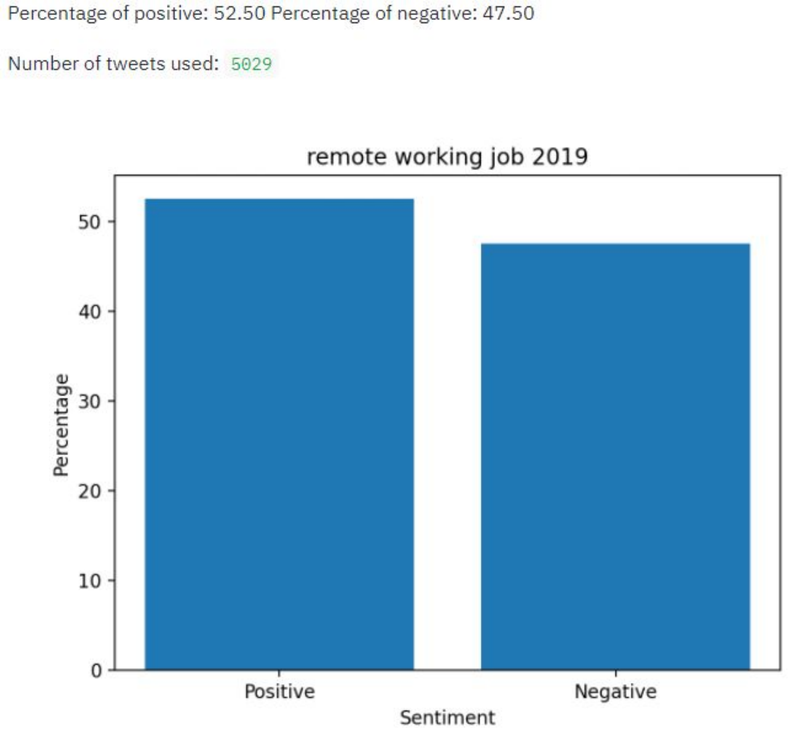
Societies closed, parties are disallowed and University life is not what it used to be

~6% and ~2% decrease is not too significant - some may take it as an advantage, no commuting required and pre-recorded lectures allow them to study at their own pace

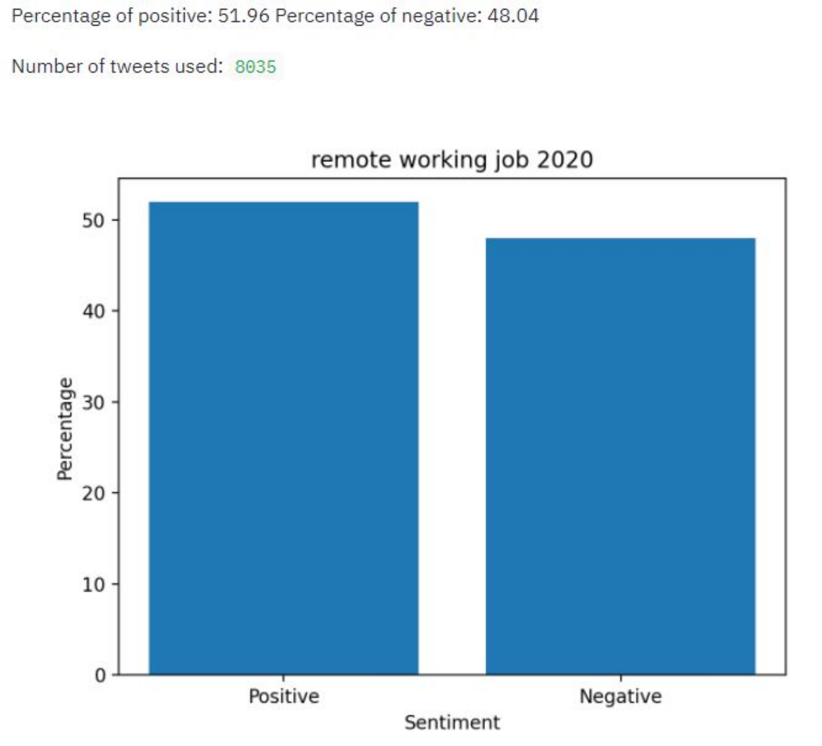
Remote Working - The New Norm

How did employees view remote working?

Tweets containing ‘remote, working, job’
~1% decrease in positive tweets in 2020
Insignificant change



```
Percentage of positive: 52.50 Percentage of negative: 47.50

Number of tweets used:  5029
```

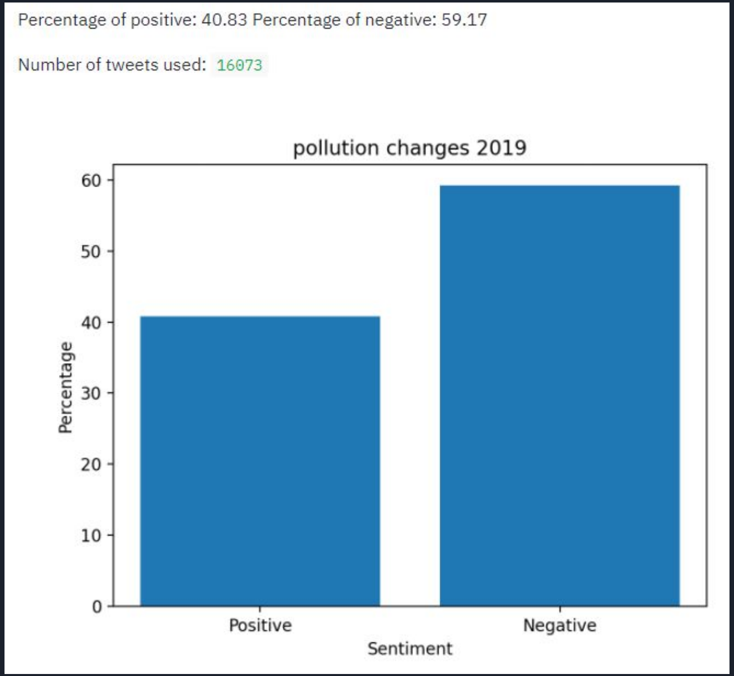


```
Percentage of positive: 51.96 Percentage of negative: 48.04

Number of tweets used:  8035
```

Staying indoors - The New Norm

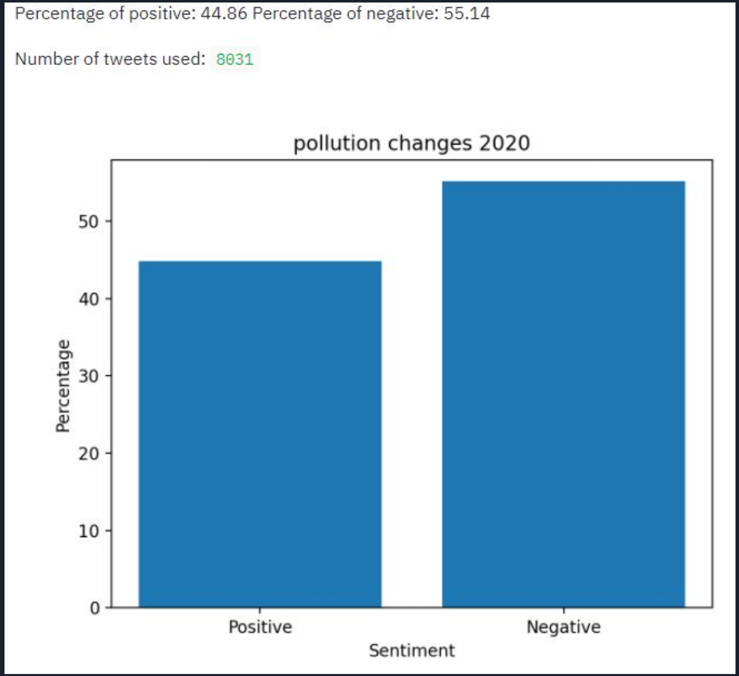
Can this be beneficial?

Tweets containing ‘pollution, changes’
~4% increase in positive tweets in 2020 due to reduction in pollution
Source: https://www.bbc.com/news/uk-scotland-53015092



Percentage of positive: 40.83 Percentage of negative: 59.17

Number of tweets used: `16073`



Percentage of positive: 44.86 Percentage of negative: 55.14

Number of tweets used: `8031`

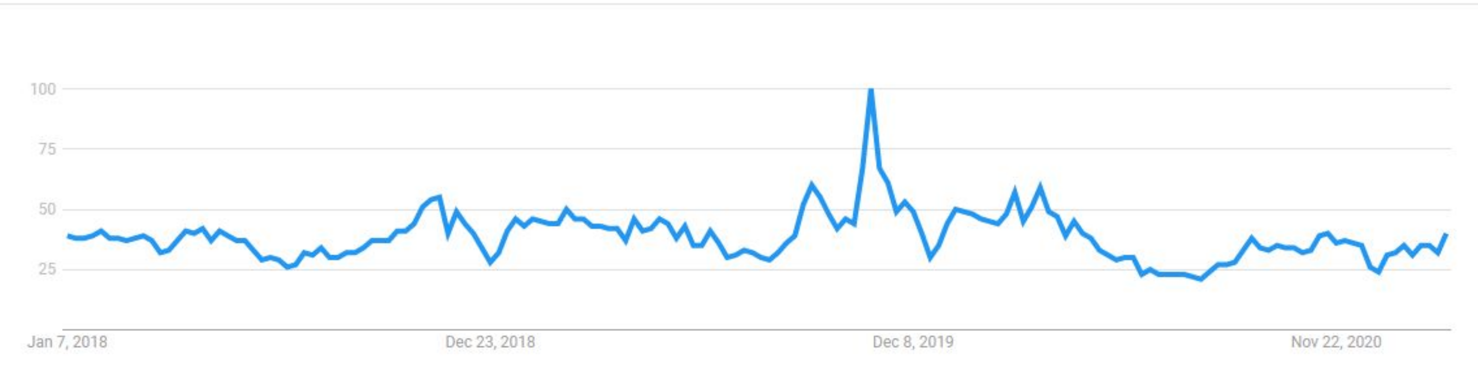
Pollution Changes - did we reduce pollution intentionally?

Pollution Web Searches

No increase in pollution searches on Google trends during the pandemic

Possible interpretation: no changes in the trend suggest no major policy or action was taken to reduce pollution; people did not intentionally increase their efforts to reduce pollution; rather, the reduction in pollution was a result of lockdowns

Alternative interpretation: one does not necessarily need to do web searches when he/she decides to contribute to pollution reduction, hence this may not be portrayed by Google trends



Further Improvement

Create a word cloud to display the most common words that occur in these tweets to obtain deeper insight into the motives behind sentiment changes

Limit number of tweets to be extracted is not consistent (e.g. number of tweets sometimes cross the limit) - potential error in the Twint library as it was noted by some developers (source: https://github.com/twintproject/twint/issues/237)

Conclusion

COVID has changed the way we live and introduced several new norms:

- Zoom became one of our main tools
- Online education became the norm
- Working remotely became the norm
- Staying indoors on regular basis became the norm

Stay home, stay safe!

Datathon Challenge Takeaway

Learnt how to use TWINT for Web Scraping tweets from Twitter

Learnt how to use Streamlit for deploying an app

Sharpened NLP skills

Gained rich insight into the life of a Data Scientist through discussion panels

All within 1 day!

Thank you so much for organising this event, it was a pleasure!