Twitter Sentiment Analysis

Background

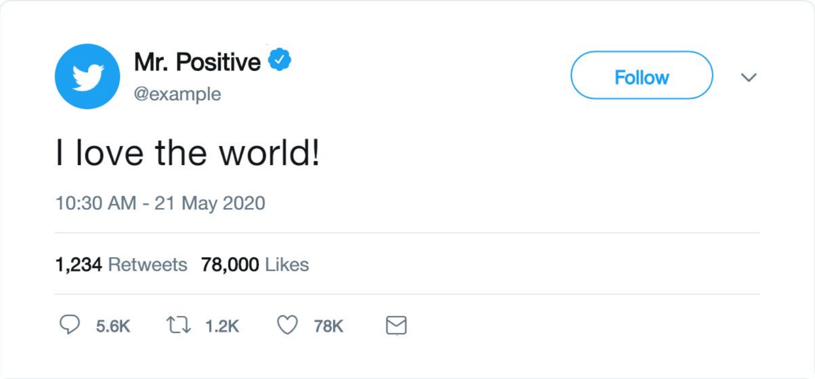
- Social media has created a new way for individuals to express their thoughts and opinions
- This medium is used by an estimated 2.95 billion people worldwide
- Sentiment analysis is the process of retrieving textual information and discerning which emotions are exhibited by the author

Introduction

- Assume that each tweet falls into one of three categories
  - Negative
  - Neutral
  - Positive

- Recognizing each sentiment does not have the same level of difficulty

- Real people can only agree on sentiment 70-90% of the time



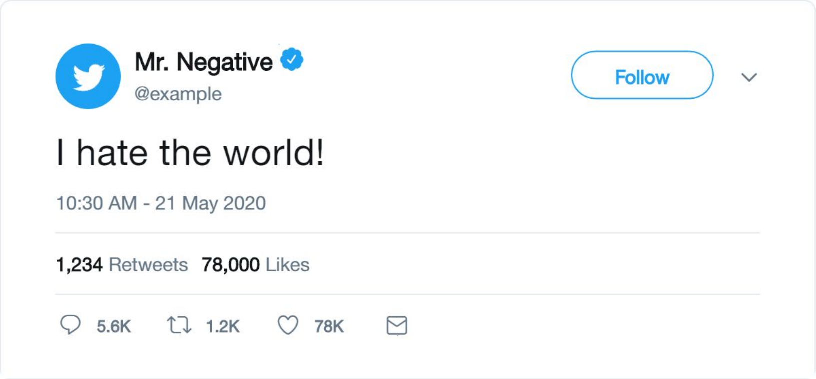
Mr. Positive

@example

I love the world!

10:30 AM - 21 May 2020

1,234 Retweets 78,000 Likes



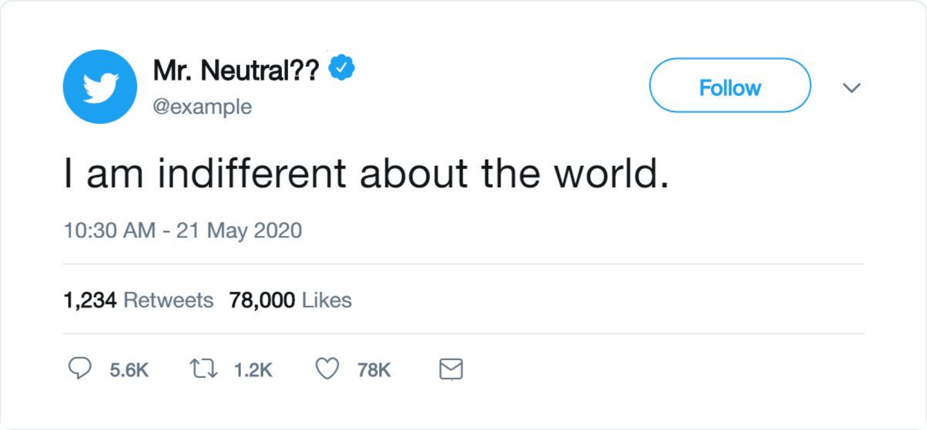
Mr. Negative

@example

I hate the world!

10:30 AM - 21 May 2020

1,234 Retweets 78,000 Likes



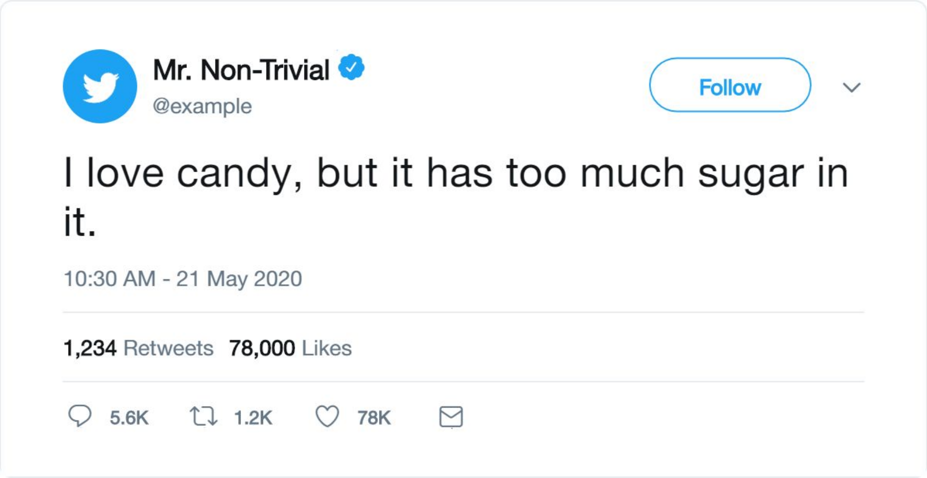
Mr. Neutral??

@example

I am indifferent about the world.

10:30 AM - 21 May 2020

1,234 Retweets 78,000 Likes



Mr. Non-Trivial

@example

I love candy, but it has too much sugar in it.

10:30 AM - 21 May 2020

1,234 Retweets 78,000 Likes

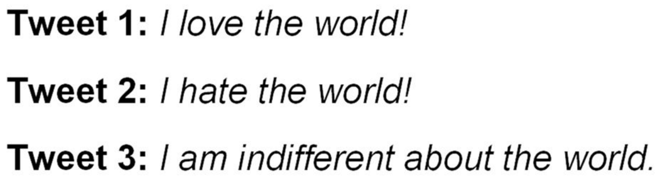
Problem Statement

- For this project, we aim to classify each tweet as either
    - Negative
    - Neutral
    - Positive

- *Accuracy* will be measured in two ways
    - % of tweets correctly classified
    - Precision of the average (aggregate) score of a basket of tweets

Methodology

Methodology

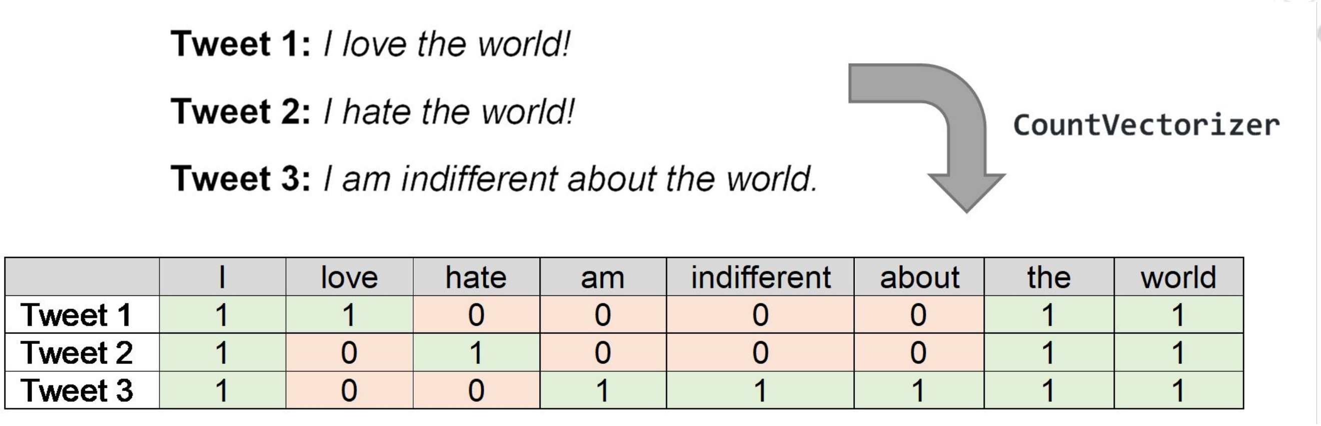


Tweet 1: *I love the world!*

Tweet 2: *I hate the world!*

Tweet 3: *I am indifferent about the world.*

Methodology



Tweet 1: *I love the world!*

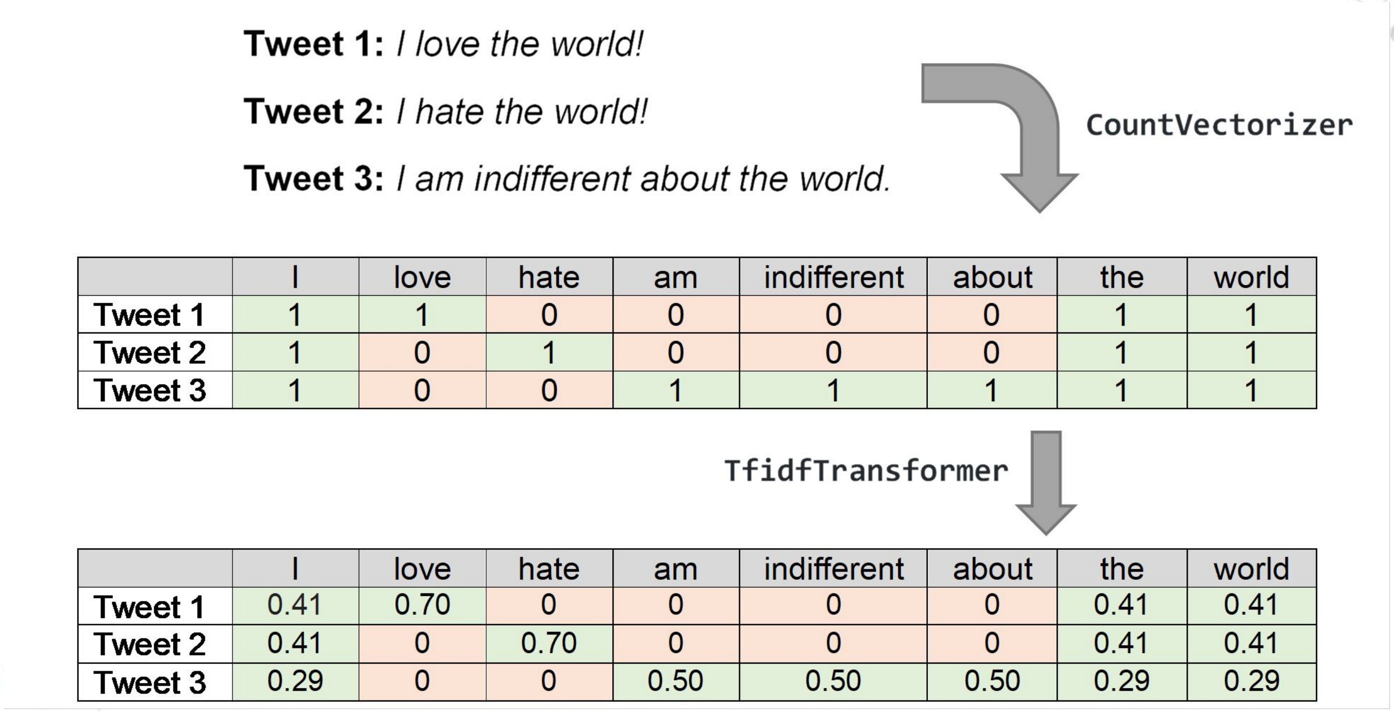
Tweet 2: *I hate the world!*

Tweet 3: *I am indifferent about the world.*

`CountVectorizer`

	I	love	hate	am	indifferent	about	the	world
Tweet 1	1	1	0	0	0	0	1	1
Tweet 2	1	0	1	0	0	0	1	1
Tweet 3	1	0	0	1	1	1	1	1

Methodology



Tweet 1: *I love the world!*

Tweet 2: *I hate the world!*

Tweet 3: *I am indifferent about the world.*

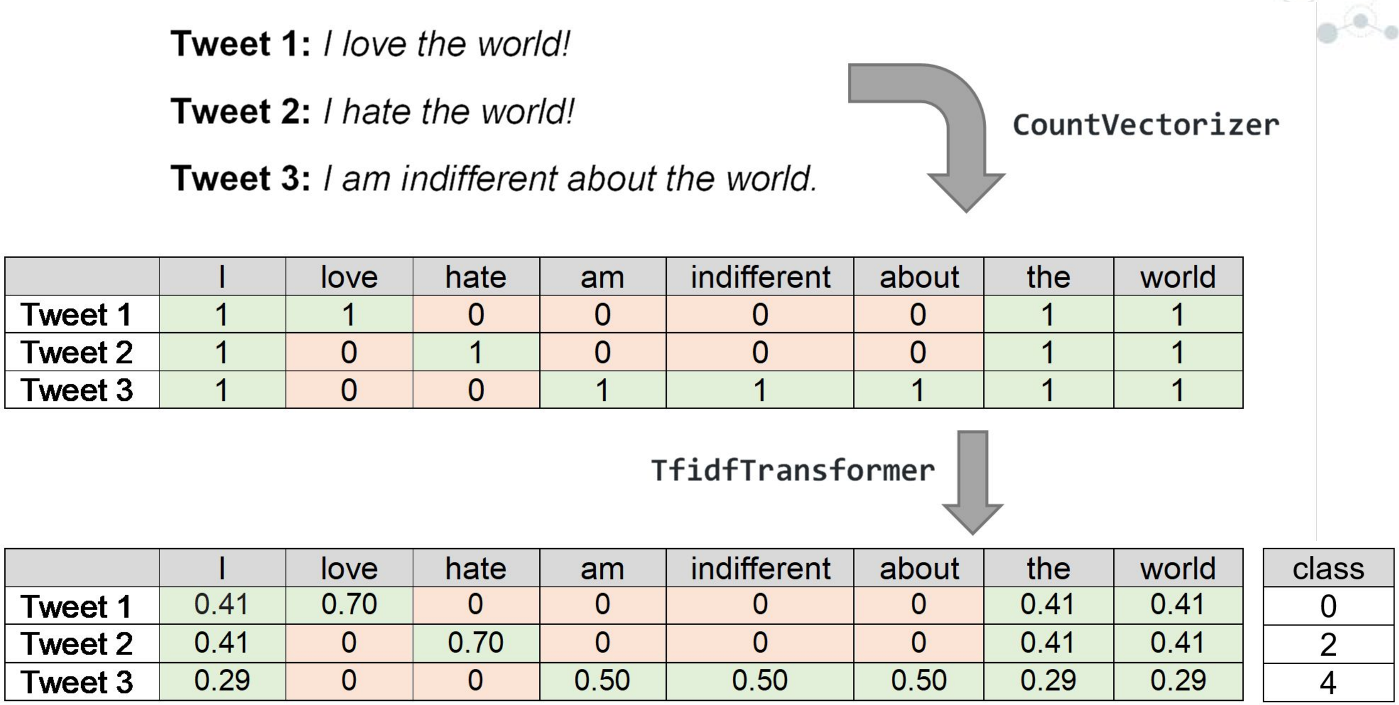
`CountVectorizer`

	I	love	hate	am	indifferent	about	the	world
Tweet 1	1	1	0	0	0	0	1	1
Tweet 2	1	0	1	0	0	0	1	1
Tweet 3	1	0	0	1	1	1	1	1

`TfidfTransformer`

	I	love	hate	am	indifferent	about	the	world
Tweet 1	0.41	0.70	0	0	0	0	0.41	0.41
Tweet 2	0.41	0	0.70	0	0	0	0.41	0.41
Tweet 3	0.29	0	0	0.50	0.50	0.50	0.29	0.29

Methodology



Tweet 1: *I love the world!*

Tweet 2: *I hate the world!*

Tweet 3: *I am indifferent about the world.*

	I	love	hate	am	indifferent	about	the	world
Tweet 1	1	1	0	0	0	0	1	1
Tweet 2	1	0	1	0	0	0	1	1
Tweet 3	1	0	0	1	1	1	1	1

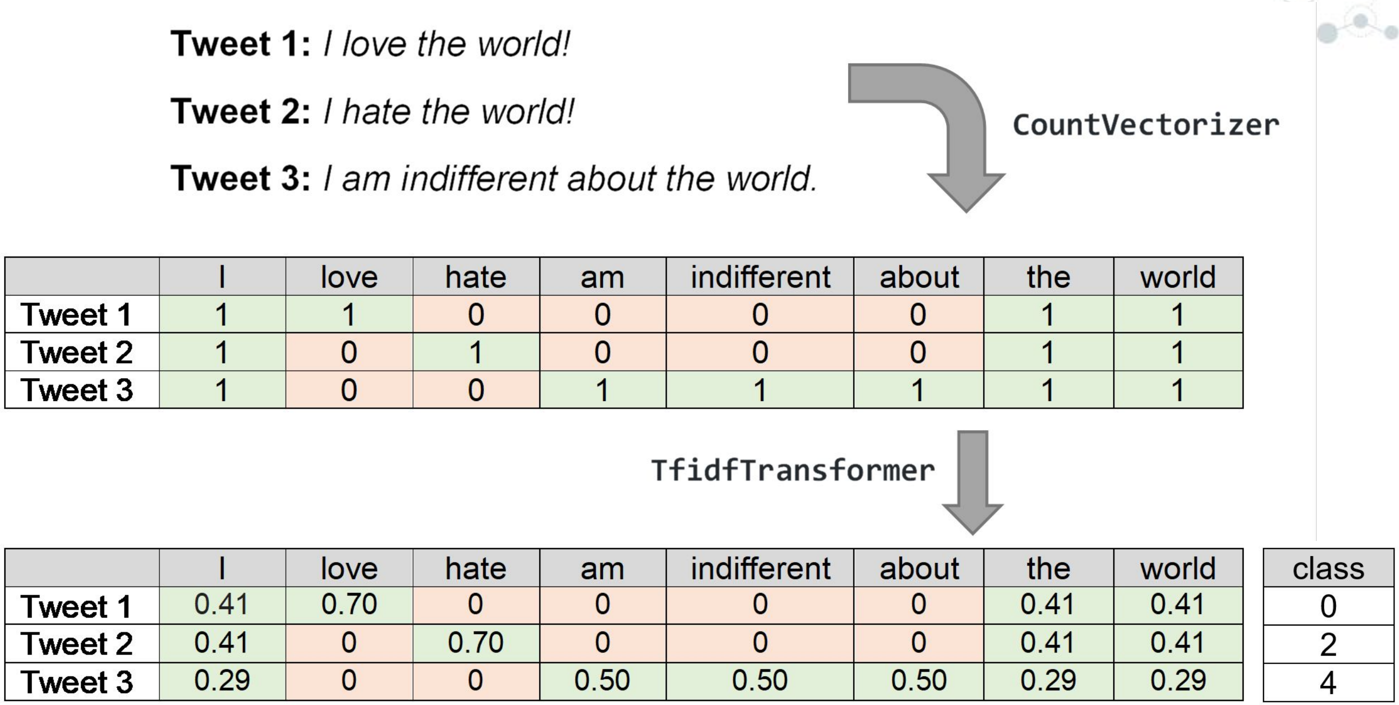
	I	love	hate	am	indifferent	about	the	world	class
Tweet 1	0.41	0.70	0	0	0	0	0.41	0.41	0
Tweet 2	0.41	0	0.70	0	0	0	0.41	0.41	2
Tweet 3	0.29	0	0	0.50	0.50	0.50	0.29	0.29	4

Predict class using Random Forest, KNeighbors, Logistic Regression, etc.

Methodology

Pipeline

allows for automated cross validation



Tweet 1: *I love the world!*

Tweet 2: *I hate the world!*

Tweet 3: *I am indifferent about the world.*

`CountVectorizer`

	I	love	hate	am	indifferent	about	the	world
Tweet 1	1	1	0	0	0	0	1	1
Tweet 2	1	0	1	0	0	0	1	1
Tweet 3	1	0	0	1	1	1	1	1

`TfidfTransformer`

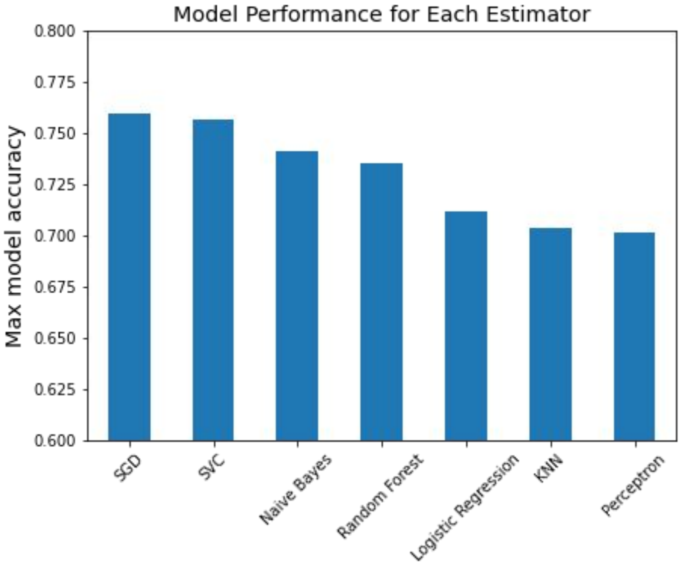
	I	love	hate	am	indifferent	about	the	world
Tweet 1	0.41	0.70	0	0	0	0	0.41	0.41
Tweet 2	0.41	0	0.70	0	0	0	0.41	0.41
Tweet 3	0.29	0	0	0.50	0.50	0.50	0.29	0.29

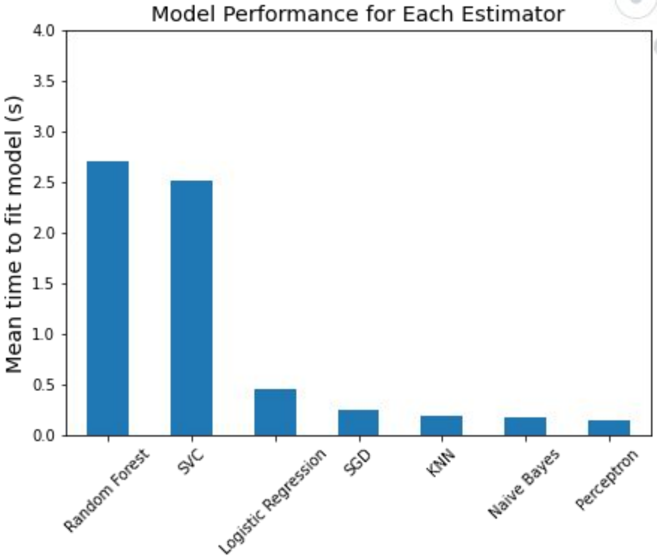
class
0
2
4

Predict class using Random Forest, KNeighbors, Logistic Regression, etc.

Results

Results: Estimator Performance





- Best accuracy: Stochastic Gradient Descent (SGD)
- Worst accuracy: Perceptron

- Fastest: Perceptron
- Slowest: Random Forest

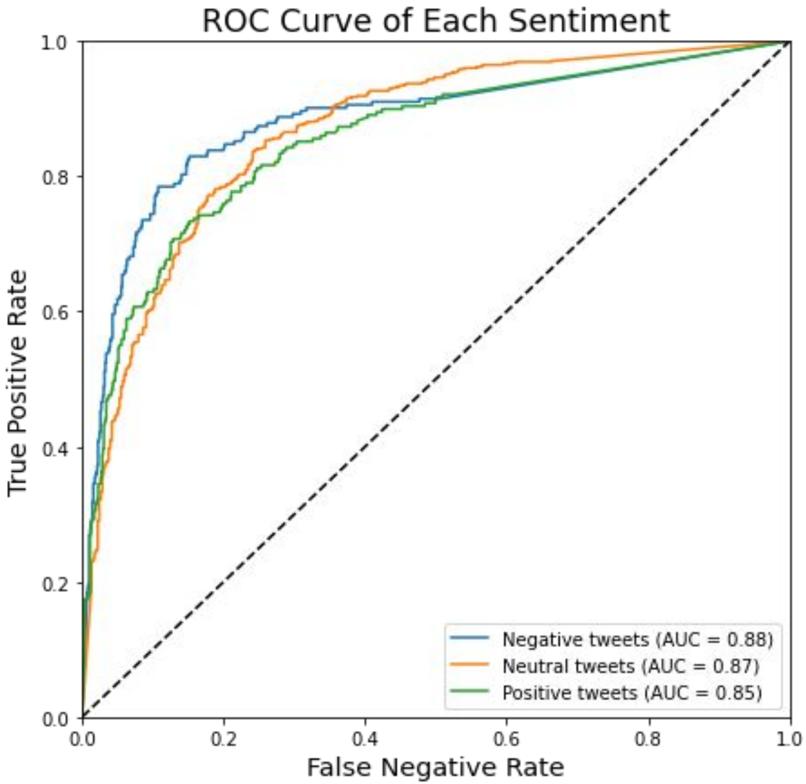
Results: Best Model Performance

- Accuracy was very good considering the subjective nature of the problem
  - Recall a “perfect” model can only achieve 70-90% accuracy

Cross validation accuracy	Test set accuracy
75.99%	77.94%

Results: Sentiment Validation

The ROC curve implies model is relatively good at predicting all three sentiments equally

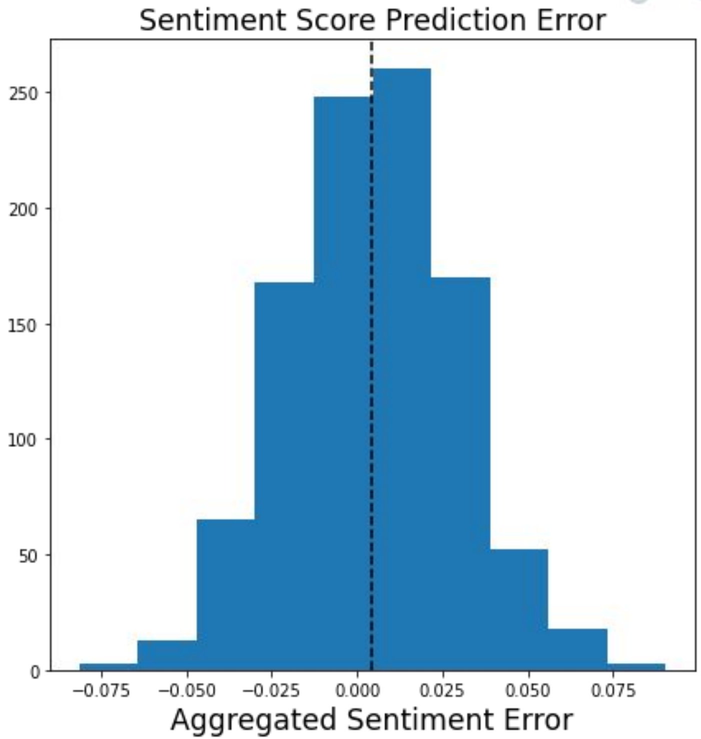


Results: Model Validation

Tweet	Negative probability	Neutral probability	Positive probability
I love the world	0%	0%	100%
I hate the world	100%	0%	0%
I am indifferent about the world	16.7%	81.1%	2.3%

Results: Aggregate Score Validation

- We created 1,000 bootstrap samples and calculated the aggregate sentiment score
- 95% of predictions were within ± 0.05 of the actual score
- Now we can test it on real tweets



Exploration

Coronavirus

Aggregate sentiment score: -3%

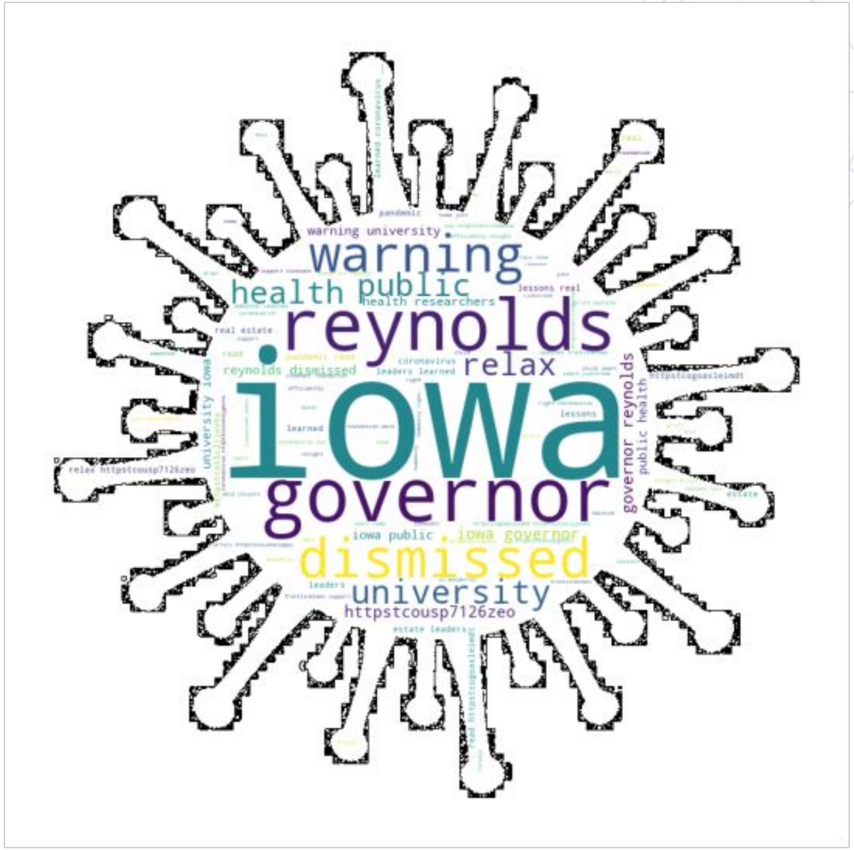
Key terms:

➢ Iowa

➢ (Iowa) Governor (Kim) Reynolds

➢ Warning

➢ Health



Coronavirus

Search Term	Reopen	Economy	School	Summer	Future
Aggregate Sentiment Score	- 9%	20%	41%	- 20%	11%

Government Officials

Search term	Agg. Score (excluding retweets)	Agg. Score (including retweets)
(Donald) Trump	1 %	6 %
(Joe) Biden	-3 %	9 %
(Nancy) Pelosi	- 43 %	- 23 %
Mitch McConnell	2 %	1 %
(Barack) Obama	19 %	28 %
Republicans	- 4 %	4 %
Democrats	- 3 %	0 %

Thank you!