Iron Hack Data Analytics
Final Project
Sentiment Analysis

Matthew Batchelor, March 2024

INSPIRED BY mathspace

I will build a model to label text by sentiment with a focus on

Performance

Accuracy: measures the overall correctness of predictions

Precision: measures the accuracy of positive predictions

Recall: measures the ability to identify all positive instances

F1-score: balances precision and recall, considering false positives and false negatives.

Solving a real-world problem

Problem statement

Companies receive more and more text-based feedback and fail to leverage that input to its maximum potential through effective prioritisation

Proposed solution

My model will empower companies to unleash the potential of their customer-base through categorising customer feedback at scale

Useability

Model should:

- be accessible to users with simple import of a csv pathway
- handle large sets of text data (as opposed to just individual lines)
- return output in useful format (CSV file)

Business case expanded

Customer Success Co-Pilot

Enables Customer Success Managers to prioritise workload

Audience Sentiment Manager

Empower companies to unleash the potential of their customer feedback

Value proposition

We empower companies to take a ‘temperature check’ of their customer sentiment quickly and at scale

Interpret sentiment of 1000s of instances of customer feedback in seconds

IMPORT

User imports (large) CSV
of customer reviews or feedback

SENTIMENT PREDICTED

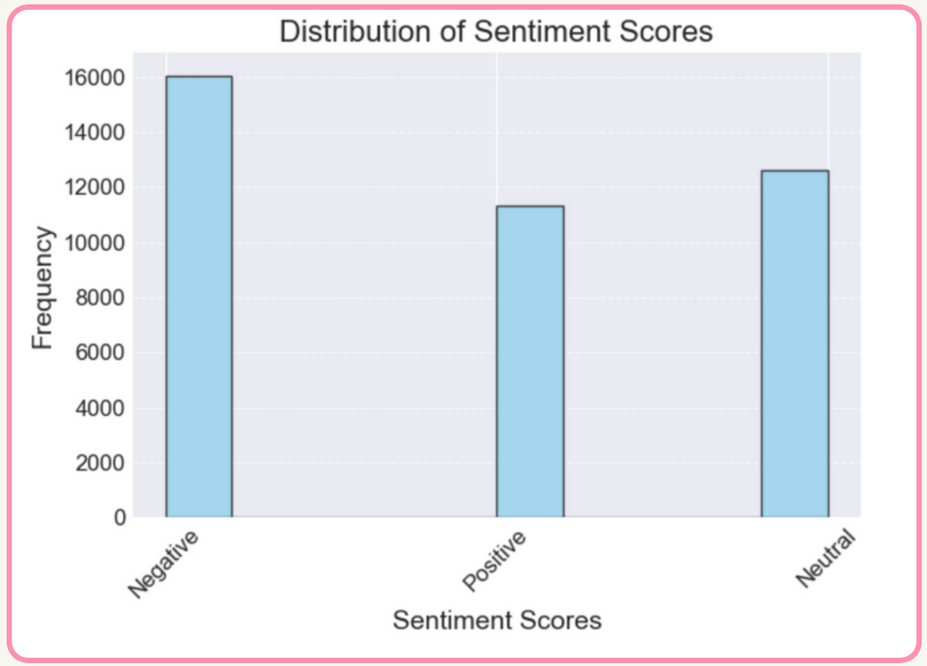
User input is preprocessed and sentiment
of each instance is predicted

OUTPUT OF LABELLED DATA

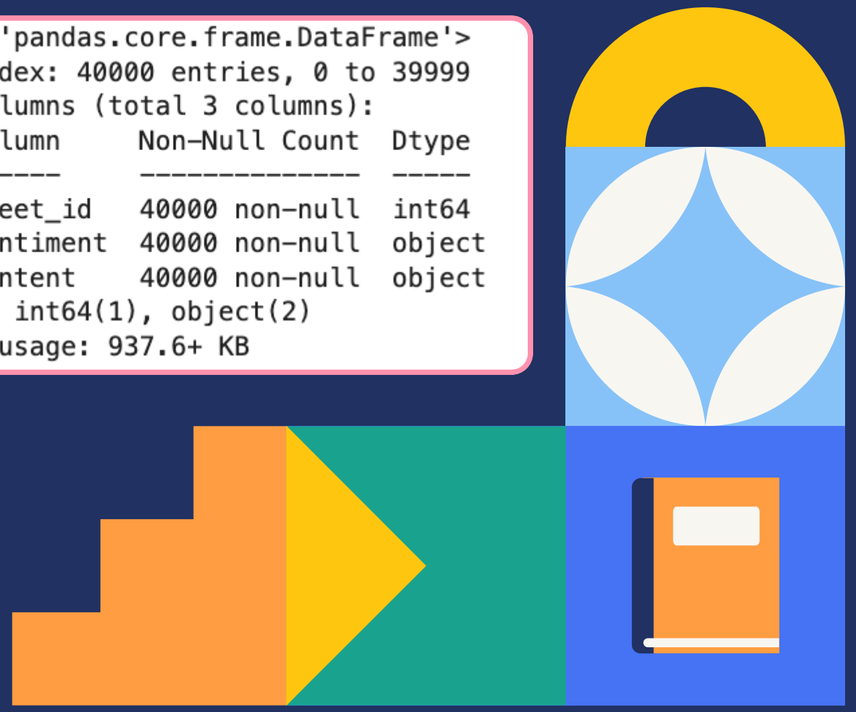
User receives table of input data with
additional sentiment prediction label

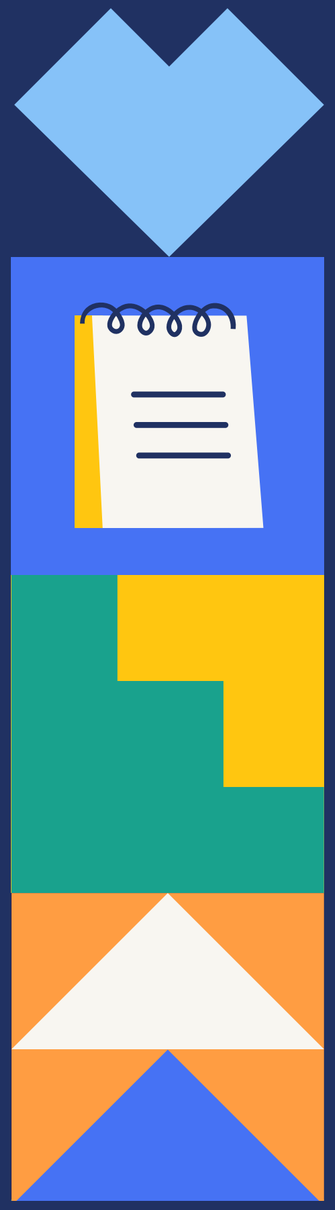
Training data

- 40 000 tweets
  - 13 sentiments
    - ['empty', 'sadness', 'enthusiasm', 'neutral', 'worry', 'surprise', 'love', 'fun', 'hate', 'happiness', 'boredom', 'relief', 'anger']
  - Text length of 73 characters on average
- Source: Kaggle




```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 40000 entries, 0 to 39999
Data columns (total 3 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   tweet_id   40000 non-null  int64
 1   sentiment  40000 non-null  object
 2   content    40000 non-null  object
dtypes: int64(1), object(2)
memory usage: 937.6+ KB
```





Tech Stack

Machine Learning (ML)

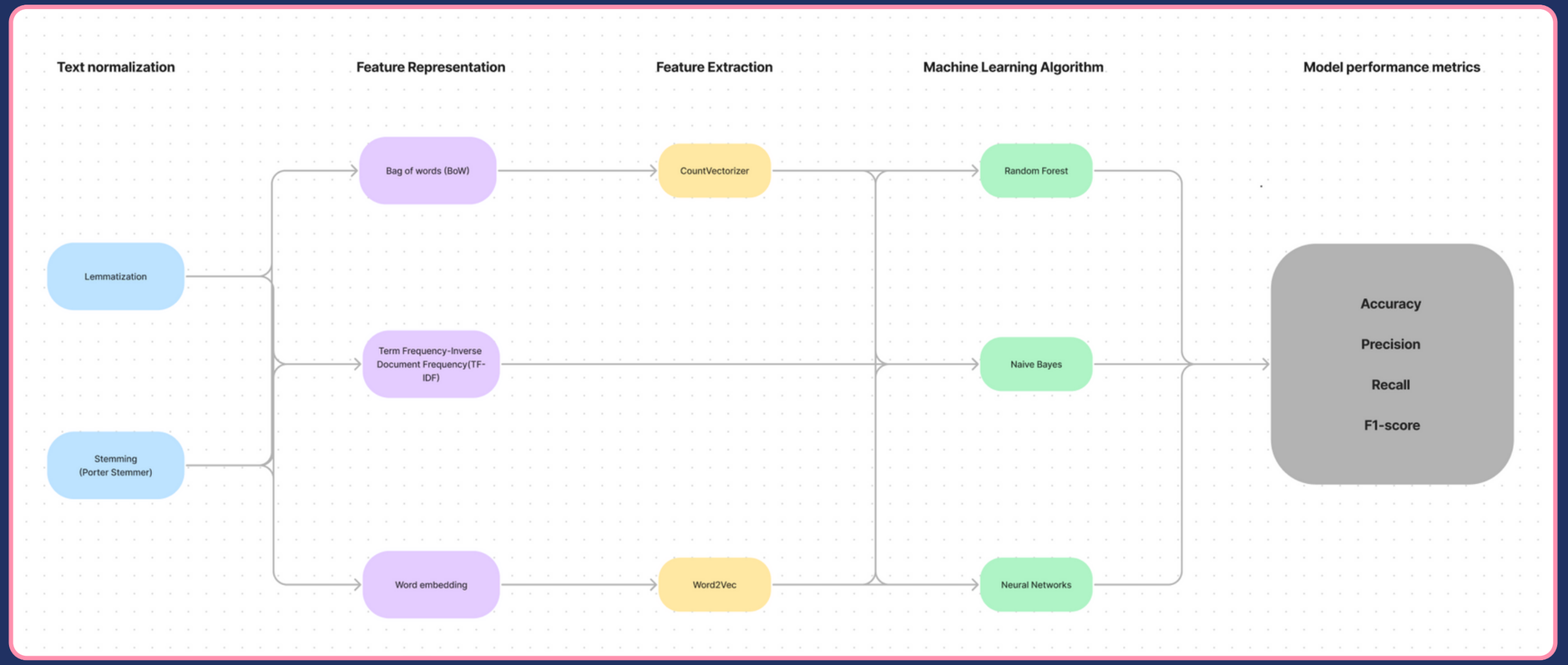
- **ML Algorithms**
  - Random Forest
  - Naive Bayes
  - Neural Networks
- **ML Techniques**
  - **KerasClassifier:** A wrapper provided by scikeras.wrappers for integrating Keras models into scikit-learn pipelines
  - **GridSearchCV:** A technique for hyperparameter tuning, used to systematically search for the best parameters

Python libraries

- **numpy:** for numerical computing with arrays
- **pandas:** for data manipulation and analysis
- **scikit-learn:** for machine learning tasks such as model selection, metrics calculation, and data preprocessing
- **tensorflow:** for building and training neural networks
- **joblib:** for saving/loading models
- **pickle:** for serializing Python objects

Building the model

18 possible models, 3 in focus



3 models in focus

	Text normalization	Feature Representation	Text vectorization	ML algorithm
Model 1	Porter Stemmer	Bag of words	Count vectorizer	Naive Bayes
Model 2	Porter Stemmer	Bag of words	Count vectorizer	Random Forest
Model 3	Porter Stemmer	Word embedding	Word2Vec	Neural networks

What's the benchmark?

This is a multi-classification task with 3 possible classifications: positive, negative, neutral

If you were to guess the sentiment of a given sentence as positive, negative, neutral at random you would guess correctly 33.3% of the time (= 33.3% accuracy)

For application in business, this model needs to exceed 80% accuracy

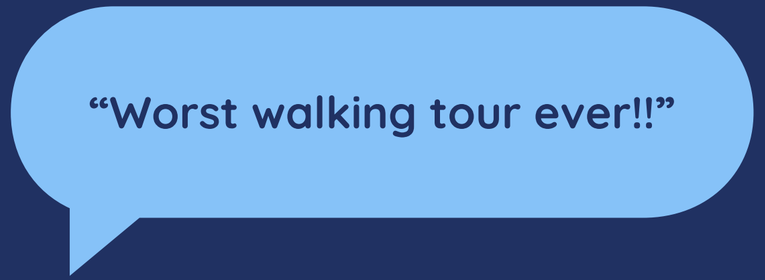
"The service at the restaurant was fabulous!"

Multi-classification scores

PRECISION =

- ratio of correctly predicted positive observations to the total predicted positives
- ratio of correctly predicted negative observations to the total predicted negatives
- ratio of correctly predicted neutral observations to the total predicted neutrals

Weighted average of 3 precision scores



Multi-classification scores

RECALL (SENSITIVITY) =

- the ratio of correctly predicted positive observations to all the truly positive observations (ability of model to identify all the positives)
- the ratio of correctly predicted negative observations to all the truly negative observations (ability of model to identify all the negatives)
- the ratio of correctly predicted neutral observations to all the truly neutral observations (ability of model to identify all the neutrals)

Weighted average of 3 recall scores

“This club plays terrible music”

Multi-classification scores

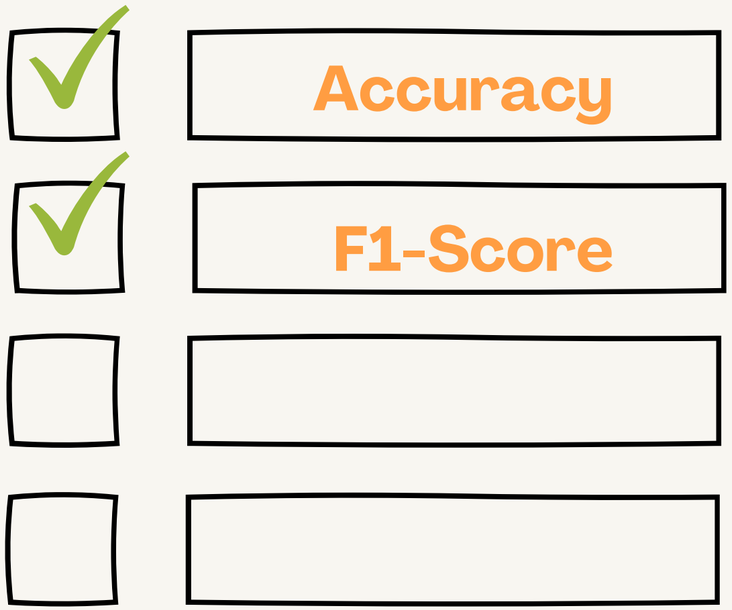
F1-SCORE =

- the harmonic mean of precision and recall

Weighted average of 3 F1-scores

Multi-classification scores

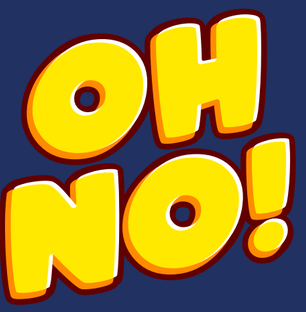
In this case, I will optimise for:

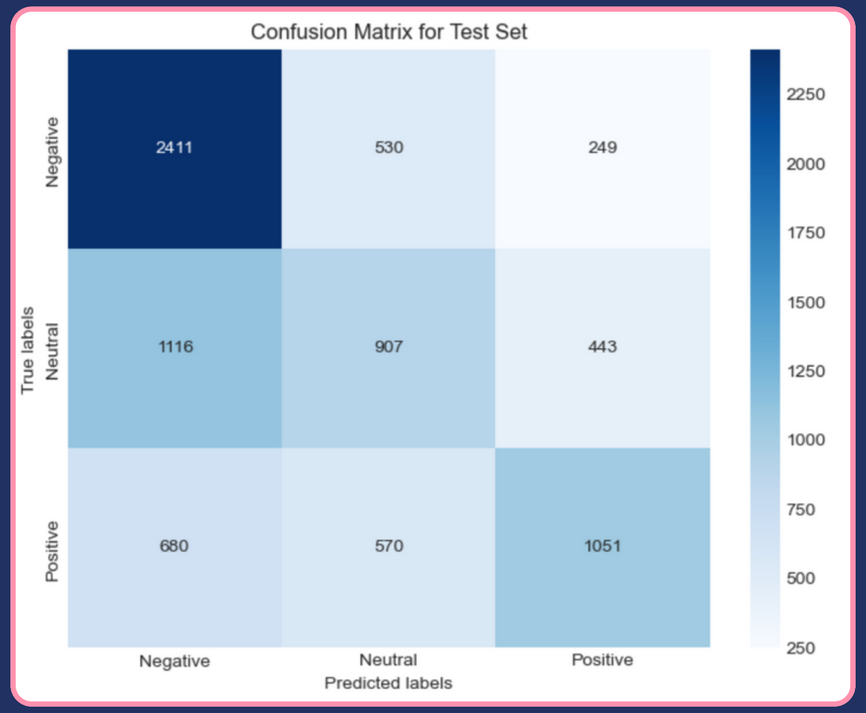


3 models in focus

	MODEL DETAILS				PERFORMANCE			
	Text normalization	Feature Representation	Text vectorization	ML algorithm	Train accuracy	Test accuracy	Train F1-score	Test F1-score
Model 1	Porter Stemmer	Bag of words	Count vectorizer	Naive Bayes	0.71	0.55	0.71	0.54
Model 2	Porter Stemmer	Bag of words	Count vectorizer	Random Forest	-	0.40	-	-
Model 3	Porter Stemmer	Word embedding	Word2Vec	Neural networks	0.48	0.45	0.45	0.41

Initial results were disappointing



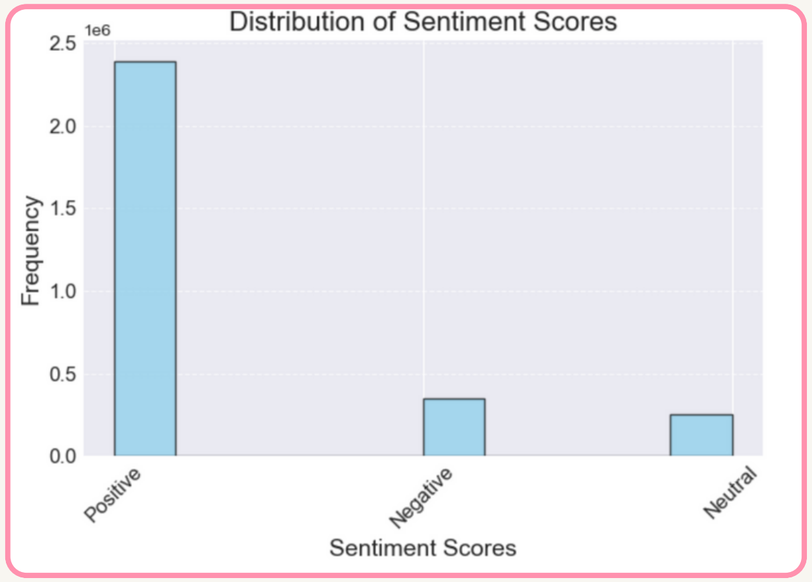


Confusion Matrix on Model 1 Test Set

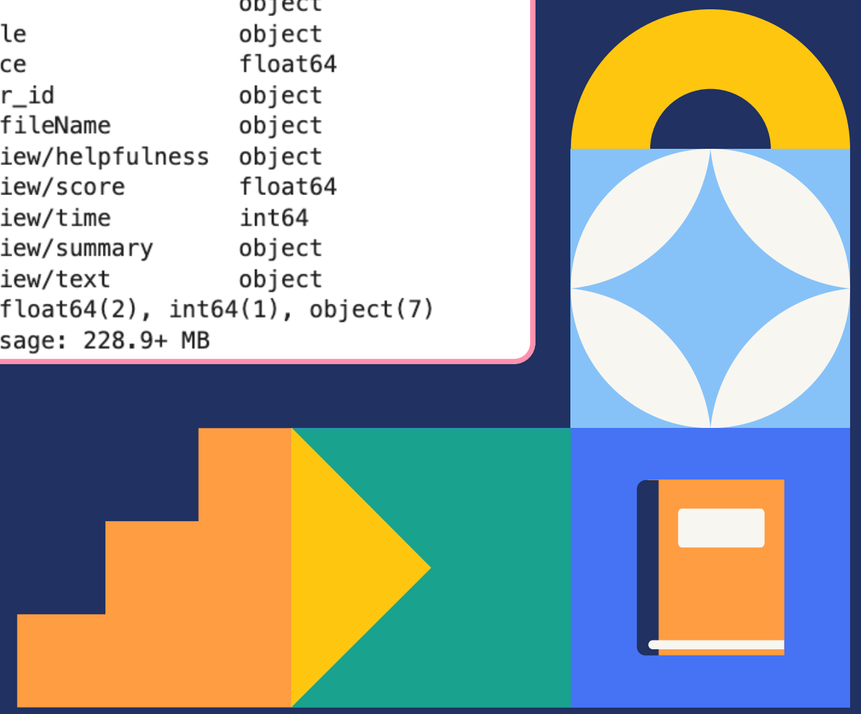
Hypothesis: The text on which the algorithm was trained was too short. By training on longer texts, we will improve our model's performance.

Training data v2

- 2 999 998 book reviews (~96 000 used for training/testing)
  - Score of 1-5
  - Text review of ~800 characters
- Source: Amazon via Kaggle
- Reviews skew positive




```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2999998 entries, 0 to 2999997
Data columns (total 10 columns):
 #   Column              Dtype
---  ------              -----
 0   Id                  object
 1   Title               object
 2   Price               float64
 3   User_id             object
 4   profileName         object
 5   review/helpfulness  object
 6   review/score        float64
 7   review/time         int64
 8   review/summary      object
 9   review/text         object
dtypes: float64(2), int64(1), object(7)
memory usage: 228.9+ MB
```



Neural network model

- computational model inspired by the structure and functioning of the human brain's neural networks
- capable of learning complex patterns and relationships from data

```python
# Define the neural network model
def create_model(optimizer='adam'):
    model = tf.keras.Sequential([
        tf.keras.layers.Dense(128, activation='relu', input_shape=(X_train.shape[1],)),
        tf.keras.layers.Dropout(0.2),
        tf.keras.layers.Dense(3, activation='softmax')  # Assuming 3 classes for sentiment_subgroup
    ])
    model.compile(optimizer=optimizer,
                  loss='sparse_categorical_crossentropy',
                  metrics=['accuracy'])
    return model

# Create a KerasClassifier object
keras_model = KerasClassifier(model=create_model, verbose=0)

# Define the hyperparameters grid
param_grid = {
    'batch_size': [16],
    'epochs': [30],
    'optimizer': ['adam'],
}
```

param_grid defined in previous iterations using GridSearchCV

- batch size = # training examples processed before the model's parameters are updated
- epochs = # times the entire training dataset is passed forward and backward through the neural network during training
- optimizer = algorithm responsible for updating the model's parameters to minimize the loss function

Neural network model

Key preprocessing steps

- **Data balancing:** RandomOverSampler on Negative and Neutral instances
- **Tokenizing:** breaking down a sequence of text into individual tokens or words

Before tokenizing

	review/text
0	This is only for Julie Strain fans. It's a col...
1	I don't care much for Dr. Seuss but after read...
2	If people become the books they read and if "t...
3	Theodore Seuss Geisel (1904-1991), aka "D...
4	Philip Nel - Dr. Seuss: American IconThis is b...

After tokenizing

	tokens
0	[This, is, only, for, Julie, Strain, fans, ., ...
1	[I, do, n't, care, much, for, Dr., Seuss, but,...
2	[If, people, become, the, books, they, read, a...
3	[Theodore, Seuss, Geisel, (, 1904-1991,), ,, ...
4	[Philip, Nel, -, Dr., Seuss, :, American, Icon...

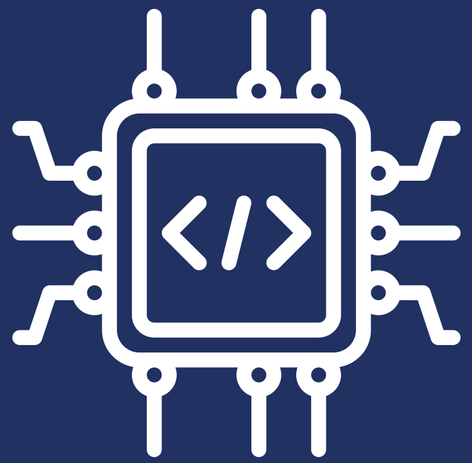
- **Token normalization:** transforming tokens into a standardized format to facilitate sentiment analysis

After token normalization

tokens
juli strain fan collect photo page worth nice ...
nt care much dr seuss read philip nel book cha...
peopl becom book read child father man dr seus...
theodor seuss geisel aka quot dr seuss quot on...
philip nel dr seuss american iconthi basic aca...

Neural network model

Key technique: word embedding



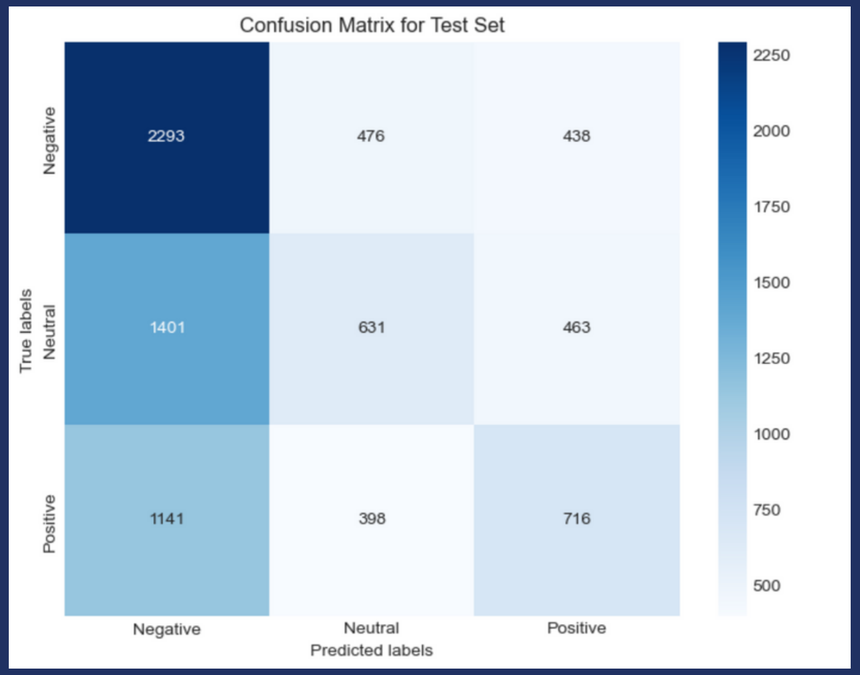
Word embedding is a technique in natural language processing (NLP) that represents words as dense vectors in a continuous vector space, capturing semantic relationships between words based on their context and meaning

tokens	word_embeddings
[This, is, only, for, Julie, Strain, fans, ., ...	[0.036981437, -0.4918617, -1.5036561, 0.653894...
[I, do, n't, care, much, for, Dr., Seuss, but,...	[-0.34708437, -0.25579467, -1.2326527, 0.35956...
[If, people, become, the, books, they, read, a...	[-0.35601878, -0.20433958, -1.0501455, 0.30435...
[Theodore, Seuss, Geisel, (, 1904-1991,), ,, ...	[-0.45737424, -0.3882762, -1.0788666, 0.477615...
[Philip, Nel, -, Dr., Seuss, :, American, Icon...	[-0.16913952, -0.43138853, -1.1637785, 0.42877...

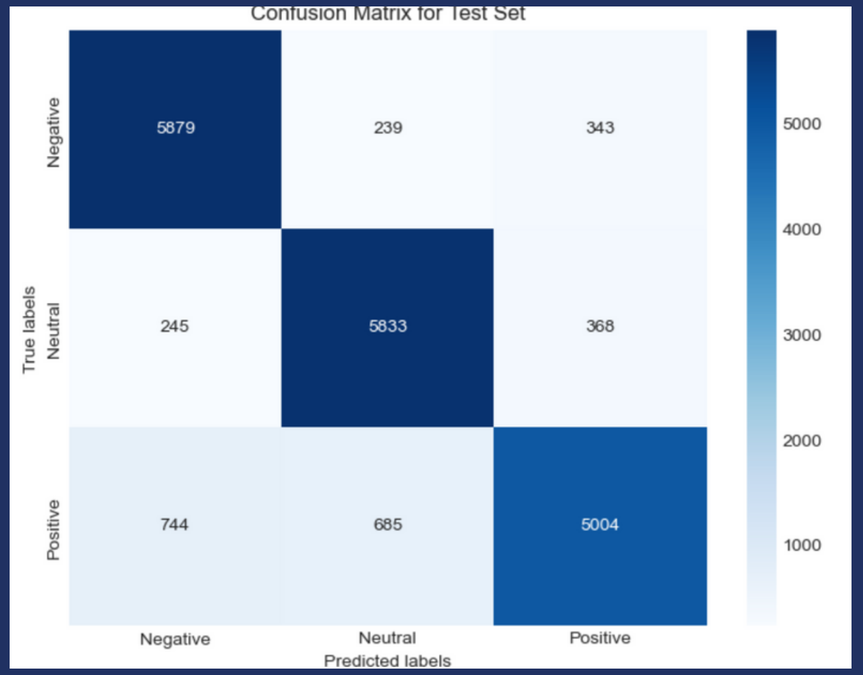
4th model (3.1) trained on longer text data

	MODEL DETAILS				PERFORMANCE			
	Text normalization	**Feature Representation**	**Text vectorization**	**ML algorithm**	**Train accuracy**	**Test accuracy**	**Train F1-score**	**Test F1-score**
Model 1	Porter Stemmer	Bag of words	Count vectorizer	Naive Bayes	0.71	0.55	0.71	0.54
Model 2	Porter Stemmer	Bag of words	Count vectorizer	Random Forest	-	0.40	-	-
Model 3	Porter Stemmer	Word embedding	Word2Vec	Neural networks	0.48	0.45	0.45	0.41
Model 3.1 NEW DATA	**Porter Stemmer**	**Word embedding**	**Word2Vec**	**Neural networks**	**0.9**	**0.86**	**0.89**	**0.86**

Performance significantly improved when trained on longer text data



Confusion Matrix on Model 3.0 Test Set
TRAINED ON TEXT = av. 73 characters



Confusion Matrix on Model 3.1 Test Set
TRAINED ON TEXT = av. 823 characters

Hypothesis: The text on which the algorithm was trained was too short. By training on longer texts, we will improve our model's performance.

Reminder: Value proposition

We empower businesses to take a ‘temperature check’ of their customer sentiment quickly and at scale

Interpret setiment of 1000s of instances of customer feedback in seconds

IMPORT

User imports (large) CSV
of customer reviews or feedback

SENTIMENT PREDICTED

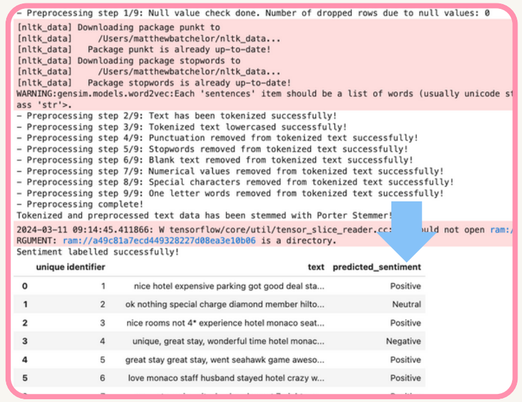
User input is preprocessed and sentiment
of each instance is predicted

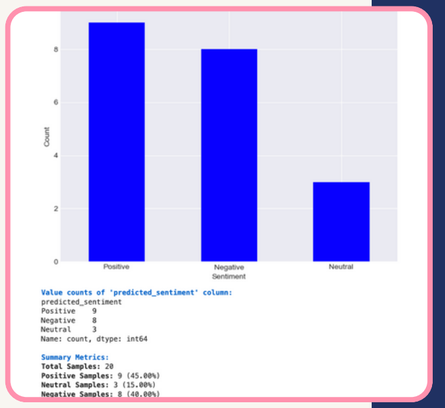
OUTPUT OF LABELLED DATA

User receives table of input data with
additional sentiment prediction label

Reminder: Value proposition

We empower businesses to take a ‘temperature check’ of their customer sentiment quickly and at scale





IMPORT

User imports (large) CSV of customer reviews or feedback

SENTIMENT PREDICTED

User input is preprocessed and sentiment of each instance is predicted

OUTPUT OF LABELLED DATA

User receives table of input data with additional sentiment prediction label

TESTING THE MODEL WITH NEW DATA ITERATION 1

```
# Make predictions on the new data
sentiment = model.predict(X_new_scaled)
```

unique identifie	text
72489	The service was exceptional! The staff was friendly and helpful, and the food was delicious. Will definitely be coming back!
53072	I had a terrible experience at your store. The staff was rude, and the product quality was poor. I won't be returning.
16947	I'm neutral about my experience at your establishment. The service was average, and the product selection was okay.
87532	The online ordering process was smooth and hassle-free. I received my package earlier than expected. Thank you!
41293	I'm disappointed with the quality of the products I received. They didn't meet my expectations.
63801	The customer service representatives were polite and knowledgeable. They assisted me in finding exactly what I needed.
29746	The product exceeded my expectations. It's high-quality and durable. I'm extremely satisfied with my purchase.
85460	I'm undecided about whether I'll return to your store. The prices were reasonable, but the selection was limited.
36109	The delivery was prompt, and the items were well-packaged. I couldn't be happier with my purchase.
50283	The checkout process was seamless, and I appreciated the variety of payment options available.
94670	The atmosphere of your establishment is cozy and inviting. It's the perfect place to relax and unwind after a long day.
20835	I encountered an issue with my order, but the customer service team resolved it promptly and courteously.
71948	The prices are reasonable, and the quality of the products is excellent. I'll definitely be a repeat customer.
58201	The staff went above and beyond to ensure that my experience was enjoyable. I highly recommend this establishment.
34576	I received personalized attention from the staff, which made me feel valued as a customer. Thank you!
90128	The ambiance of your establishment is delightful. It's the perfect spot for a romantic dinner or a casual lunch.
46739	I'm dissatisfied with the level of customer service provided. The staff seemed indifferent to my concerns.

Neural Network

20 AI generated customer reviews

Av characters per review: 109

OUTPUT
WITH NEW DATA ITERATION 1

Neural Network

	unique identifier	text	predicted_sentiment
0	72489	The service was exceptional! The staff was fri...	Negative
1	53072	I had a terrible experience at your store. The...	Positive
2	16947	I'm neutral about my experience at your establ...	Neutral
3	87532	The online ordering process was smooth and has...	Neutral
4	41293	I'm disappointed with the quality of the produ...	Positive
5	63801	The customer service representatives were poli...	Positive
6	29746	The product exceeded my expectations. It's hig...	Positive
7	85460	I'm undecided about whether I'll return to you...	Negative
8	36109	The delivery was prompt, and the items were we...	Positive
9	50283	The checkout process was seamless, and I appre...	Negative
10	94670	The atmosphere of your establishment is cozy a...	Positive
11	20835	I encountered an issue with my order, but the ...	Positive
12	71948	The prices are reasonable, and the quality of ...	Negative
13	58201	The staff went above and beyond to ensure that...	Positive
14	34576	I received personalized attention from the sta...	Negative
15	90128	The ambiance of your establishment is delightf...	Positive
16	46739	I'm dissatisfied with the level of customer se...	Negative
17	81254	The checkout process was confusing, and I had ...	Positive
18	63597	The food was fresh and flavorful. It's evident...	Negative
19	17480	I'm on the fence about whether I'll shop here ...	Positive

Output of first test with new customer data

TESTING THE MODEL WITH NEW DATA ITERATION 2

```
# Make predictions on the new data
sentiment = model.predict(X_new_scaled)
```

Neural Network

text
nice hotel expensive parking got good deal stay hotel anniversary, arrived late evening took advice previous reviews did valet parking, check quick easy, little disappointed non-existent vie
ok nothing special charge diamond member hilton decided chain shot 20th anniversary seattle, start booked suite paid extra website description not, suite bedroom bathroom standard h
nice rooms not 4* experience hotel monaco seattle good hotel n't 4* level.positives large bathroom mediterranean suite comfortable bed pillowsattentive housekeeping staffnegatives ac
unique, great stay, wonderful time hotel monaco, location excellent short stroll main downtown shopping area, pet friendly room showed no signs animal hair smells, monaco suite sleepi
great stay great stay, went seahawk game awesome, downfall view building did n't complain, room huge staff helpful, booked hotels website seahawk package, no charge parking got vou
love monaco staff husband stayed hotel crazy weekend attending memorial service best friend husband celebrating 12th wedding anniversary, talk mixed emotions, booked suite hotel m
cozy stay rainy city, husband spent 7 nights monaco early january 2008. business trip chance come ride.we booked monte carlo suite proved comfortable longish stay, room 905 located st
excellent staff, housekeeping quality hotel chocked staff make feel home, experienced exceptional service desk staff concierge door men maid service needs work, maid failed tuck sheets
hotel stayed hotel monaco cruise, rooms generous decorated uniquely, hotel remodeled pacific bell building charm sturdiness, everytime walked bell men felt like coming home, secure, g
excellent stayed hotel monaco past w/e delight, reception staff friendly professional room smart comfortable bed, particularly liked reception small dog received staff guests spoke loved,
poor value stayed monaco seattle july, nice hotel priced 100- 150 night not, hotel takes beating quotient, experience simply average, nothing exceptional paying 300+ n't ca n't terribly dis
nice value seattle stayed 4 nights late 2007. looked comparable hilton marriott westin area points/miles n't gave monaco shot, pleasantly surprised nice room service quick tasty bed espe
nice hotel good location hotel kimpton design whimsical vibe fun, staff young casual problem hotel busy stay friendly helpful, group reserved rooms gave connecting rooms fuss, not busy
nice hotel not nice staff hotel lovely staff quite rude, bellhop desk clerk going way make things difficult, waited forever check heavy bags no help getting throught double doors room, wor
great hotel night quick business trip, loved little touches like goldfish leopard print robe, complaint wifi complimentary not internet access business center, great location library service fa
horrible customer service hotel stay february 3rd 4th 2007my friend picked hotel monaco appealing website online package included champagne late checkout 3 free valet gift spa weeke
disappointed say anticipating stay hotel monaco based reviews seen tripadvisor, definitely disppointment, decor room hotel envisioned nice, housekeeping staff impressive extremely poli
fantastic stay monaco seattle hotel monaco holds high standards kimpton hotel line, having stayed kimpton hotels cities easily say seattle hotel monaco best seen, service attentive promp
good choice hotel recommended sister, great location room nice, comfortable bed- quiet- staff helpful recommendations restaurants, pike market 4 block walk, stay,
hmmmmm say really high hopes hotel monaco chose base girlfriend shopping trip seattle, stay say given competition seattle just okay, hotel lot nice features little things detract bedding

20 hotel reviews

Av characters per review: 790

OUTPUT
WITH NEW DATA ITERATION 2

Neural Network

	unique identifier	text	predicted_sentiment
0	1	nice hotel expensive parking got good deal sta...	Positive
1	2	ok nothing special charge diamond member hilto...	Negative
2	3	nice rooms not 4* experience hotel monaco seat...	Positive
3	4	unique, great stay, wonderful time hotel monac...	Negative
4	5	great stay great stay, went seahawk game aweso...	Negative
5	6	love monaco staff husband stayed hotel crazy w...	Positive
6	7	cozy stay rainy city, husband spent 7 nights m...	Positive
7	8	excellent staff, housekeeping quality hotel ch...	Negative
8	9	hotel stayed hotel monaco cruise, rooms genero...	Positive
9	10	excellent stayed hotel monaco past w/e delight...	Neutral
10	11	poor value stayed monaco seattle july, nice ho...	Negative
11	12	nice value seattle stayed 4 nights late 2007. ...	Negative
12	13	nice hotel good location hotel kimpton design ...	Positive
13	14	nice hotel not nice staff hotel lovely staff q...	Positive
14	15	great hotel night quick business trip, loved l...	Neutral
15	16	horrible customer service hotel stay february ...	Positive
16	17	disappointed say anticipating stay hotel monac...	Negative
17	18	fantastic stay monaco seattle hotel monaco hol...	Neutral
18	19	good choice hotel recommended sister, great lo...	Positive
19	20	hmmmmm say really high hopes hotel monaco chos...	Negative

Output of second test with new customer data

Why might a model with supposed accuracy of 86% deliver such poor results in subsequent tests?

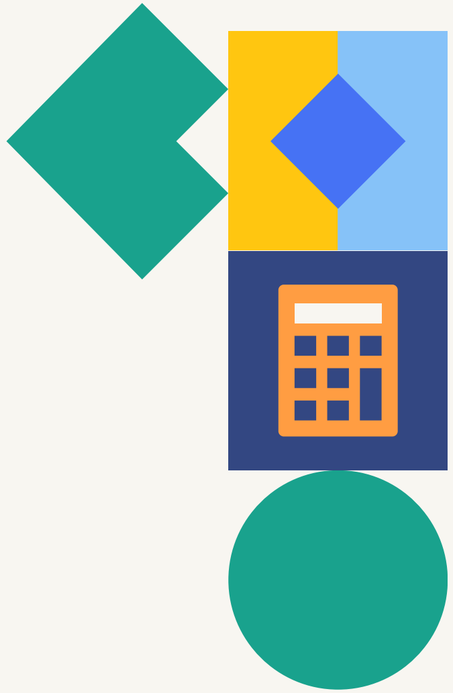
A few possible reasons that we may be observing this outcome

1 TO INVESTIGATE

- Data leakage during training
- Imbalanced classes (RandomOverSampler)
- Unseen patterns potentially not captured in training
- Limited generalization - trained on book reviews, predicting sentiment on hotel/restaurant reviews

2 RULED OUT

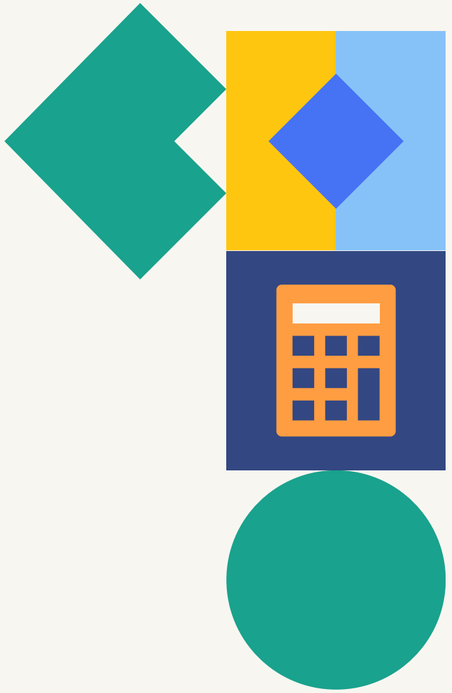
- Overfitting - would have shown up in train-test-split
- Hyperparameter sensitivity - tuning performed
- Data drift - not at play here



How will I develop this model?

1. Improve accuracy of model with new data input, especially with text input of shorter length

2. Improve user interface (Streamlit)



Thank you for being part of this journey!

Matthew Batchelor
IronHack Data Analytics student
Presentation delivered March 11th 2024