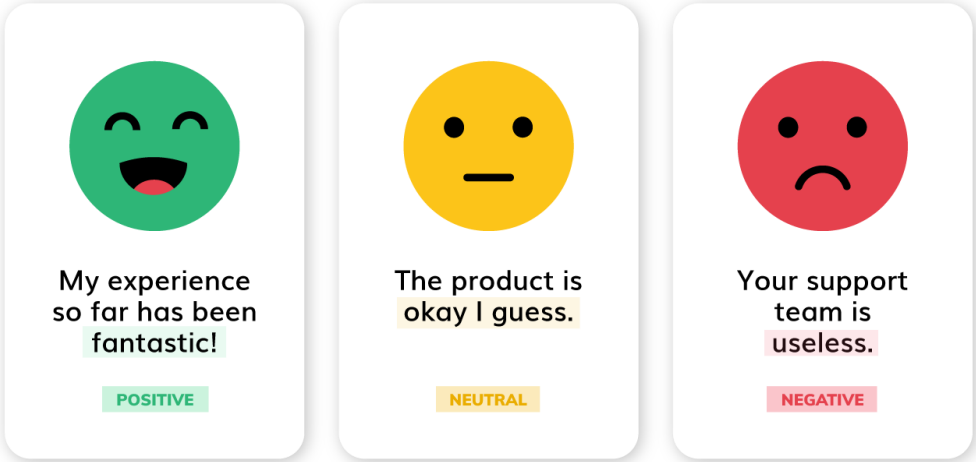
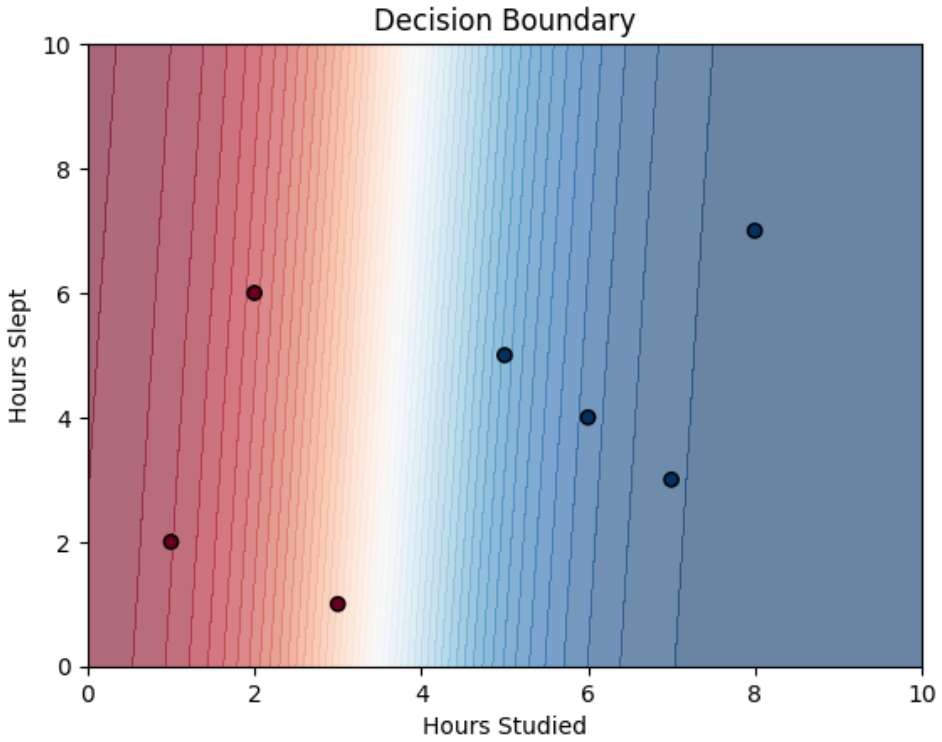
Part 2: Text Sentiment Classification

SUMMER 2025 STARTER-AI HANDS-ON WORKSHOP SERIES

The University of Texas Rio Grande Valley





Classification on Text

Task: We want to learn a decision boundary to decide if text is positive or negative

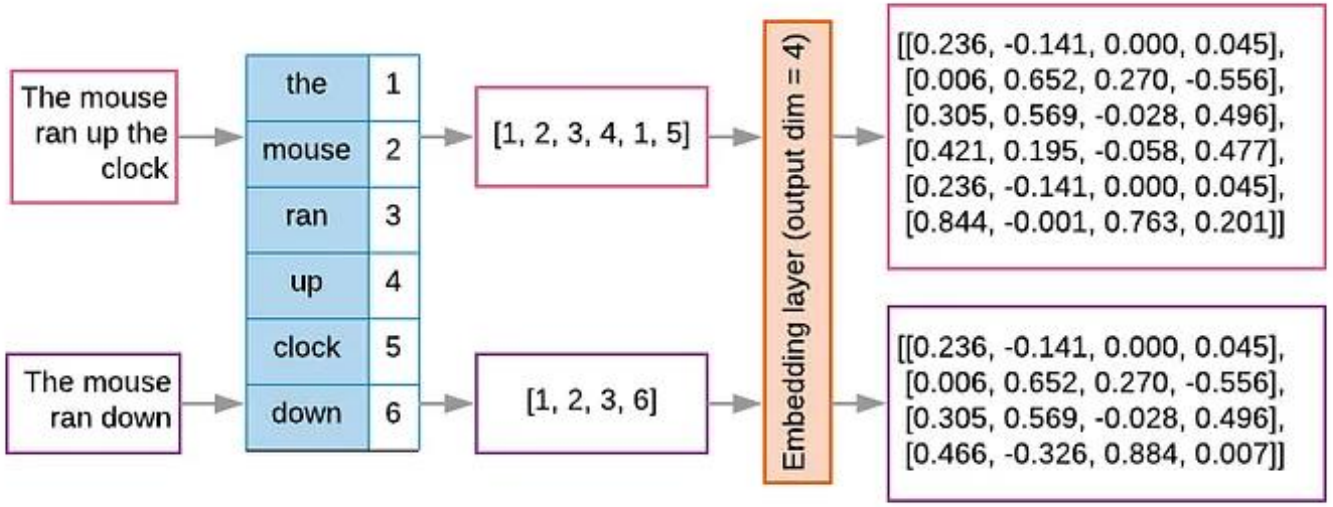
- Great! This is much more flexible. Applicable to anything involving language/opinions

Problem: Text is not numerical, so it does not match a format our model can understand!

Solution: Convert text into numerical representations

https://centricconsulting.com/wp-content/uploads/2022/08/Blog_DA_Sentiment_Customer_08052022.png

Convert Text to Numbers



This is called "Embedding" in Machine learning-based Natural Language Processing (NLP) tasks.

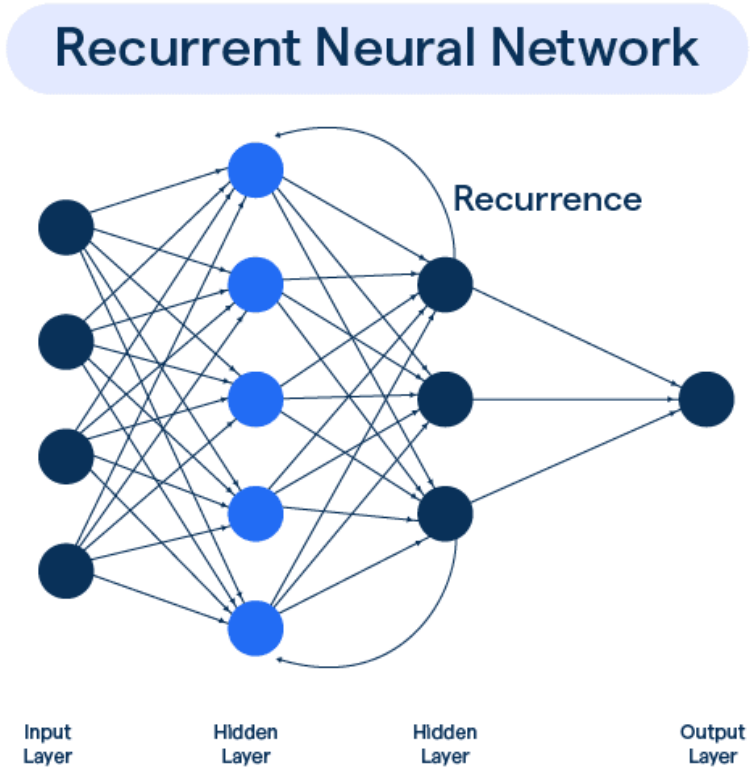
There are many approaches, but let's focus on a simple example

Steps:

- "Tokenize" text (e.g. into words)
- Assign IDs to tokens
- Convert token IDs into embedded vectors to be fed into a model
  - Could be one-hot encoded, but also learned

https://airbyte.com/data-engineering-resources/tokenization-vs-embeddings

	about	bird	heard	is	the	word	you
About the bird, the bird, bird bird bird	1	5	0	0	2	0	0
You heard about the bird	1	1	1	0	1	0	1
The bird is the word	0	1	0	1	2	1	0



Three Approaches

We will look at the following 3 models:

- "Bag-of-words" + Linear classifier
  - Will skip the learned embedding step, and will simply keep a count of words present in a statement
- Embeddings + Mean-pooling
  - Mean-pooling takes the average of the learned embedding of all words in a statement
- Gated-Recurrent Unit (GRU) Network
  - Will account for the order of words

https://cdn.botpenguin.com/assets/website/Recurrent_Neural_Network_508b372642.png

Task: Sentiment Classification of Yelp Reviews

Dataset available through HuggingFace: ”Yelp Polarity”

Includes text of reviews along with a label indicating a positive or negative review

Example:

Text:

- `Wing sauce is like water. Pretty much a lot of butter and some hot sauce (franks red hot maybe). The whole wings are good size and crispy, but for $1 a wing the sauce could be better. The hot and extra hot are about the same flavor/heat. The fish sandwich is good and is a large portion, sides are decent.`

label:

- `0 {negative}`

Yelp dataset info

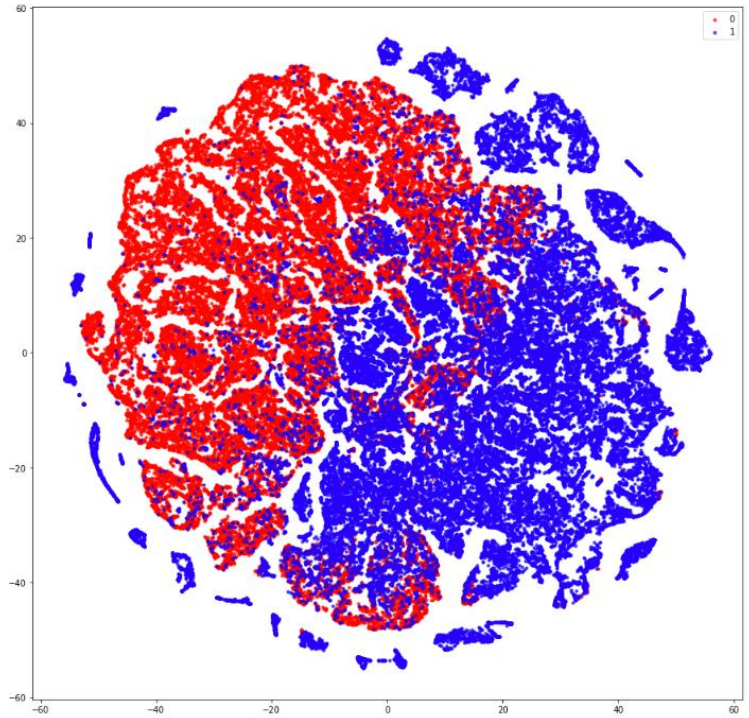
Visualization with t-SNE

t-SNE is is a tool that helps us visualize high-dimensional data — like the embedded sentence vectors the model produces

The model embeds statements into high-dimensional vector that is impossible to visualize directly.

- Here it's 50D for mean-pooling and 64D for GRU

t-SNE helps us project those complex vectors down into 2D so we can see how sentences are grouped.



https://i.sstatic.net/tZWGT.png

Training Metrics

To assess our models, we will use two metrics

- Accuracy (how often is the model correct?)
- F1-score
  - takes into account precision and recall to give a better picture of performance
  - **Precision example:** if the model predicts it is positive, how likely is it to be positive?
  - **Recall example:** of the total tested positive examples, what percentage does the model correctly identify as positive?

The results might be surprising!

Time to Train!

Let’s work on our classification approaches