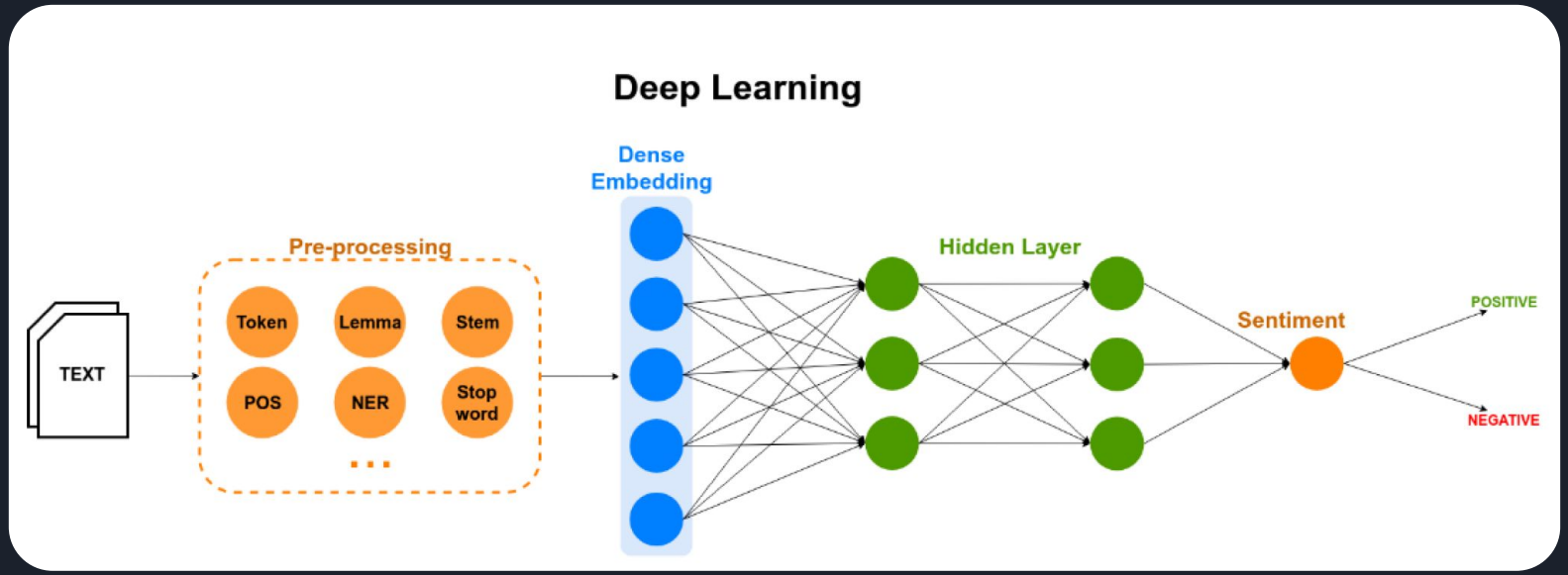
COMP 6630: Project Proposal Group 6

Matthew Freestone*
Will Humphlett
Matthew Shiplett

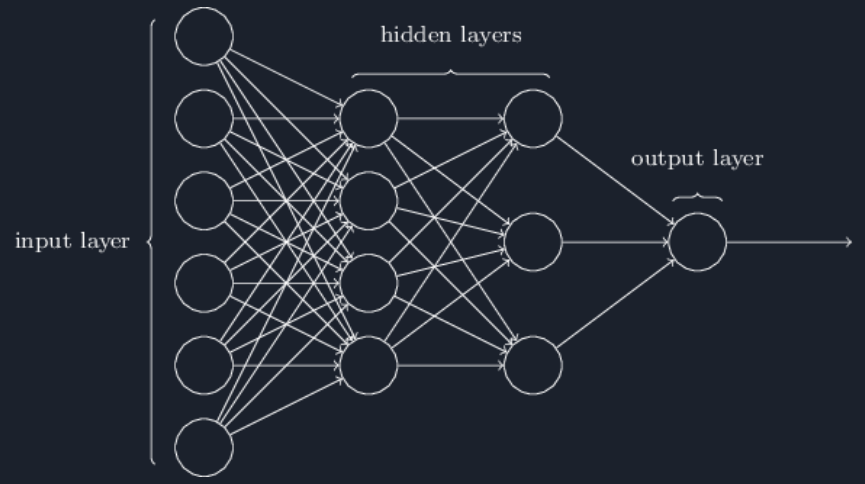
Define the problem.

- Multilayer Perceptron with from-scratch backpropagation
- Sentiment Analysis
- Given a text input, determine sentiment in {-1:, negative, 1: positive}



How can MLP be applied to solve the problem?

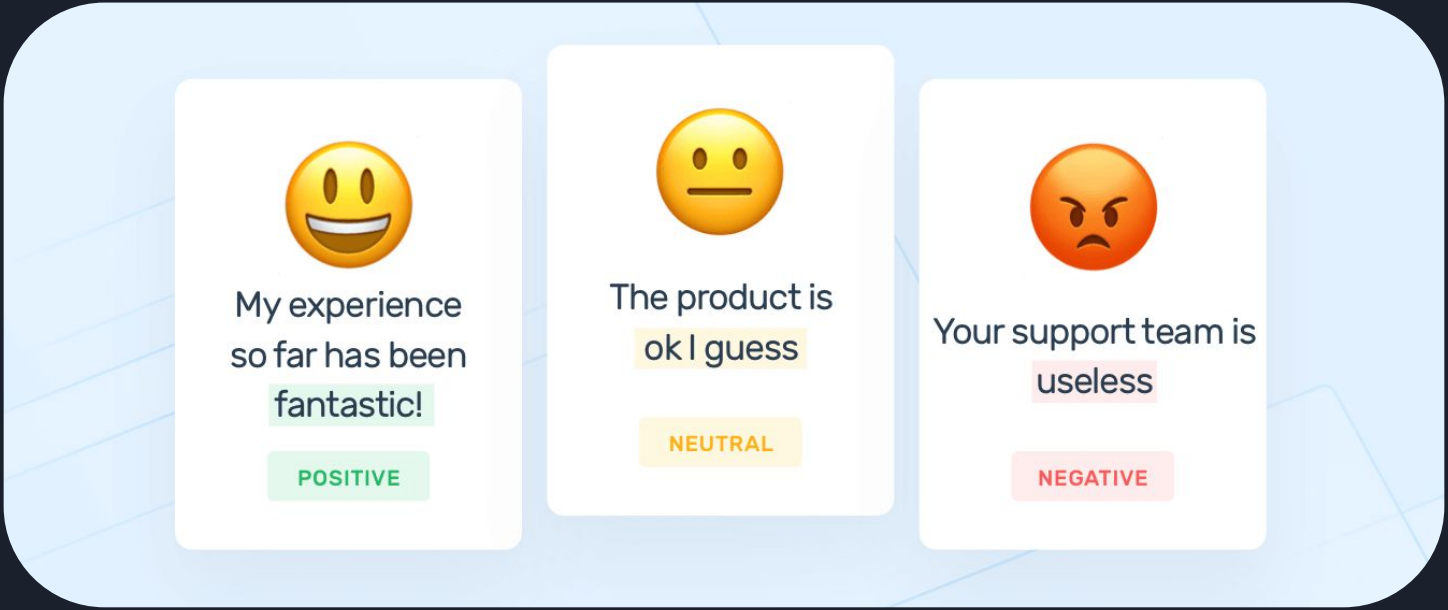
- Highly non-linear decision plane
  - Combination of perceptron units allows this
- Backpropagation
- Stochastic Gradient Descent
  - Allows dealing with datasets too large to fit in memory



Users of the classifier

Data scientists: Label data quickly and accurately

Corporate Entities: Analysis on feedback or online discussions of company



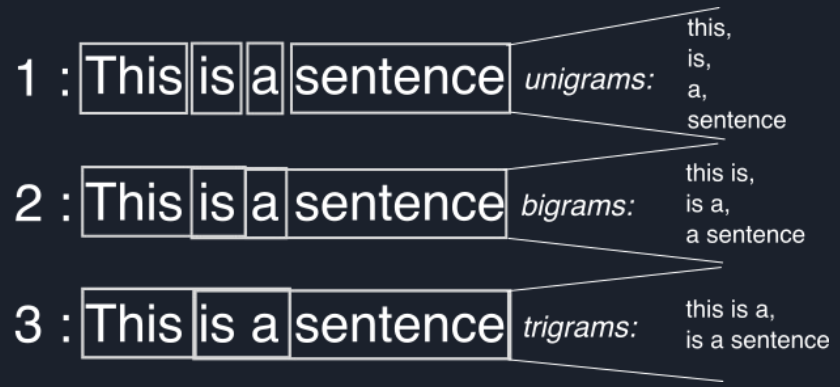
Potential Challenges

Large input set, so very large first layer

- Computationally Expensive
  - GPU vs CPU
- Memory Limitations

How should text be represented?

- Word count vectors
- N-Grams
- TF-IDF



	the	cat	sat	on	hat	dog	ate	and
Document 1	2	1	1	1	1	0	0	0
Document 2	3	1	0	0	1	1	1	1

Dataset

Kaggle “Highly Polar”Labeled Dataset:
https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews

review	sentiment
49582 unique values	2 unique values
One of the other reviewers has mentioned that after watching just 1 Oz episode you'll be hooked. The...	positive

For neutral points, we can web scrape IMDB for reviews with 5 - 7 stars

Implementation: Technologies, and Libraries

Version Control: Github

Language: Python

Implementation Libraries: Numpy, Pandas

Benchmarking: sklearn, PyTorch

Hyperparameters and Tuning

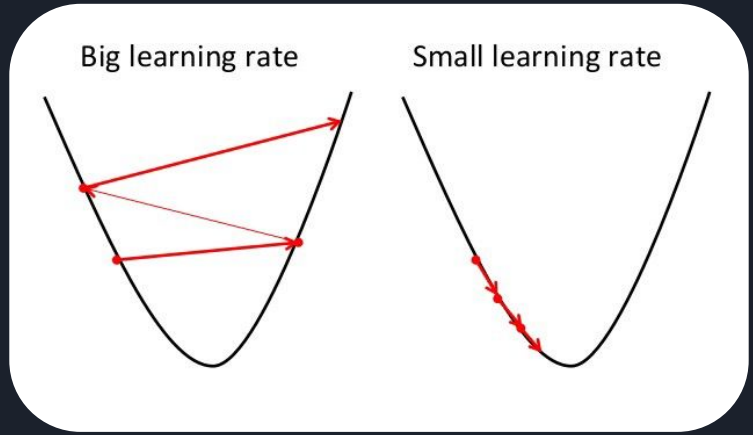
Hyperparameters

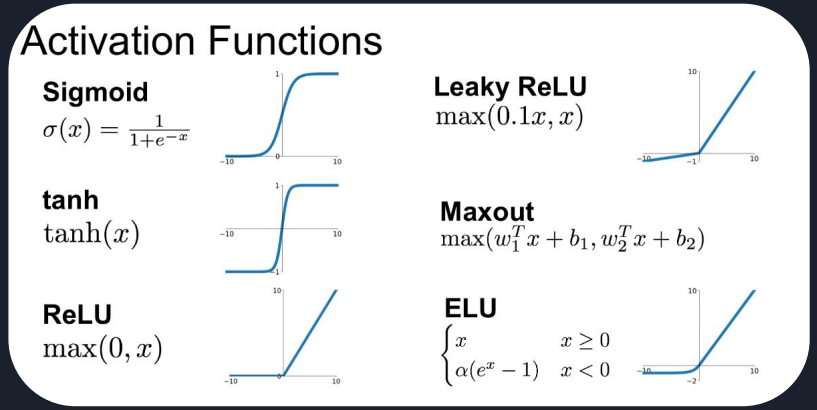
Number of hidden layers and size

Learning Rate

Activation Function

SGD Batch Size





Validation

K-fold Cross Validation

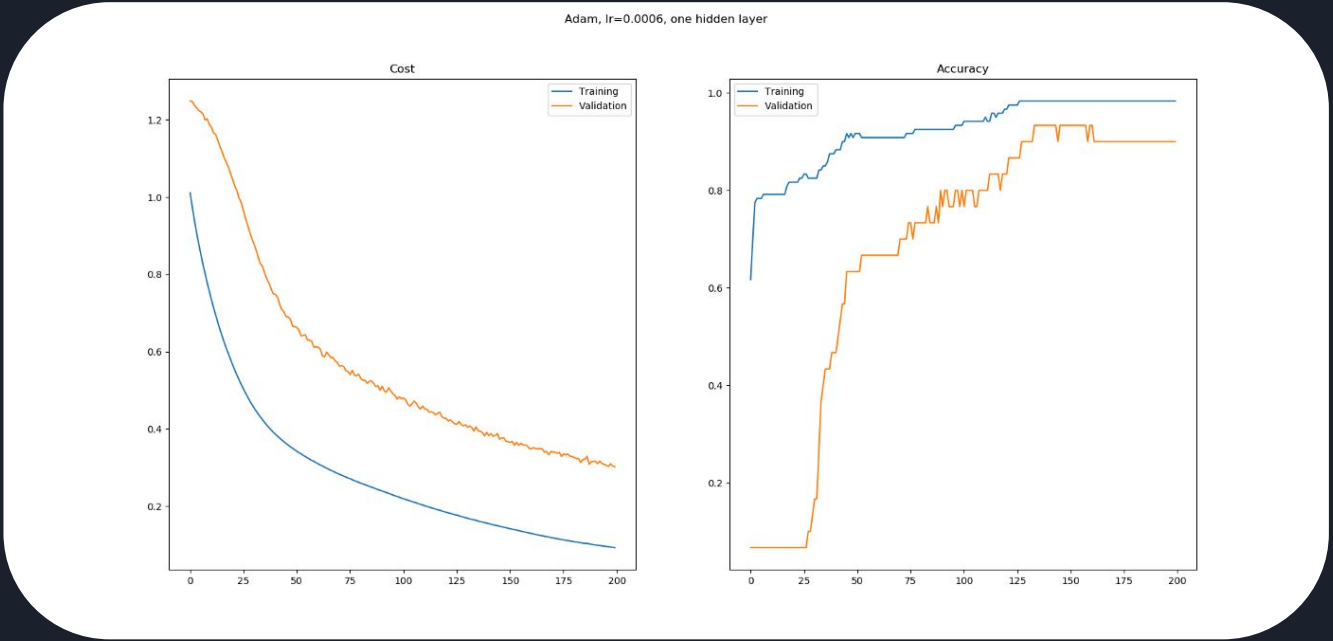
Grid Search

Random Search

Demonstrating Usefulness

Perform sentiment analysis on a large body of unlabeled text

Avoid overfitting on movie-related topics



Rough Timeline

Project Proposal - Due : October 26th, 2022

Initial Commit of Repository - Expected : October 26th,2022

Proposal Feedback - Expected : After October 26th, 2022

Dataset acquisition and pre-processing - Expected November 7th, 2022

Initial implementation of MLP - Expected : November 9th,2022

Initial training run of MLP - Expected : November 10th,2022

Verification and recording of results - Expected : November 11th, 2022

Initialization of Benchmark implementation(s) - Expected November 11th, 2022

Verification and recording of benchmark results - Expected November 15th, 2022

Creation of Final report documentation - Expected November 16th, 2022

Finalization of Final report documentation - Expected November 18th, 2022

Rehearsals of Final report - Expected November 28th -December 1st, 2022

Project presentation - Expected December 1st, 202