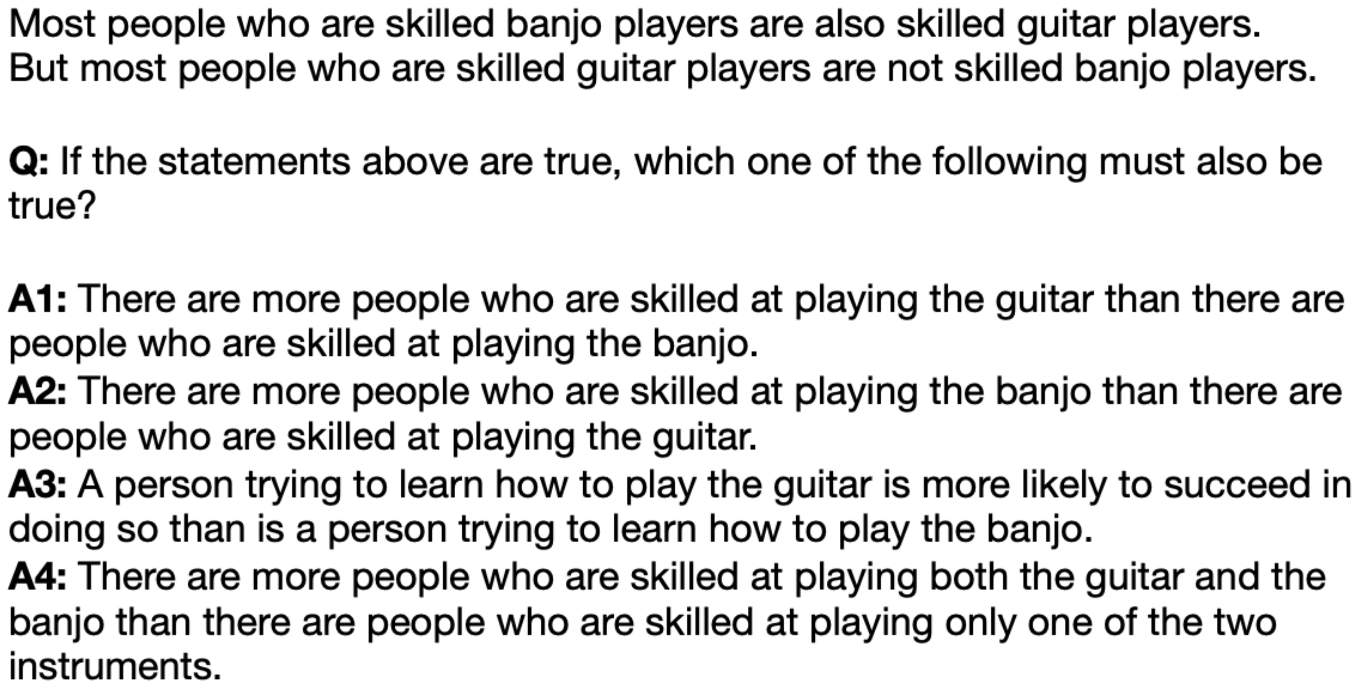
Project 1: Reading comprehension with logic
Multi-class text classification



Most people who are skilled banjo players are also skilled guitar players.
But most people who are skilled guitar players are not skilled banjo players.

Q: If the statements above are true, which one of the following must also be true?

A1: There are more people who are skilled at playing the guitar than there are people who are skilled at playing the banjo.
A2: There are more people who are skilled at playing the banjo than there are people who are skilled at playing the guitar.
A3: A person trying to learn how to play the guitar is more likely to succeed in doing so than is a person trying to learn how to play the banjo.
A4: There are more people who are skilled at playing both the guitar and the banjo than there are people who are skilled at playing only one of the two instruments.

https://www.kaggle.com/t/9027d6a7c14f45d2a8d07dd583543634

Th

White throated Sparrow

Green Jay

White breasted kingfisher

Yellow bellied flycatcher

Final project 2: Feather in Focus

Classifying images of bird species

https://www.kaggle.com/t/3ff6add39118421b81c42865b7e54809

Project 3: Unified tabular learning

Learning to classify multiple tabular datasets

Forest Cover Dataset

Credit Card Fraud Dataset

Bank Marketing Dataset

https://www.kaggle.com/t/1647fea77ea84eea9d8b6e062959c8e0

Teams and timeline

Each team consists of four people.

Final deliverable: poster with main outcomes, findings, novelties, etc.

Each poster is graded by two people.

Lab sessions vital for progress! Communicate often with many TAs.

Spam Filter

Giulia Donker: g.i.e.donker@student.vu.nl
Roos Slingerland: roos.slingerland@student.uva.nl

The Project

Previous research

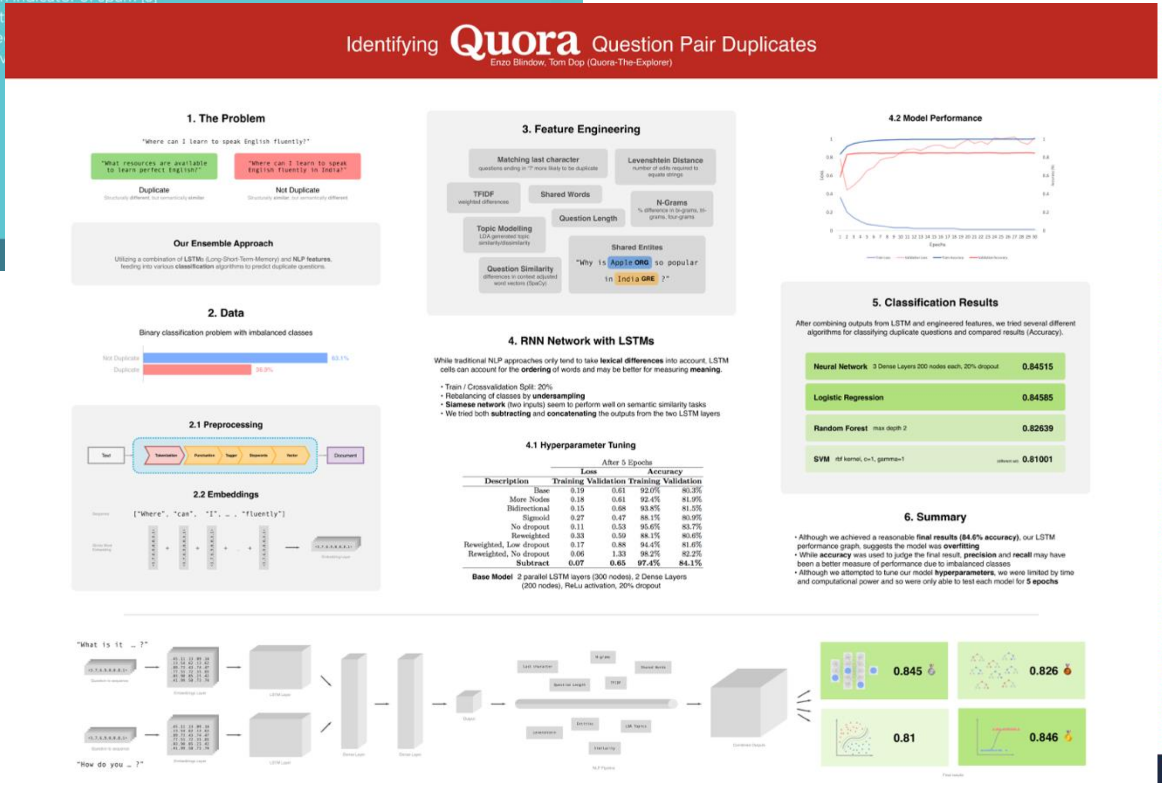
- Spam: sell product or services to customers available on the internet via email, also bulk-email [7]
- Because of the increase of email use, bulk-email increased as well [4]
- Research is often done, but spam keeps developing [4] and labelled data is often an issue [7]
- Lenght could be an indicator of spam [5]
- Metadata such as t
- Mail is often forme
- Decision trees prov field [7, 8]

About the data

4021 training examples

24% spam

76% not spam



Identifying Quora Question Pair Duplicates

Enzo Blindow, Tom Dop (Quora-The-Explorer)

1. The Problem

"Where can I learn to speak English fluently?"

"What resources are available to learn perfect English?"
Duplicate
Structurally different, but semantically similar

"Where can I learn to speak English fluently in India?"
Not Duplicate
Structurally similar, but semantically different

Our Ensemble Approach

Utilizing a combination of **LSTMs** (Long-Short-Term-Memory) and **NLP features**, feeding into various **classification** algorithms to predict duplicate questions.

2. Data

Binary classification problem with imbalanced classes

2.1 Preprocessing

2.2 Embeddings

["Where", "can", "I", … , "fluently"]

3. Feature Engineering

Matching last character questions ending in '?' more likely to be duplicate

Levenshtein Distance number of edits required to equate strings

TFIDF weighted differences

Shared Words

N-Grams % difference in bi-grams, tri-grams, four-grams

Question Length

Topic Modelling LDA generated topic similarity/dissimilarity

Shared Entites "Why is Apple ORG so popular in India GRE ?"

Question Similarity differences in context adjusted word vectors (SpaCy)

4. RNN Network with LSTMs

While traditional NLP approaches only tend to take **lexical differences** into account, LSTM cells can account for the **ordering** of words and may be better for measuring **meaning**.

- Train / Crossvalidation Split: 20%
- Rebalancing of classes by **undersampling**
- **Siamese network** (two inputs) seem to perform well on semantic similarity tasks
- We tried both **subtracting** and **concatenating** the outputs from the two LSTM layers

4.1 Hyperparameter Tuning

	After 5 Epochs			
	Loss		Accuracy	
Description	Training	Validation	Training	Validation
Base	0.19	0.61	92.0%	80.3%
More Nodes	0.18	0.61	92.4%	81.9%
Bidirectional	0.15	0.68	93.8%	81.5%
Sigmoid	0.27	0.47	88.1%	80.9%
No dropout	0.11	0.53	95.6%	83.7%
Reweighted	0.33	0.59	88.1%	80.6%
Reweighted, Low dropout	0.17	0.88	94.4%	81.6%
Reweighted, No dropout	0.06	1.33	98.2%	82.2%
Subtract	**0.07**	**0.65**	**97.4%**	**84.1%**

Base Model 2 parallel LSTM layers (300 nodes), 2 Dense Layers (200 nodes), ReLu activation, 20% dropout

4.2 Model Performance

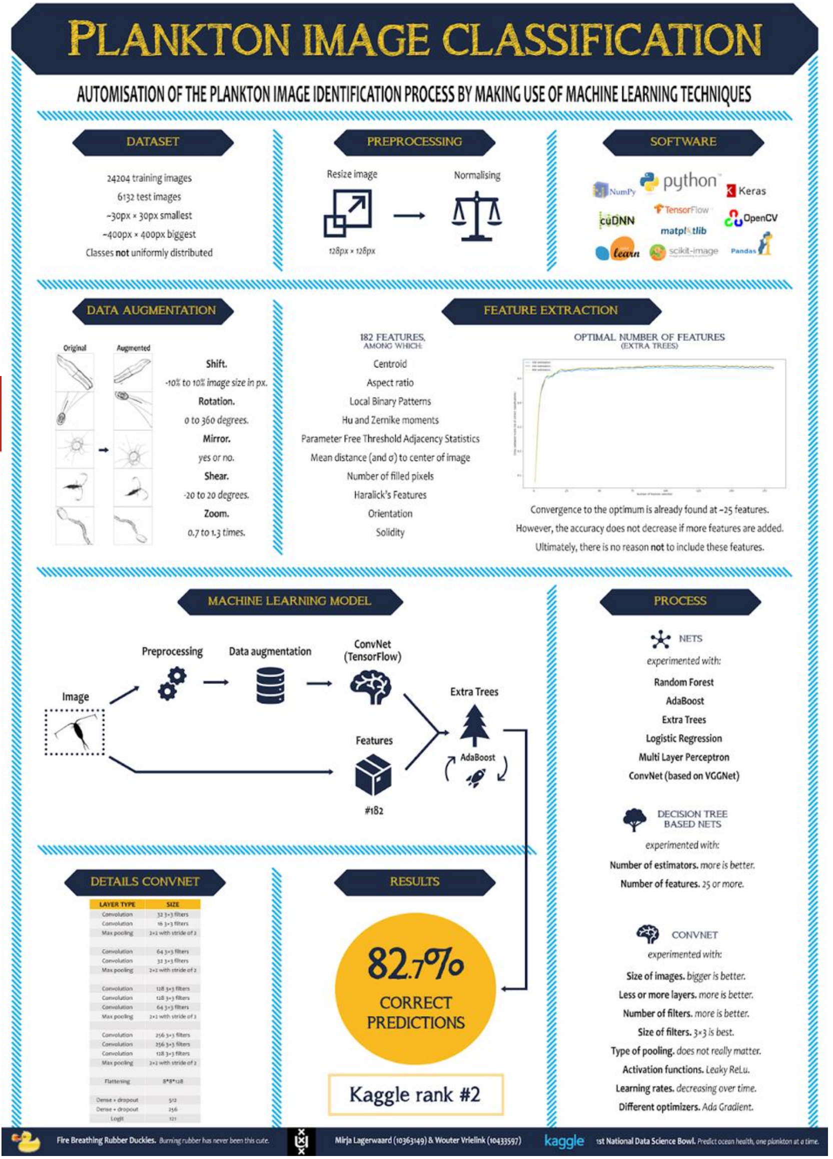
5. Classification Results

After combining outputs from LSTM and engineered features, we tried several different algorithms for classifying duplicate questions and compared results (Accuracy).

Neural Network	3 Dense Layers 200 nodes each, 20% dropout	**0.84515**
Logistic Regression		**0.84585**
Random Forest	max depth 2	**0.82639**
SVM	rbf kernel, c=1, gamma=1	**0.81001**

6. Summary

- Although we achieved a reasonable **final results (84.6% accuracy)**, our LSTM performance graph, suggests the model was **overfitting**
- While **accuracy** was used to judge the final result, **precision** and **recall** may have been a better measure of performance due to imbalanced classes
- Although we attempted to tune our model **hyperparameters**, we were limited by time and computational power and so were only able to test each model for **5 epochs**



Plankton Image Classification

Automisation of the plankton image identification process by making use of machine learning techniques

Dataset

24204 training images
6132 test images
~30px × 30px smallest
~400px × 400px biggest
Classes **not** uniformly distributed

Preprocessing

Resize image
128px × 128px
Normalising

Software

Data Augmentation

Shift. -10% to 10% image size in px.
Rotation. 0 to 360 degrees.
Mirror. yes or no.
Shear. -20 to 20 degrees.
Zoom. 0.7 to 1.3 times.

Feature Extraction

182 FEATURES,
AMONG WHICH:

Centroid
Aspect ratio
Local Binary Patterns
Hu and Zernike moments
Parameter Free Threshold Adjacency Statistics
Mean distance (and σ) to center of image
Number of filled pixels
Haralick's Features
Orientation
Solidity

OPTIMAL NUMBER OF FEATURES
(EXTRA TREES)

Convergence to the optimum is already found at ~25 features.
However, the accuracy does not decrease if more features are added.
Ultimately, there is no reason **not** to include these features.

Machine Learning Model

Process

NETS
experimented with:
Random Forest
AdaBoost
Extra Trees
Logistic Regression
Multi Layer Perceptron
ConvNet (based on VGGNet)

DECISION TREE BASED NETS
experimented with:
Number of estimators. *more is better.*
Number of features. *25 or more.*

CONVNET
experimented with:
Size of images. *bigger is better.*
Less or more layers. *more is better.*
Number of filters. *more is better.*
Size of filters. *3×3 is best.*
Type of pooling. *does not really matter.*
Activation functions. *Leaky ReLu.*
Learning rates. *decreasing over time.*
Different optimizers. *Ada Gradient.*

Details ConvNet

LAYER TYPE	SIZE
Convolution	32 3×3 filters
Convolution	16 3×3 filters
Max pooling	2×2 with stride of 2
Convolution	64 3×3 filters
Convolution	32 3×3 filters
Max pooling	2×2 with stride of 2
Convolution	128 3×3 filters
Convolution	128 3×3 filters
Convolution	64 3×3 filters
Max pooling	2×2 with stride of 2
Convolution	256 3×3 filters
Convolution	256 3×3 filters
Convolution	128 3×3 filters
Max pooling	2×2 with stride of 2
Flattening	8*8*128
Dense + dropout	512
Dense + dropout	256
Logit	121

Results

82.7%
CORRECT PREDICTIONS

Kaggle rank #2

Fire Breathing Rubber Duckies. *Burning rubber has never been this cute.*

Mirja Lagerwaard (10363149) & Wouter Vrielink (10433597)

kaggle **1st National Data Science Bowl.** *Predict ocean health, one plankton at a time.*