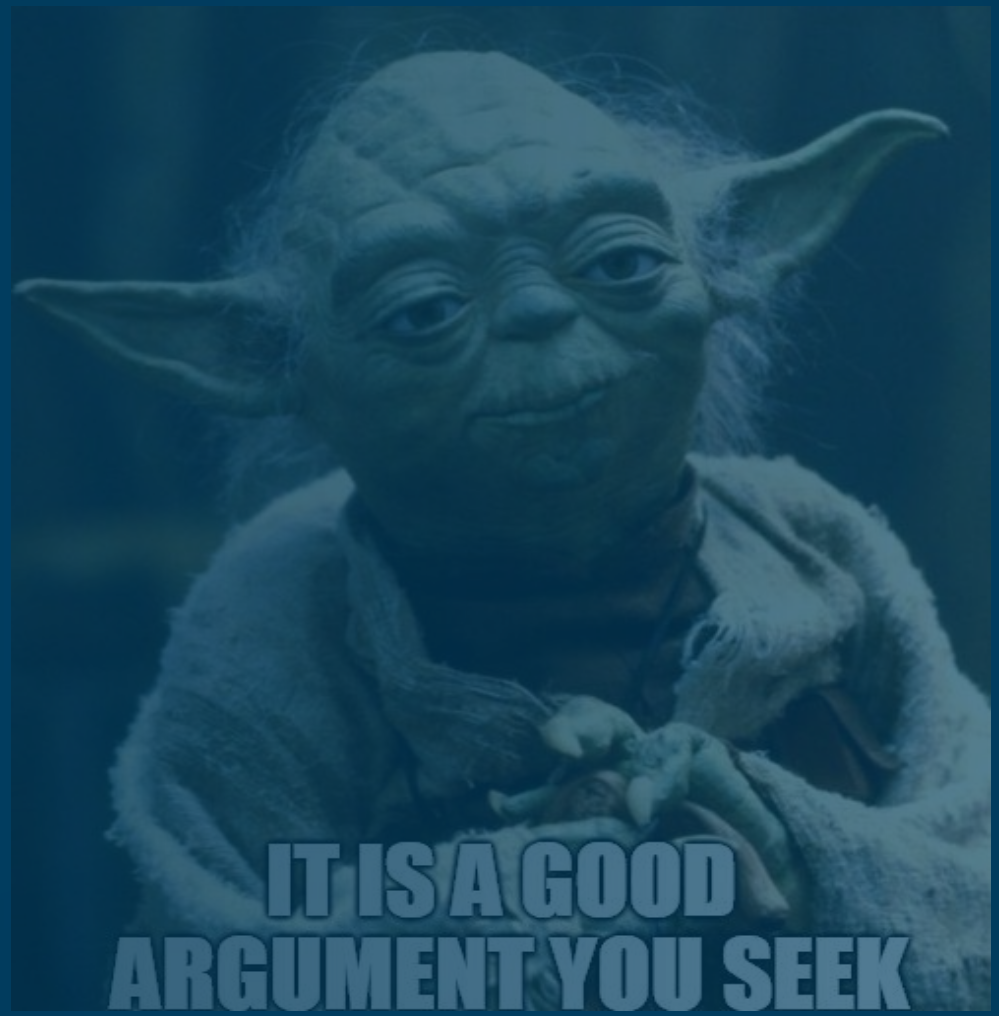
Athens University of Economics and Business

M.Sc. Business Analytics

Machine Learning & Content Analytics

Title: Argument Detection



Agenda

- The Existing Problem
- Methodology
- Results

The Problem

Content Quality Evaluation of existing bibliography (scientific documents, essays, e-books , articles etc) regarding the achievement or not of the sustainable development goals.

The majority of papers do not include the necessary information in order to elaborate their findings. Need for a mechanism to detect valid papers that include complete arguments, meaning evidences that support existing claims.

Methodology

- Data Acquisition
- Data Preprocessing
- Data Cleansing
- Data preparation

- Pre-Trained Glove Embeddings
- Model Structure
- Hyper - Parameters
- Performance Measurements
- Class Weights
- Optimizers and Learning Rates

✓ F1 – Score (focused on claim and evidence classes)

✓ Confusion Matrix

✗ Accuracy

Predictions of the best model on blind data set.

Showcasing the best model and its performance with plots, curves and matrices.

Hyper-Parameters Testing

CNN

- Number of Filters
- Extra Layers
- Kernel Size
- Batch Size
- Number of Epochs
- Dropout Rate (before/after pooling layer)
- Class Weights
- Optimizers
- Learning Rate

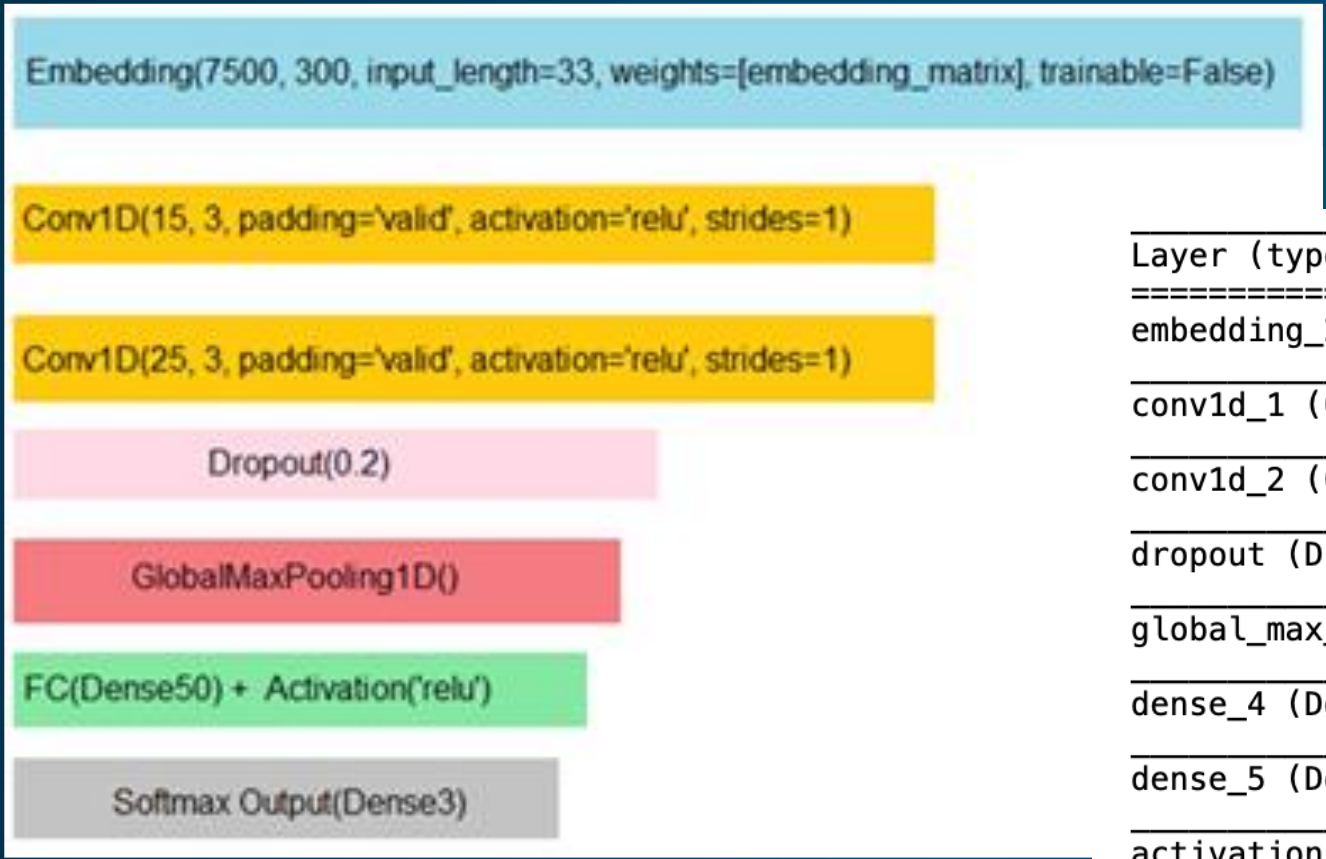
RNN

- Batch Size
- Number of Epochs
- Dropout Rate
- Class Weights
- Optimizers
- Learning Rate

Best CNN Model

Characteristics		Values
Hyper-parameters	Number of filters (1st layer)	15
	Number of filters (2nd layer)	25
	Kernel size	3
	Hidden size	50
	Epochs	100
	Batch size	32
	Dropout rate	0.2
Extras	Optimizers	Adam
	Learning Rate	1E-3
	Class Weights	-

CNN Model Architecture



Layer (type)	Output Shape	Param #
embedding_2 (Embedding)	(None, 33, 300)	2250000
conv1d_1 (Conv1D)	(None, 31, 15)	13515
conv1d_2 (Conv1D)	(None, 29, 25)	1150
dropout (Dropout)	(None, 29, 25)	0
global_max_pooling1d_1 (Glob	(None, 25)	0
dense_4 (Dense)	(None, 50)	1300
dense_5 (Dense)	(None, 3)	153
activation_2 (Activation)	(None, 3)	0

Total params: 2,266,118
Trainable params: 16,118
Non-trainable params: 2,250,000

Best CNN Model

Characteristics		Values
Hyper-parameters	Bidirectional LSTM(1st layer)	33
	Bidirectional LSTM(2nd layer)	33
	Epochs	60
	Batch size	256
	Dropout rate	0.1
Extras	Optimizers	SGD
	Learning Rate	1E-3
	Class Weights	-

RNN Model Architecture

Embedding(7500,300, input_length = 33 , weights = [embedding_matrix], trainable = False)

LSTM(33, return_sequences=False)

LSTM(33, return_sequences=False)

Dropout in [0.1, 0.5]

Softmax Output (Dense3)

```
Model: "functional_11"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
input_6 (InputLayer)         [(None, 33)]              0
_________________________________________________________________
embedding_4 (Embedding)      (None, 33, 300)           2250000
_________________________________________________________________
bidirectional_5 (Bidirection (None, 66)                88176
_________________________________________________________________
dropout_2 (Dropout)          (None, 66)                0
_________________________________________________________________
dense_7 (Dense)              (None, 3)                 201
=================================================================
Total params: 2,338,377
Trainable params: 88,377
Non-trainable params: 2,250,000
```

Results – Model Comparison

From Classification Report

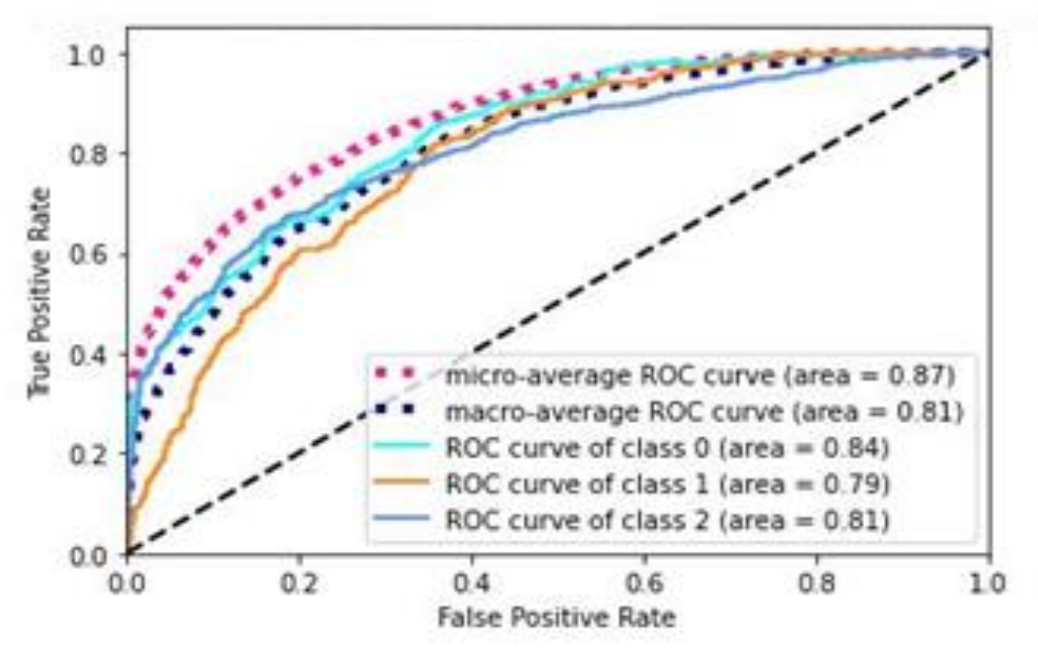
Best Chosen Model	f1-score (Class 0)	f1-score (Class 1)	Accuracy
RNN	.403	.370	.736
CNN	.399	.407	.694

From Confusion Matrix

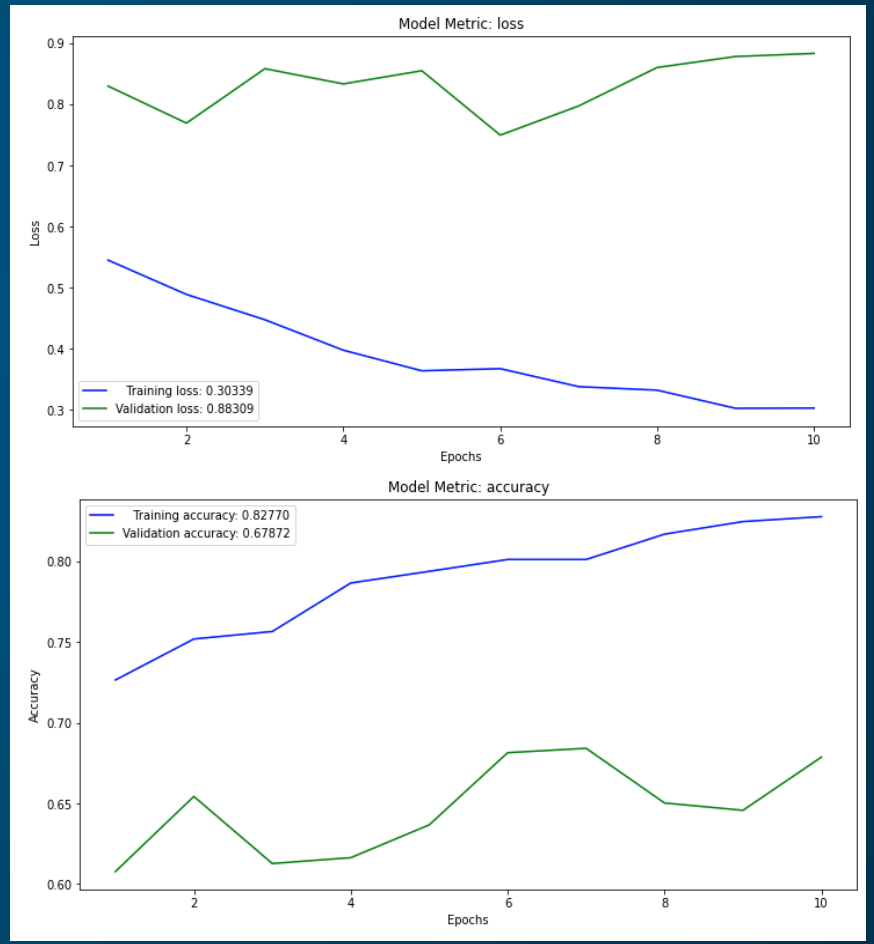
Best Chosen Model	Claims	Evidences
RNN	82/198 = 41%	117/297 = 39%
CNN	108/198 = 54%	153/297 = 51%

Results – Best Model

Roc Curves



Loss and Accuracy plots



Questions and Thoughts?

Thank you!