

Natural Language Processing for Finance with Transformer Models

Lawrence Johny

ljohny@mathworks.com

Valerio Sperandeo

vsperand@mathworks.com

MATLAB Apps

Friendly Interface to Perform Iterative Compute-Heavy Workflows...

```

10
11 % Fit Model
12
13 numNeighbors = 5 : 5 : 50 ;
14 distanceWeight = [ "equal" ; "inverse" ; "squaredinverse" ] ;
15
16 vnames = [ "numNeighbors" , "distanceWeight" , "loss" ] ;
17 resultSet = array2table( [NaN, "", NaN], "VariableNames", vnames ) ;
18 idxTable = 1 ;
19
20 for x = 1 : numel( distanceWeight )
21     for y = 1 : numel( numNeighbors )
22
23         Mdl = fitcknn( predictors, response, ...
24             'NumNeighbors', numNeighbors(y), ...
25             'DistanceWeight', distanceWeight(x) );
26
27         loss = resubLoss(Mdl) ;
28         resultSet(idxTable, :) = { numNeighbors(y), string(distanceWeight(x)), loss } ;
29         idxTable = idxTable + 1 ;
30
31     end
32 end
33

```

Command Window

New to MATLAB? See resources for [Getting Started](#).

numNeighbors	distanceWeight	loss
"5"	"equal"	"0.024845"
"10"	"equal"	"0.024845"
"15"	"equal"	"0.018634"
"20"	"equal"	"0.031056"
"25"	"equal"	"0.037267"
"30"	"equal"	"0.031056"
"35"	"equal"	"0.024845"
"40"	"equal"	"0.037267"
"45"	"equal"	"0.086957"
"50"	"equal"	"0.13665"
"5"	"inverse"	"0"
"10"	"inverse"	"0"
"15"	"inverse"	"0"
"20"	"inverse"	"0"
"25"	"inverse"	"0"



Experiment Manager

EXPERIMENT MANAGER

FILE ENVIRONMENT PARALLEL RUN

Experiment Browser

demo-experiment-manager

Experiment1

Result1

Description

Test multiple parameters of Classifiers

Hyperparameters

Name	Values
numNeighbors	5:5:50
DistanceWeight	["equal","inverse","squaredInverse"]

Training Function

Experiment1_training1

... and Offload Repetitive Tasks !

Experiment Manager

EXPERIMENT MANAGER

New Open Save Duplicate FILE

Layout ENVIRONMENT

Use Parallel PARALLEL

Restart All Canceled RUN

Stop

Training Plot REVIEW RESULTS

Confusion Matrix

Filter FILTER

Annotations ANNOTATIONS

Export EXPORT

Experiment Browser

demo-experiment-manager

Experiment1

Result1

Experiment1 | Result1

Exhaustive Sweep Result

Experiment1 Start: 23/08/2021 à 11:12:36 30/30 Trials

(View Experiment Source)

Test multiple parameters of Classifiers

Complete 30 Stopped 0 Error 0

Running 0 Queued 0 Canceled 0

Trial	Status	Progress	Elapsed Time	numNeighbors	DistanceWeight	Loss
1	Complete	100.0%	0 hr 0 min 2 sec	5.0000	equal	0.0248
2	Complete	100.0%	0 hr 0 min 2 sec	10.0000	equal	0.0248
3	Complete	100.0%	0 hr 0 min 1 sec	15.0000	equal	0.0186
4	Complete	100.0%	0 hr 0 min 2 sec	20.0000	equal	0.0311
5	Complete	100.0%	0 hr 0 min 1 sec	25.0000	equal	0.0373
6	Complete	100.0%	0 hr 0 min 1 sec	30.0000	equal	0.0311
7	Complete	100.0%	0 hr 0 min 2 sec	35.0000	equal	0.0248
8	Complete	100.0%	0 hr 0 min 1 sec	40.0000	equal	0.0373
9	Complete	100.0%	0 hr 0 min 2 sec	45.0000	equal	0.0870
10	Complete	100.0%	0 hr 0 min 1 sec	50.0000	equal	0.1366
11	Complete	100.0%	0 hr 0 min 2 sec	5.0000	inverse	0.0000
12	Complete	100.0%	0 hr 0 min 1 sec	10.0000	inverse	0.0000
13	Complete	100.0%	0 hr 0 min 1 sec	15.0000	inverse	0.0000
14	Complete	100.0%	0 hr 0 min 2 sec	20.0000	inverse	0.0000
15	Complete	100.0%	0 hr 0 min 1 sec	25.0000	inverse	0.0000
16	Complete	100.0%	0 hr 0 min 2 sec	30.0000	inverse	0.0000
17	Complete	100.0%	0 hr 0 min 1 sec	35.0000	inverse	0.0000
18	Complete	100.0%	0 hr 0 min 2 sec	40.0000	inverse	0.0000
19	Complete	100.0%	0 hr 0 min 1 sec	45.0000	inverse	0.0000
20	Complete	100.0%	0 hr 0 min 2 sec	50.0000	inverse	0.0000
21	Complete	100.0%	0 hr 0 min 1 sec	5.0000	squaredInverse	0.0000
22	Complete	100.0%	0 hr 0 min 2 sec	10.0000	squaredInverse	0.0000
23	Complete	100.0%	0 hr 0 min 1 sec	15.0000	squaredInverse	0.0000
24	Complete	100.0%	0 hr 0 min 1 sec	20.0000	squaredInverse	0.0000

We rely on a dictionary to identify sentiment of a sentence

Sentences	GenericNLP
"The book was VERY good!!!!"	0.73
"The book was not very good."	-0.39

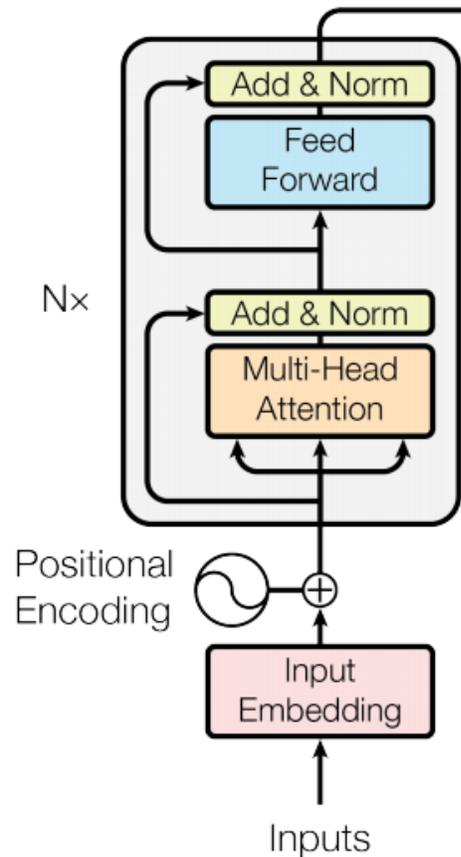
Generic dictionary does not understand financial lexicon

Sentences	GenericNLP
"The book was VERY good!!!!"	0.73
"The book was not very good."	-0.39
"The EBIT was VERY good!!!!"	0.73
"The EBIT was not very good."	-0.39

A Transformer Model like FinBERT knows Financial Jargon

Sentences	GenericNLP	FinBERT
"The book was VERY good!!!!"	0.73	0.74
"The book was not very good."	-0.39	-0.47
"The EBIT was VERY good!!!!"	0.73	0.83
"The EBIT was not very good."	-0.39	-0.88

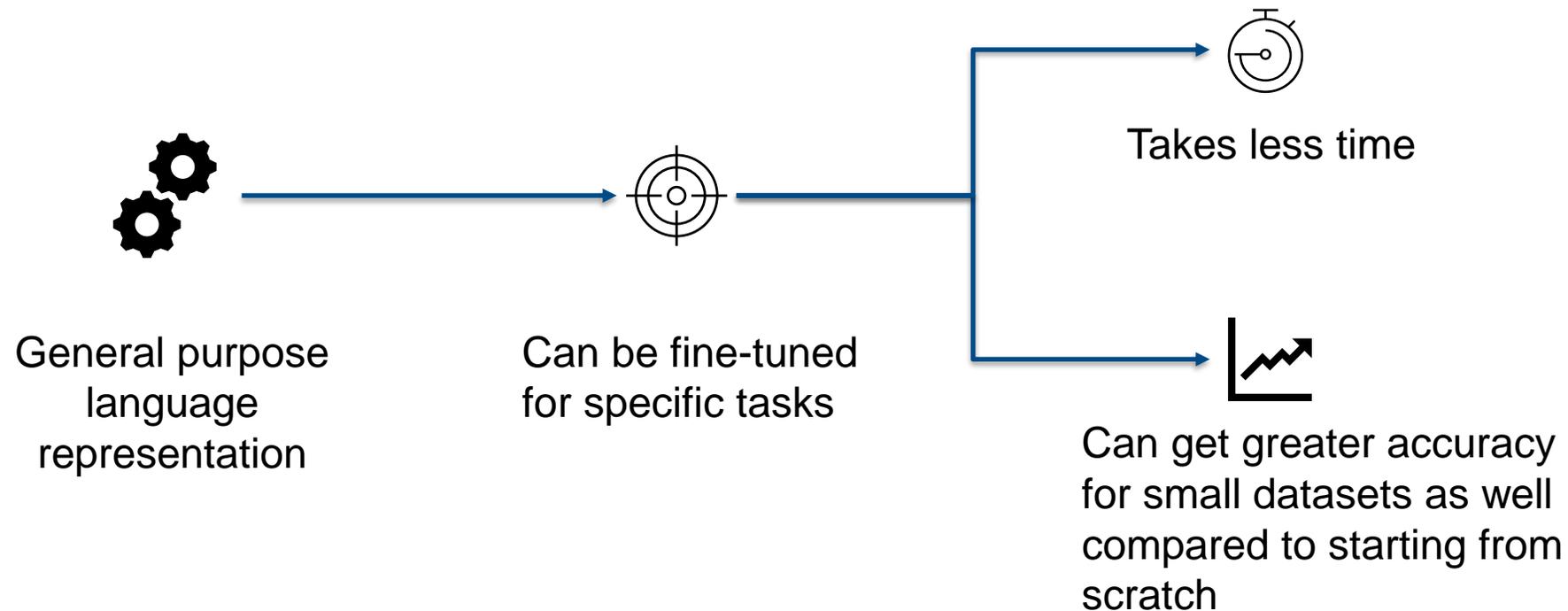
Make context-aware NLP Models by adding attention mechanism



- Bidirectional Encoder Representations from Transformers
- Custom Word-Piece tokenizer
- Transformer-encoder architecture – based on self-attention.

Going from BERT to FinBERT

FinBERT = Trained Model to Analyze Sentiment of Financial Text



What are Transformer Models used for in Finance?

Sentiment Analysis

Identify and score sentiments expressed in text

Word	Label
"convient"	Positive
"ulterior"	Negative
"undecided"	Negative
"grating"	Negative
"partisans"	Negative
"wellbeing"	Positive
"excite"	Positive
"smug"	Negative
"bleeding"	Negative
"interfere"	Negative

Text Classification

Classify documents into pre-determined categories for efficient information retrieval and prediction

FundX promotes manager to head new research team

Low Risk

FundX hit by more PM departures, AAA-rated managers out

High Risk

Topic Modeling

Identify topics from a collection of documents that show underlying patterns and relationships in raw text data



Summarization

Extract/create a summary from one or more documents automatically

The San Bernardino County Sheriff's Department said it was a matter of public safety when deputies opened fire from a helicopter on a wrong-way chase suspect on the 215 Freeway.

On Friday, a suspect led authorities on a dangerous high-speed chase through surface streets in Fontana and San Bernardino before he drove the wrong way in the northbound lanes of the freeway. He reached speeds as high as 100 mph and deputies deemed him a threat to public safety.

Lt. Mitch Datillo said there is a lot for deputies to take into account before firing from a helicopter.

"If you're near schools, if you're in residential areas. We have had times where we fully intended that it was getting to that level and we could not deploy, based on the backdrop of that level safety there," he said.

But when the suspect jumped out of his SUV at the car continued to move and slammed head-on into the SUV, injuring three people. A relative of that far from the freeway, he questions the decision to open fire from the air.

While shooting from a helicopter is rare, it does a deputy with the Orange County Sheriff's Department opened fire from their helicopter on a suspect firing at deputies at a recycling plant near Irvine.

Eyewitness News spoke with a use-of-force expert with the County Sheriff's Deputy Ed Yamashita, who said the decision to shoot, based on the fact of speed while heading in the wrong direction on the freeway.

"It could have been worse," he said. "It could have been an immediate grave danger to the public if the suspect had been able to get to the freeway."

The Los Angeles County Sheriff's Special Enforcement Bureau is authorized and trains deputies to shoot from a helicopter.

Ventura County sheriff's deputies are authorized to shoot from a helicopter after hearing about the deadly chase, officials are considering changing the policy and training deputies to shoot from a helicopter.

The San Bernardino County Sheriff's Department said it was a matter of public safety when deputies opened fire from a helicopter on a wrong-way chase suspect on the 215 Freeway.

Lt. Mitch Datillo said there is a lot for deputies to take into account before firing from a helicopter.

In 2004, a deputy with the Orange County Sheriff's Department opened fire from their helicopter on a suspect who was firing at deputies at a recycling plant near Irvine.

The Los Angeles County Sheriff's Special Enforcement Bureau is authorized and trains deputies to shoot from a helicopter.

How are Transformer Models used in practice?

Building Custom Applications – Using FinBERT for Text Classification

Text Classification

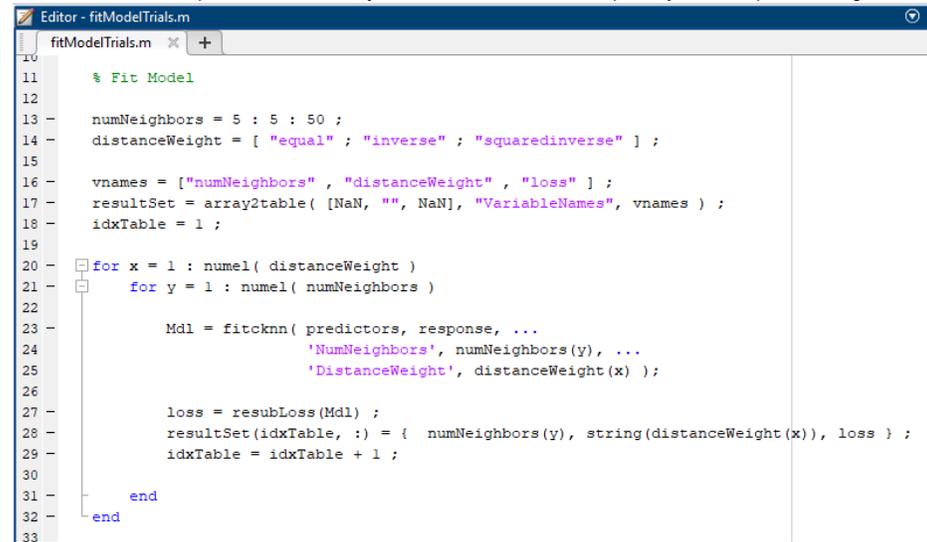
Classify documents into pre-determined categories for efficient information retrieval and prediction



- Use Transformer Model as embedding and Train New Application-Specific Classifiers with MATLAB Apps
 - Visually Work with Various Classifier Families using [Classification Learner App](#)
 - Create & Analyze Networks with [Deep Network Designer](#)
- Fine Tune FinBERT with Hyper-Parameters and Social & Economic Text
 - Seek higher accuracy by testing with [Experiment Manager](#)

Challenges

- Code multiple families of classifiers.
- Manage potentially thousands of parameters in Deep Learning networks.
- Save time by parallelization.



```

10
11 % Fit Model
12
13 numNeighbors = 5 : 5 : 50 ;
14 distanceWeight = [ "equal" ; "inverse" ; "squaredinverse" ] ;
15
16 vnames = [ "numNeighbors" , "distanceWeight" , "loss" ] ;
17 resultSet = array2table( [NaN, "", NaN], "VariableNames", vnames ) ;
18 idxTable = 1 ;
19
20 for x = 1 : numel( distanceWeight )
21     for y = 1 : numel( numNeighbors )
22
23         Mdl = fitcknn( predictors, response, ...
24             'NumNeighbors', numNeighbors(y), ...
25             'DistanceWeight', distanceWeight(x) );
26
27         loss = resubLoss( Mdl ) ;
28         resultSet( idxTable, :) = { numNeighbors(y), string( distanceWeight(x) ), loss } ;
29         idxTable = idxTable + 1 ;
30     end
31 end
32
33

```

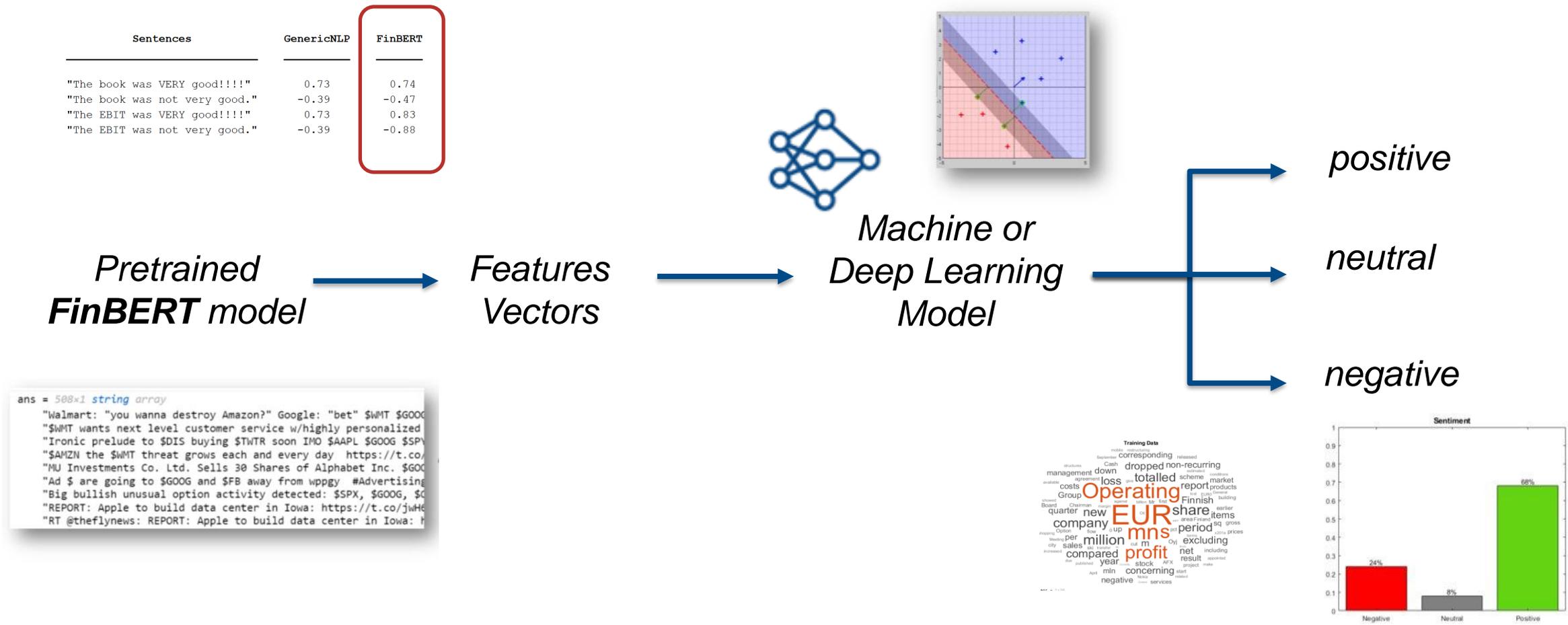
Command Window

New to MATLAB? See resources for [Getting Started](#).

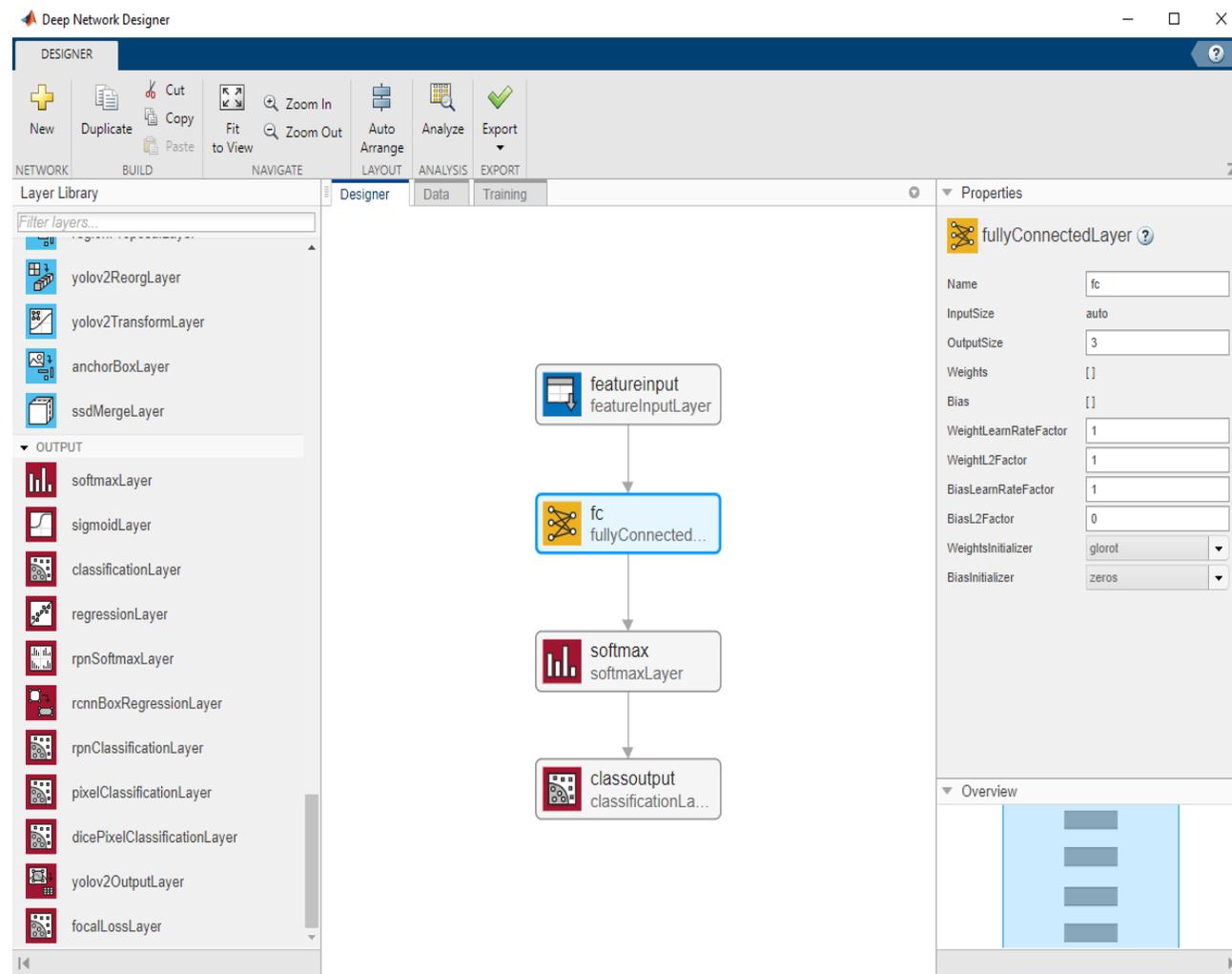
numNeighbors	distanceWeight	loss
"5"	"equal"	"0.024845"
"10"	"equal"	"0.024845"
"15"	"equal"	"0.018634"
"20"	"equal"	"0.031056"
"25"	"equal"	"0.037267"
"30"	"equal"	"0.031056"
"35"	"equal"	"0.024845"
"40"	"equal"	"0.037267"
"45"	"equal"	"0.086957"
"50"	"equal"	"0.13665"
"5"	"inverse"	"0"
"10"	"inverse"	"0"

Classify Sentiment of Financial Statements using Transformer Models

FinBERT can be used as **feature extractor**. These features are a numeric representation of contextual relations between words which can be used to train a machine or deep learning classification model.



Deep Learning **Sentiment Models** with MATLAB's Ease-of-Use



[Deep Network Designer](#)
[Network Analyzer](#)

**Visualize pretrained model or
 design model from scratch**

Deep Learning Sentiment Models with MATLAB's Ease-of-Use

Analyze Model

Deep Learning Network Analyzer

Igraph3
Analysis date: 30-Apr-2018 14:43:00

143 layers 0 warnings 4 errors

ISSUES

FOUND IN	MESSAGE
Network	Missing input layer. The network must have one input layer.
conv1-7x7_s2	Missing input. Each layer input must be connected to the output of another layer.
inception_4b-3x3	Unused output. Each layer output must be connected to the input of another layer.
inception_4b-relu_3x3	Missing input. Each layer input must be connected to the output of another layer.

ANALYSIS RESULT

NAME	TYPE	ACTIVATIONS	LEARNABLES
inception_4a-relu_5x5_reduce	ReLU	Unknown	-
inception_4a-5x5	Convolution	Unknown	Weights 5x5x16x48 Bias 1x1x48
inception_4a-relu_5x5	ReLU	Unknown	-
inception_4a-pool	Max Pooling	Unknown	-
inception_4a-pool_proj	Convolution	Unknown	Weights 1x1x480x64 Bias 1x1x64
inception_4a-relu_pool_proj	ReLU	Unknown	-
inception_4a-output	Depth concatenation	Unknown	-
inception_4b-1x1	Convolution	Unknown	Weights 1x1x512x160 Bias 1x1x160
inception_4b-relu_1x1	ReLU	Unknown	-
inception_4b-3x3_reduce	Convolution	Unknown	Weights 1x1x512x112 Bias 1x1x112
inception_4b-relu_3x3_reduce	ReLU	Unknown	-
inception_4b-3x3	Convolution	Unknown	Weights 3x3x112x224 Bias 1x1x224

[Deep Network Designer](#)
[Network Analyzer](#)

Monitor Training Progress and automatically generate MATLAB code

Deep Network Designer

TRAINING

Training Options TRAIN EXPORT

Designer Data Training

Training Progress (18-Aug-2021 14:27:53)

Results

Validation accuracy: 92.50%

Training finished: Reached final iteration

Training Time

Start time: 18-Aug-2021 14:27:53

Elapsed time: 37 sec

Training Cycle

Epoch: 30 of 30

Iteration: 30 of 30

Iterations per epoch: 1

Maximum iterations: 30

Validation

Frequency: 50 iterations

Other Information

Hardware resource: Single CPU

Learning rate schedule: Constant

Learning rate: 0.001

[Learn more](#)

Accuracy

- Training (smoothed)
- Training
- Validation

Loss

MATLAB is built to...

Empower People to Convert Ideas to Action

Share Knowledge Effortlessly

- Apps
 - Domain expertise → Point-&-Click App and automatically generate code from Apps
 - Classification Learner, Deep Network Designer,
- Integrates Workflows
 - Automate research to production with Unit Testing & System Testing
 - Works well with Excel & other programming languages
- Ease of scale
 - Easy parallelization and scaling up

Thank You!