HIGHER SCHOOL OF ECONOMICS

NATIONAL RESEARCH UNIVERSITY

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ

AUTOMATIC PRIVACY DETECTION IN SCANNED DOCUMENT IMAGES BASED ON DEEP NEURAL NETWORKS

Lyudmila Kopeykina, Andrey V. Savchenko
e-mail : lnkopeykina@mail.ru

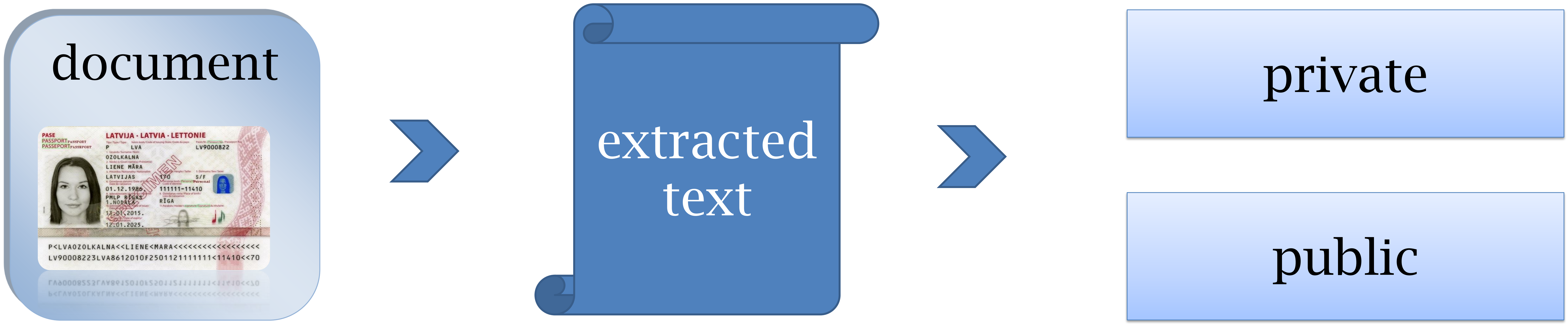
RusAutoCon2019

OUTLINE

- The problem of automatic detection of private scanned documents
- Proposed approach for classification of private and public scanned documents
- Experimental results in automatic detection of private scanned documents
- Concluding comments and future plans

MOTIVATION AND PROBLEM FORMULATION

It is required to assign an image of scanned English document to one of two possible classes (private or public) according to the extracted text from the image



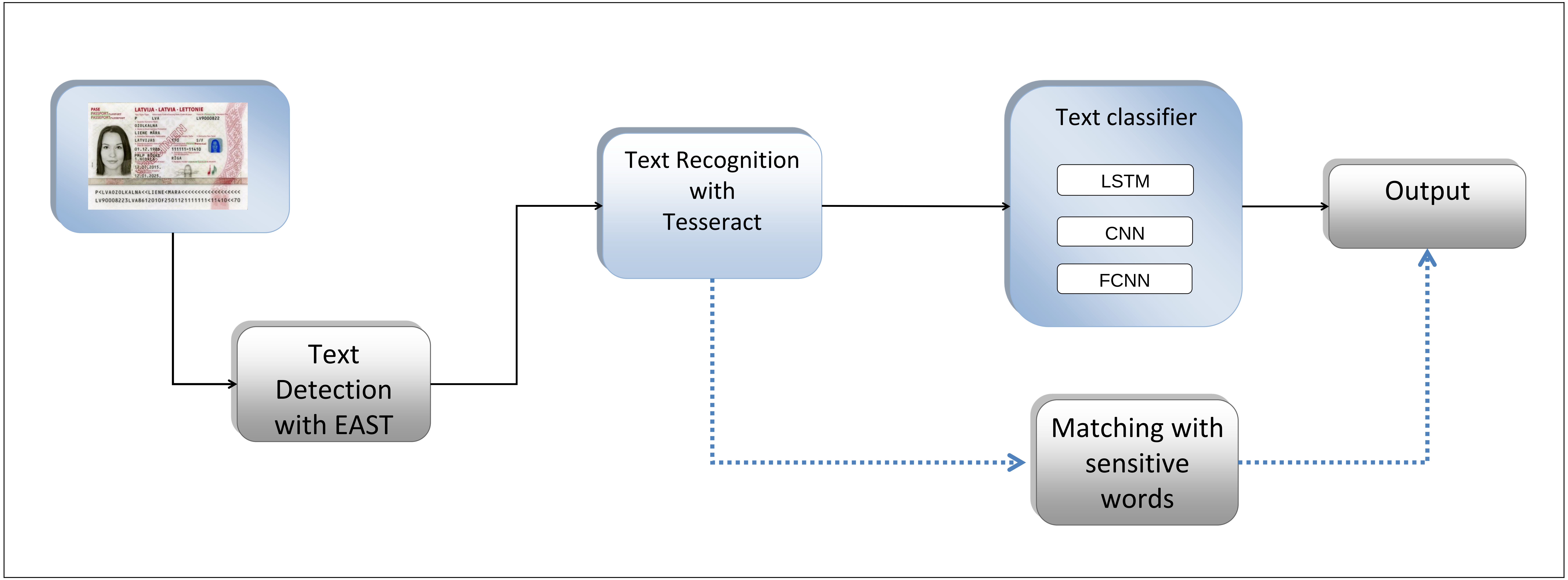
document

extracted text

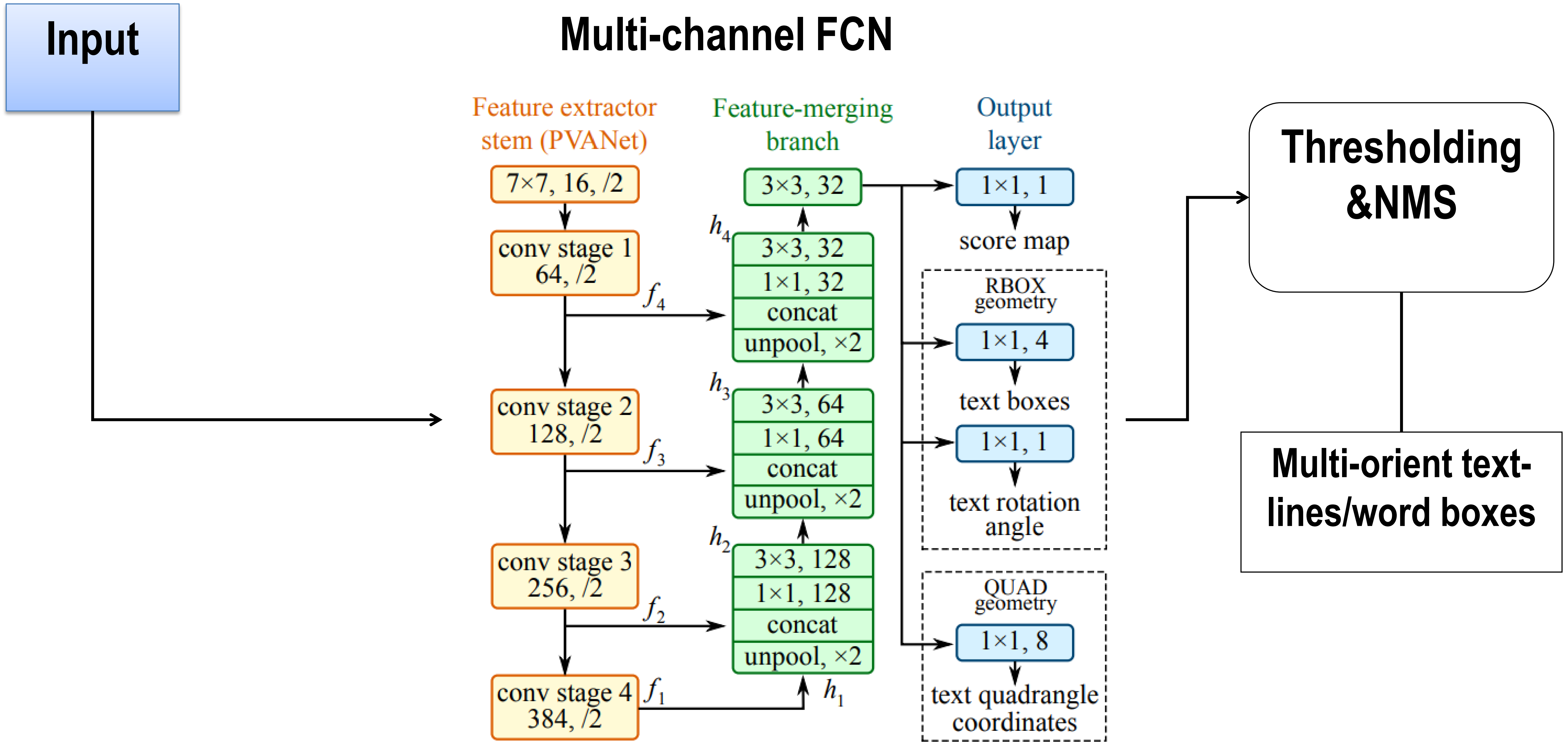
private

public

PROPOSED PIPELINE



EAST TEXT DETECTOR FROM ORIGINAL PAPER



Input

Multi-channel FCN

Thresholding &NMS

Multi-orient text-lines/word boxes

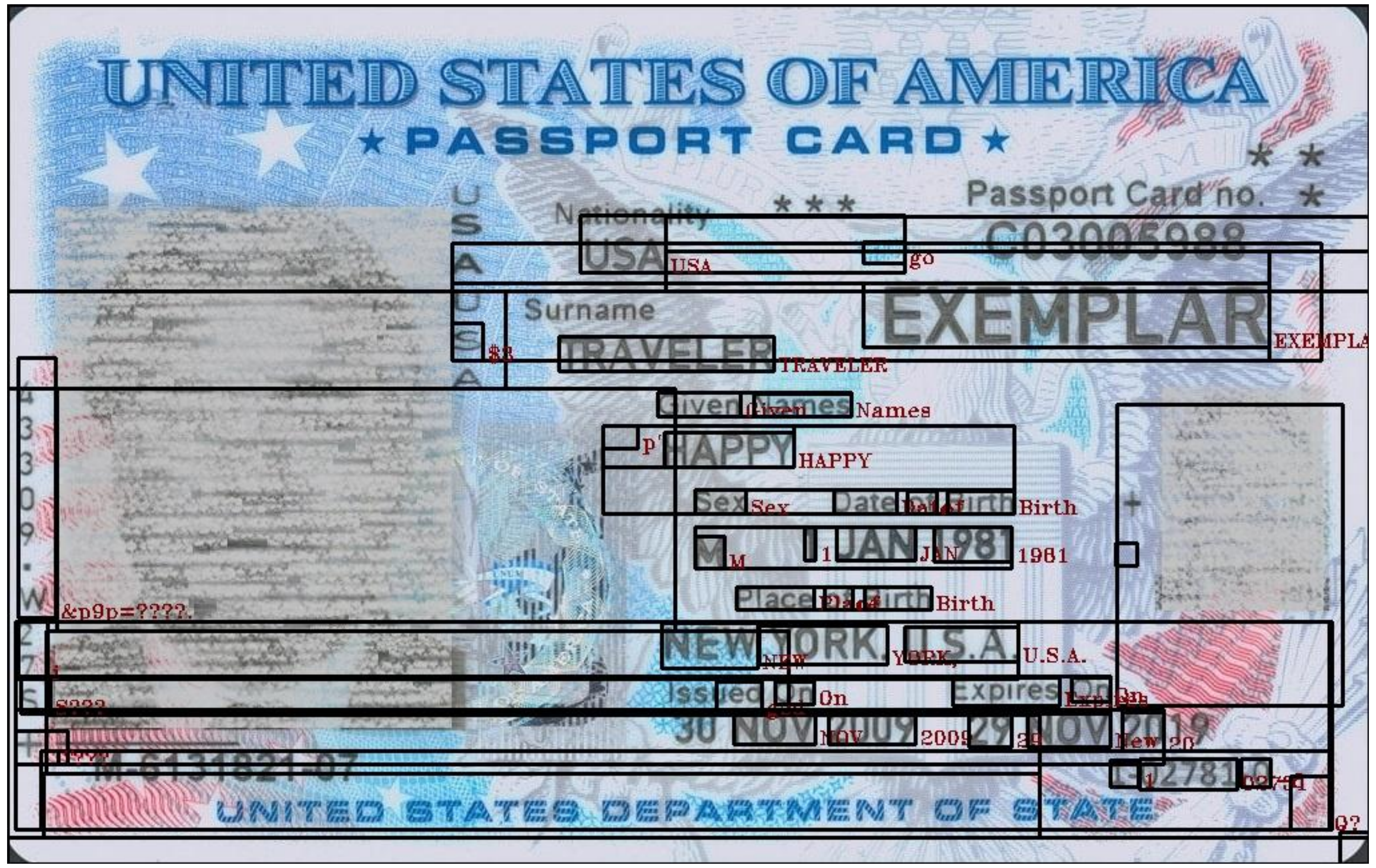
X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He, and J. Liang, "EAST: an efficient and accurate scene text detector," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5551-5560.

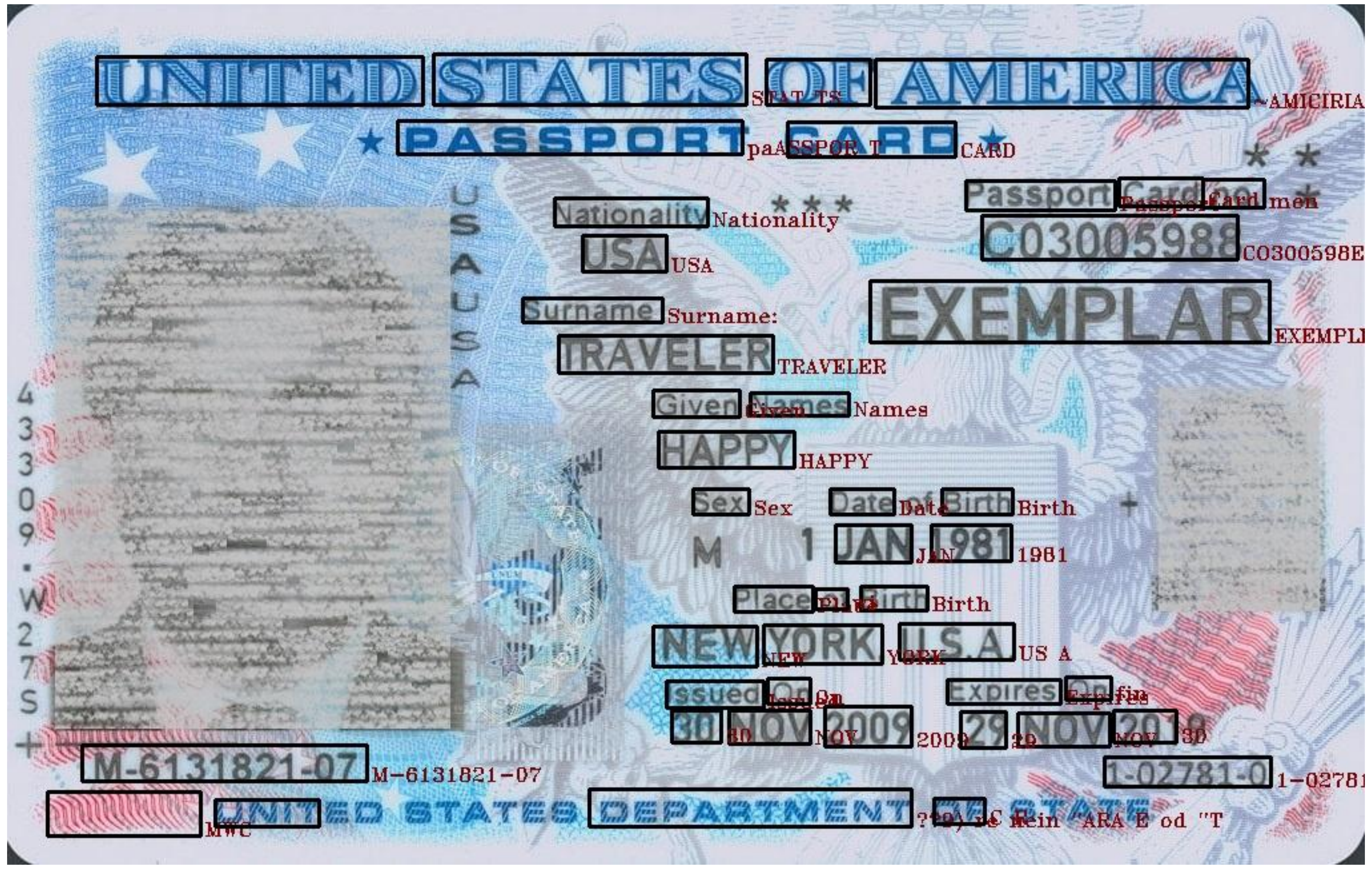
A tensorflow implementation of EAST text detector

TEXT DETECTION AND RECOGNITION

- TensorFlow re-implementation of the EAST to detect regions with text.
- As regards text recognition, we used Tesseract 4.0 in image_to_string mode. Additionally, we set LSTM Engine mode to recognize characters on our images. To recognize text in areas, which EAST detector assigned with textual data areas, we switched the page-segmentation mode to psm=8 (ROI as a single word).

TEXT DETECTION AND RECOGNITION

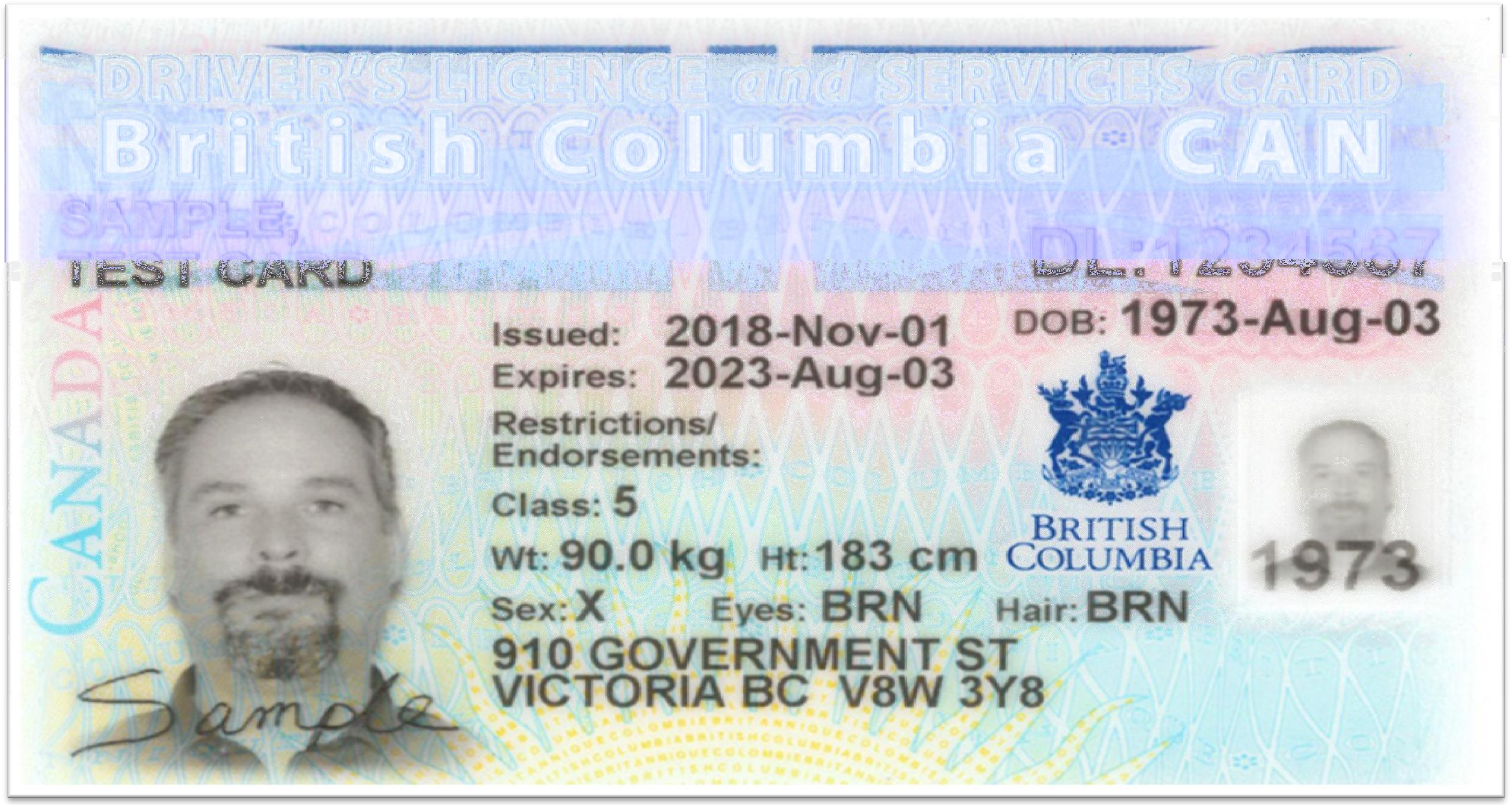




The fragments of text detection on image from MIDV-500: a) Tesseract b) EAST text-detector

DATASET

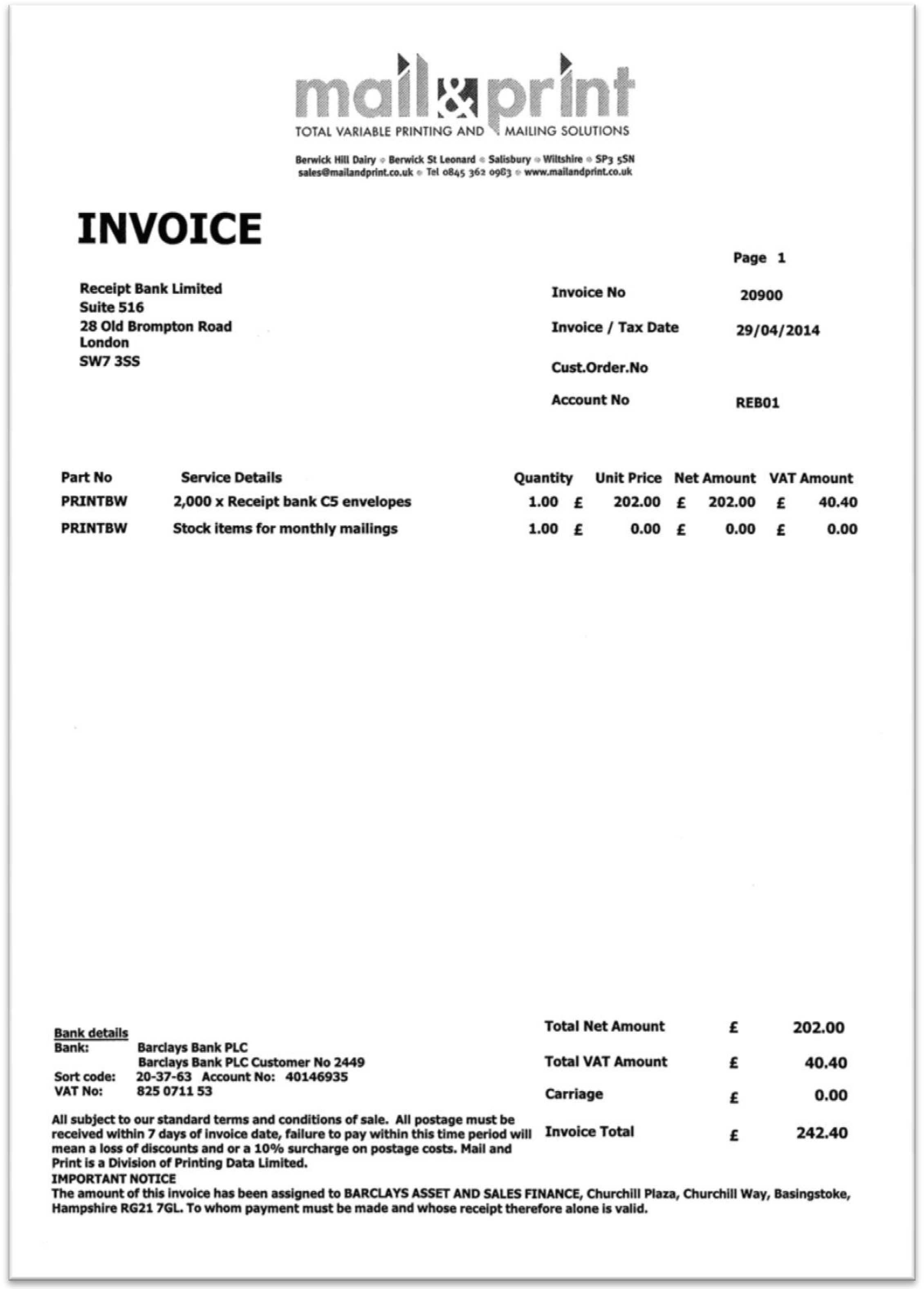
The positive class is presented by 350 images of driving license and medical insurance cards, passports and invoices from extension of the MIDV dataset, whereas negative class consists of photos from publicly available datasets for text classification tasks DIQA and Ghega.



DRIVER'S LICENCE and SERVICES CARD
British Columbia CAN
SAMPLE
TEST CARD
DL:1234567
CANADA

Issued: 2018-Nov-01
Expires: 2023-Aug-03
DOB: 1973-Aug-03
Restrictions/
Endorsements:
Class: 5
Wt: 90.0 kg Ht:183 cm
Sex: X Eyes: BRN Hair: BRN
910 GOVERNMENT ST
VICTORIA BC V8W 3Y8

BRITISH COLUMBIA
1973
Sample



mail&print
TOTAL VARIABLE PRINTING AND MAILING SOLUTIONS
Berwick Hill Dairy • Berwick St Leonard • Salisbury • Wiltshire • SP3 5SN
sales@mailandprint.co.uk • Tel 0845 362 0983 • www.mailandprint.co.uk

INVOICE

Page 1

Receipt Bank Limited
Suite 516
28 Old Brompton Road
London
SW7 3SS

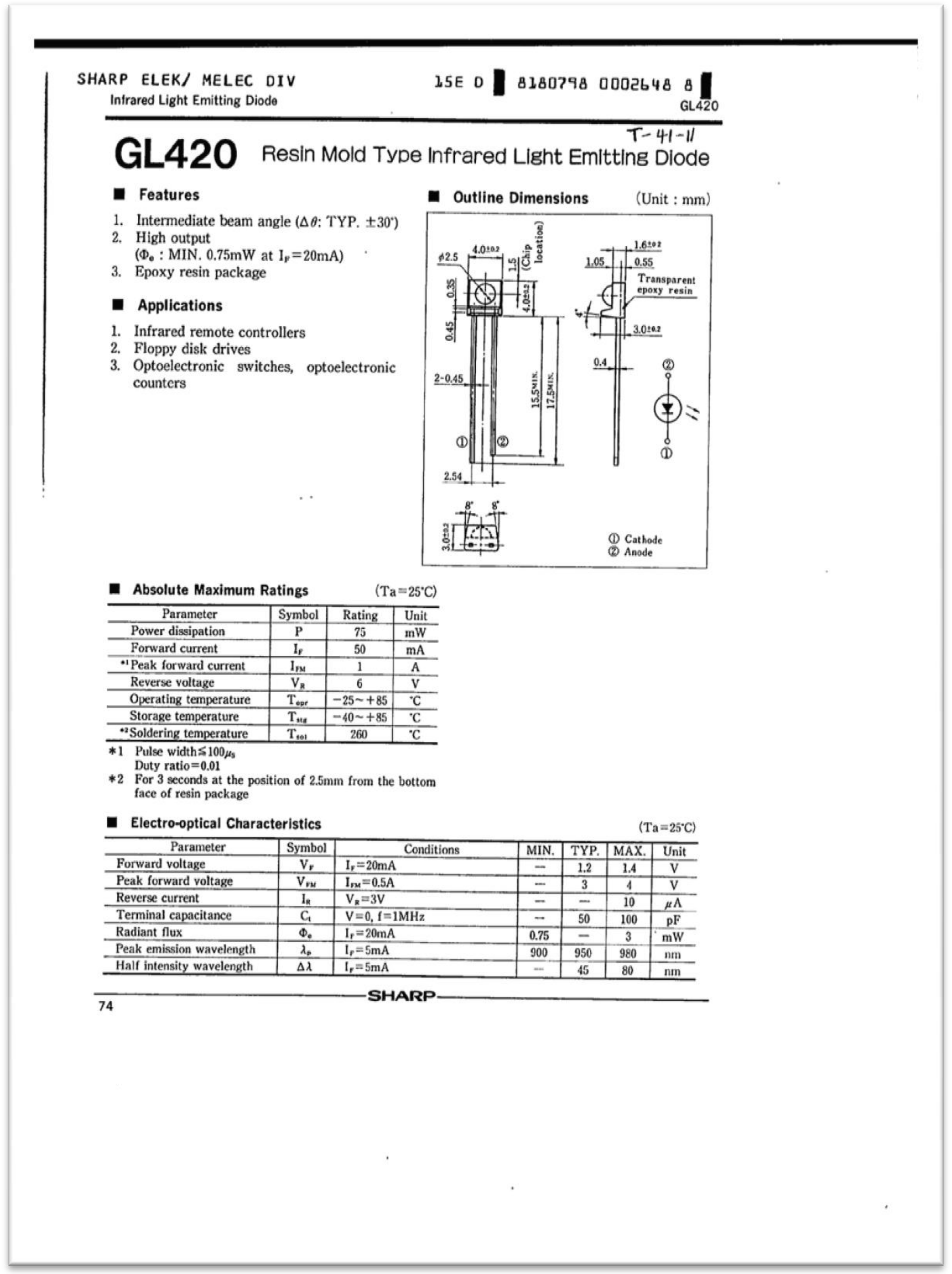
Invoice No	20900
Invoice / Tax Date	29/04/2014
Cust.Order.No	
Account No	REB01

Part No	Service Details	Quantity		Unit Price		Net Amount		VAT Amount
PRINTBW	2,000 x Receipt bank C5 envelopes	1.00	£	202.00	£	202.00	£	40.40
PRINTBW	Stock items for monthly mailings	1.00	£	0.00	£	0.00	£	0.00

Bank details
Bank: Barclays Bank PLC
Barclays Bank PLC Customer No 2449
Sort code: 20-37-63 Account No: 40146935
VAT No: 825 0711 53

Total Net Amount	£	202.00
Total VAT Amount	£	40.40
Carriage	£	0.00
Invoice Total	£	242.40

All subject to our standard terms and conditions of sale. All postage must be received within 7 days of invoice date, failure to pay within this time period will mean a loss of discounts and or a 10% surcharge on postage costs. Mail and Print is a Division of Printing Data Limited.
IMPORTANT NOTICE
The amount of this invoice has been assigned to BARCLAYS ASSET AND SALES FINANCE, Churchill Plaza, Churchill Way, Basingstoke, Hampshire RG21 7GL. To whom payment must be made and whose receipt therefore alone is valid.



SHARP ELEK/ MELEC DIV 15E D 8180798 0002648 8
Infrared Light Emitting Diode GL420
T-41-11

GL420 Resin Mold Type Infrared Light Emitting Diode

■ Features
1. Intermediate beam angle ($\Delta\theta$: TYP. ±30°)
2. High output
(Φ_e : MIN. 0.75mW at I_F=20mA)
3. Epoxy resin package

■ Applications
1. Infrared remote controllers
2. Floppy disk drives
3. Optoelectronic switches, optoelectronic counters

■ Outline Dimensions (Unit : mm)

■ Absolute Maximum Ratings (Ta=25°C)

Parameter	Symbol	Rating	Unit
Power dissipation	P	75	mW
Forward current	I_F	50	mA
*1 Peak forward current	I_{FM}	1	A
Reverse voltage	V_R	6	V
Operating temperature	T_{opr}	−25~+85	°C
Storage temperature	T_{stg}	−40~+85	°C
*2 Soldering temperature	T_{sol}	260	°C

*1 Pulse width≦100μs
Duty ratio=0.01
*2 For 3 seconds at the position of 2.5mm from the bottom face of resin package

■ Electro-optical Characteristics (Ta=25°C)

Parameter	Symbol	Conditions	MIN.	TYP.	MAX.	Unit
Forward voltage	V_F	I_F=20mA	—	1.2	1.4	V
Peak forward voltage	V_{FM}	I_{FM}=0.5A	—	3	4	V
Reverse current	I_R	V_R=3V	—	—	10	μA
Terminal capacitance	C_t	V=0, f=1MHz	—	50	100	pF
Radiant flux	Φ_e	I_F=20mA	0.75	—	3	mW
Peak emission wavelength	λ_p	I_F=5mA	900	950	980	nm
Half intensity wavelength	$\Delta\lambda$	I_F=5mA	—	45	80	nm

SHARP
74

CLASSIFICATION

Keyword spotting

•Matching of detected text with a list of sensitive attributes, such as "passport", "invoice", etc.

•The recognized word is labeled as sensitive (""1") if the similarity with one of the keywords was higher than 0.8.

•This similarity is computed as 1 minus edit distance between an input word and keyword relative to the length of a keyword.

•Each photo is associated with a vector with zeros and ones ("1" - the word is sensitive and "0" - the word is not sensitive). The input image was classified as private when there was at least one sensitive attribute.

Neural networks

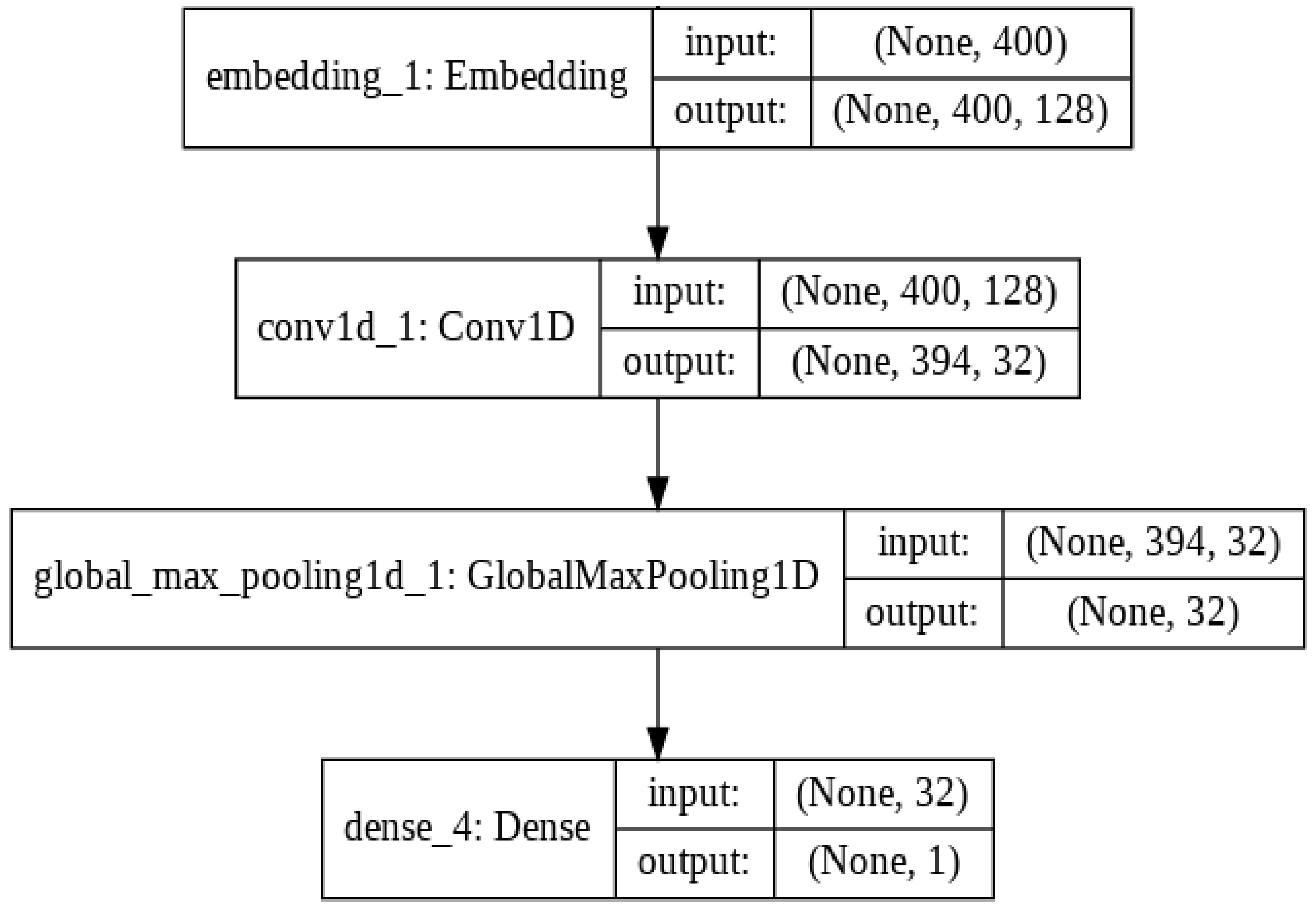
•To present input data as vectors, one-hot-encoding was applied.

•To be more exact, we created a vocabulary of D=5000 most frequent words recognized with detectors matched each word with the dictionary. Then for each image we got a list of indices of a specific word in a dictionary.

CLASSIFICATION

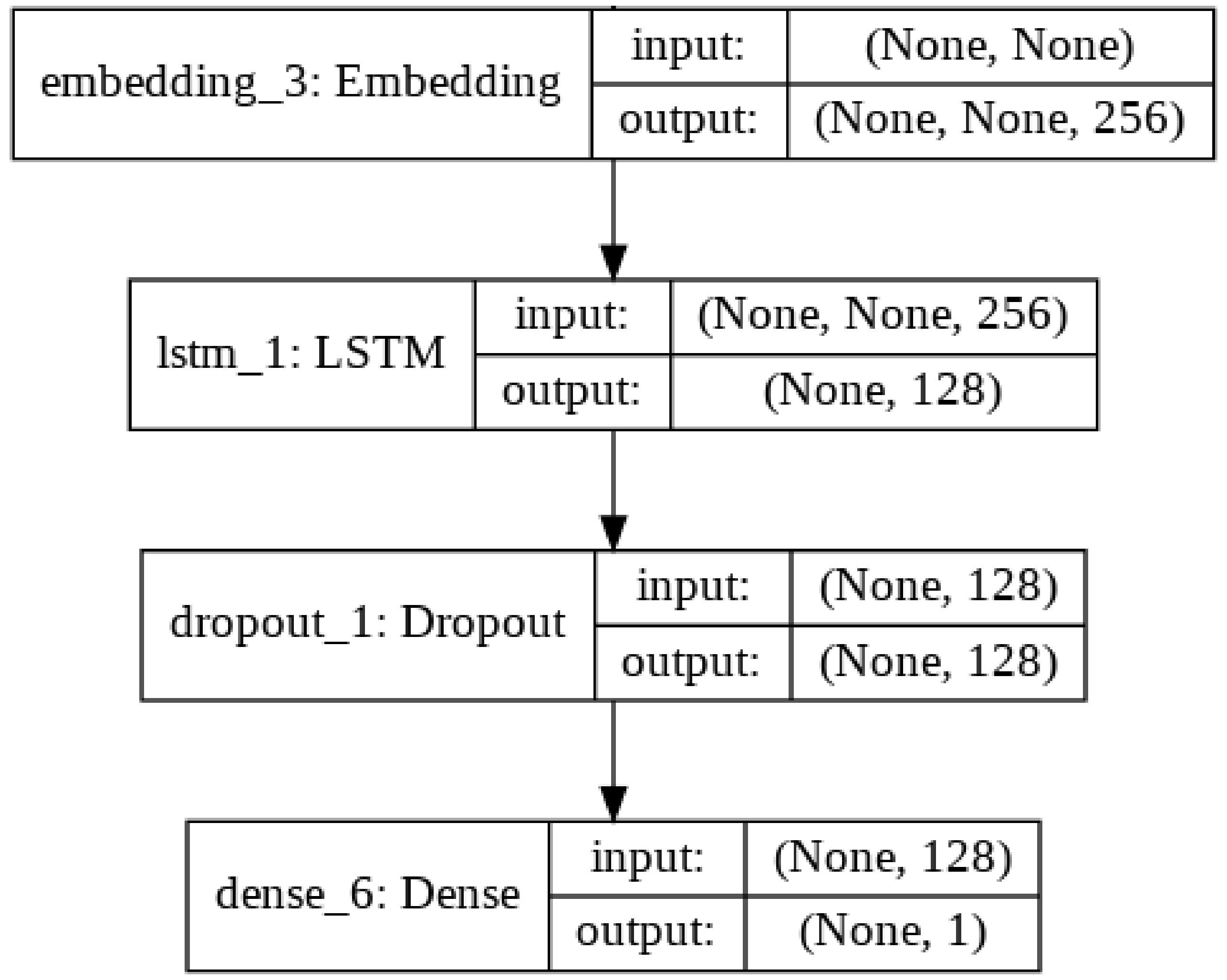
NEURAL NETWORKS

Keras + TensorFlow



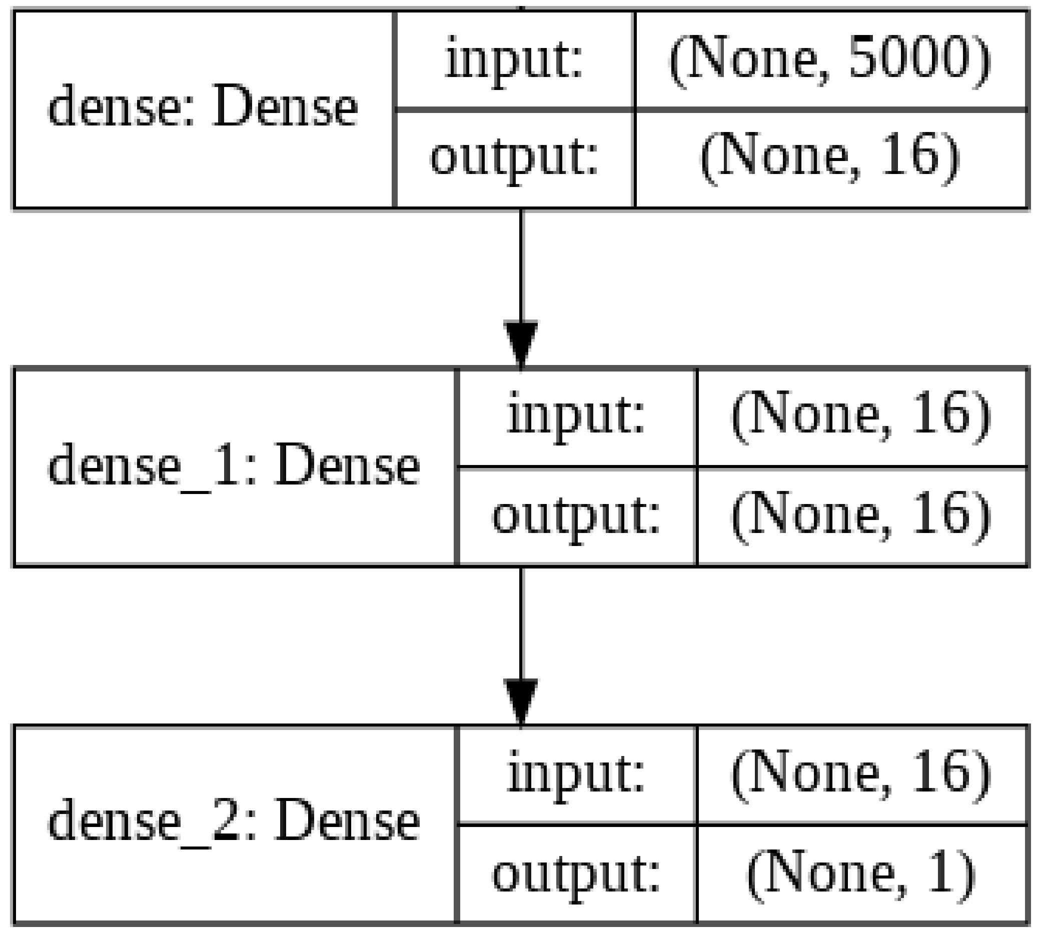
embedding_1: Embedding	input:	(None, 400)
	output:	(None, 400, 128)
conv1d_1: Conv1D	input:	(None, 400, 128)
	output:	(None, 394, 32)
global_max_pooling1d_1: GlobalMaxPooling1D	input:	(None, 394, 32)
	output:	(None, 32)
dense_4: Dense	input:	(None, 32)
	output:	(None, 1)

Convolutional neural network



embedding_3: Embedding	input:	(None, None)
	output:	(None, None, 256)
lstm_1: LSTM	input:	(None, None, 256)
	output:	(None, 128)
dropout_1: Dropout	input:	(None, 128)
	output:	(None, 128)
dense_6: Dense	input:	(None, 128)
	output:	(None, 1)

Recurrent neural network with LSTM

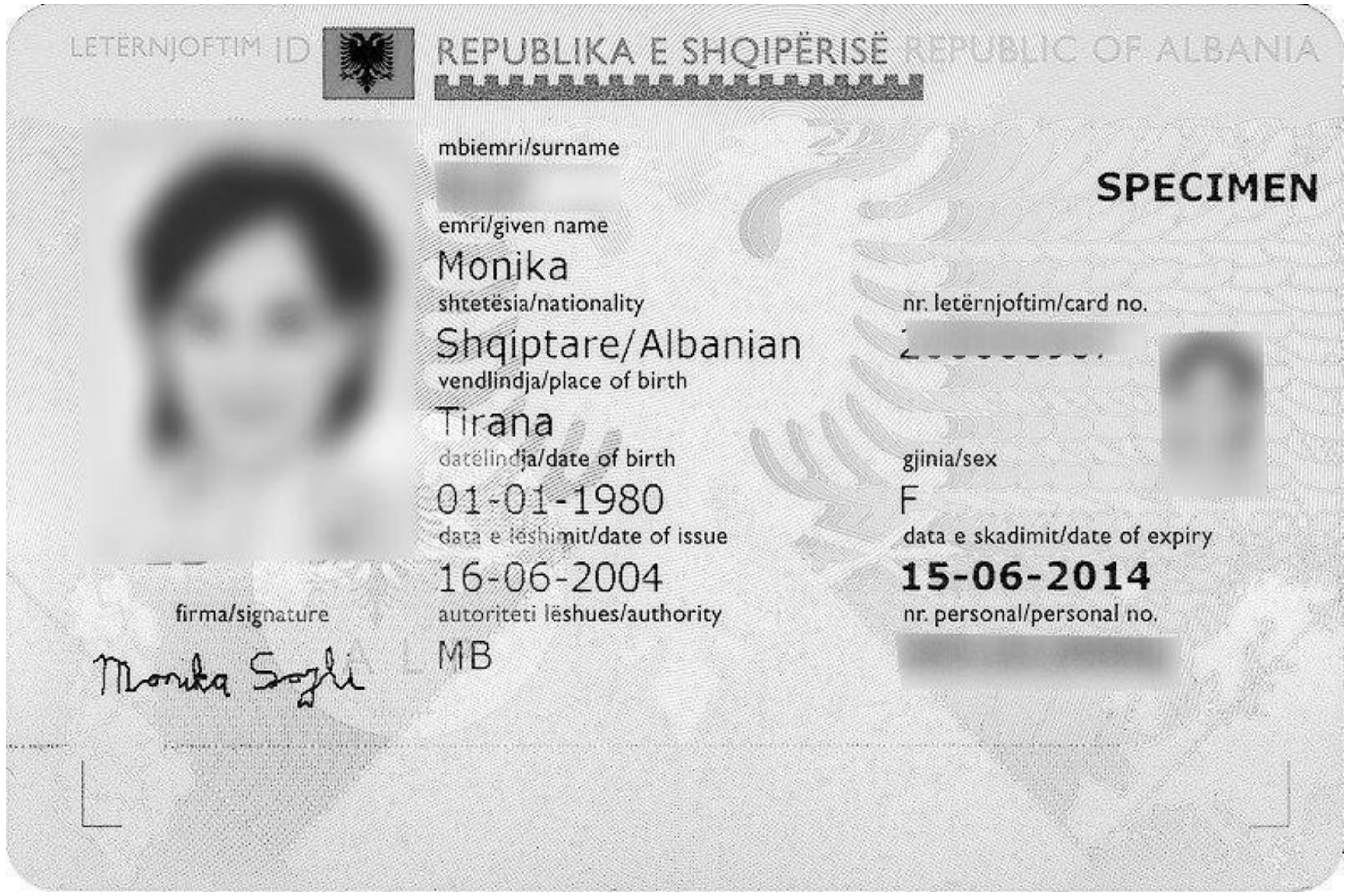


dense: Dense	input:	(None, 5000)
	output:	(None, 16)
dense_1: Dense	input:	(None, 16)
	output:	(None, 16)
dense_2: Dense	input:	(None, 16)
	output:	(None, 1)

Fully-connected neural network

EXPERIMENTS AND RESULTS

KEYWORD SPOTTING WITH TWO VARIANTS OF FEATURE EXTRACTION



LETËRNJOFTIM ID

REPUBLIKA E SHQIPËRISË REPUBLIC OF ALBANIA

mbiemri/surname

[redacted]

SPECIMEN

emri/given name

Monika

shtetësia/nationality

Shqiptare/Albanian

vendlindja/place of birth

Tirana

datëlindja/date of birth

01-01-1980

data e lëshimit/date of issue

16-06-2004

autoriteti lëshues/authority

MB

nr. letërnjoftim/card no.

[redacted]

gjinia/sex

F

data e skadimit/date of expiry

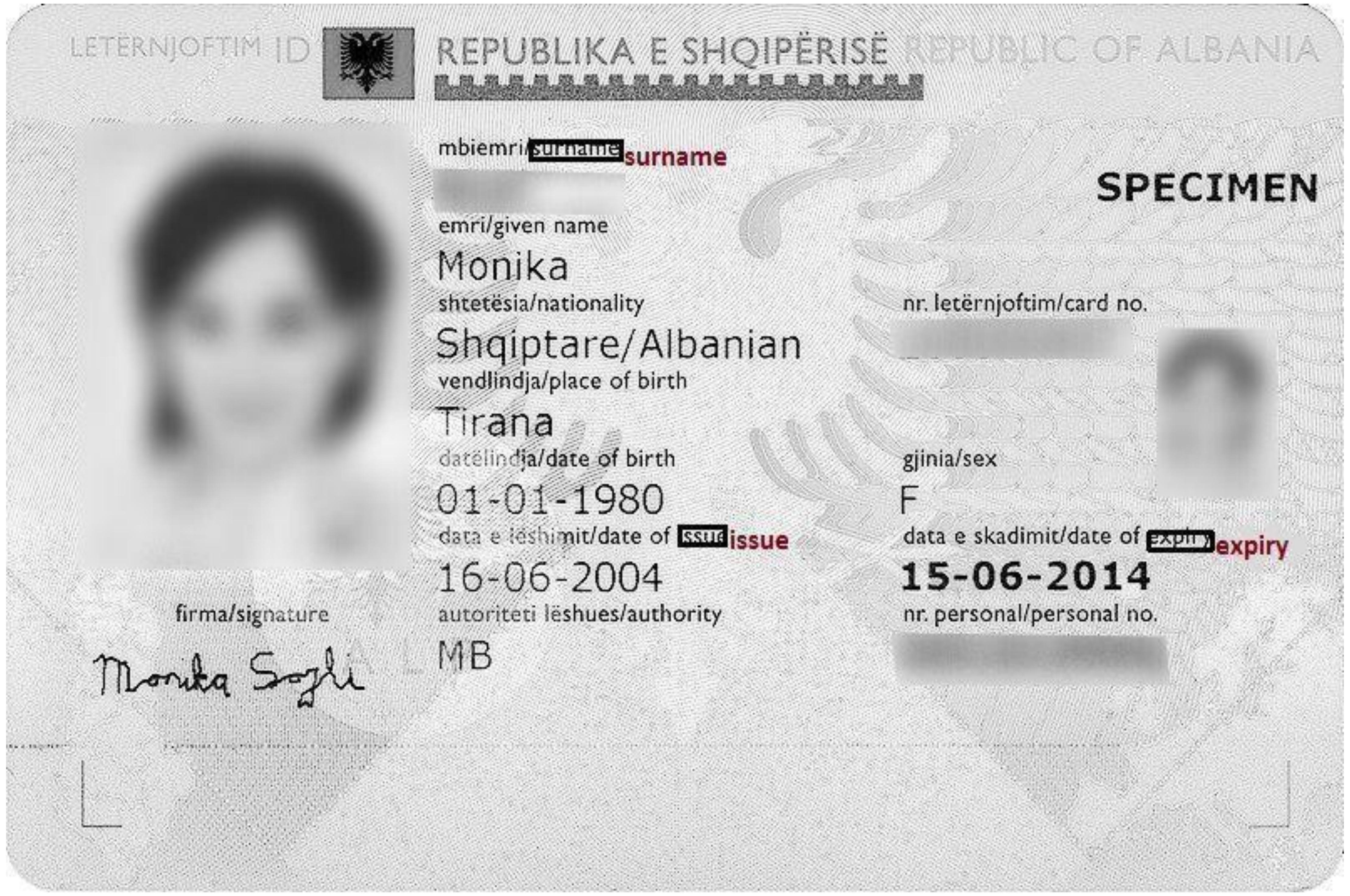
15-06-2014

nr. personal/personal no.

[redacted]

firma/signature

Monika Sogli



LETËRNJOFTIM ID

REPUBLIKA E SHQIPËRISË REPUBLIC OF ALBANIA

mbiemri/surname

[redacted]

SPECIMEN

emri/given name

Monika

shtetësia/nationality

Shqiptare/Albanian

vendlindja/place of birth

Tirana

datëlindja/date of birth

01-01-1980

data e lëshimit/date of issue

16-06-2004

autoriteti lëshues/authority

MB

nr. letërnjoftim/card no.

[redacted]

gjinia/sex

F

data e skadimit/date of expiry

15-06-2014

nr. personal/personal no.

[redacted]

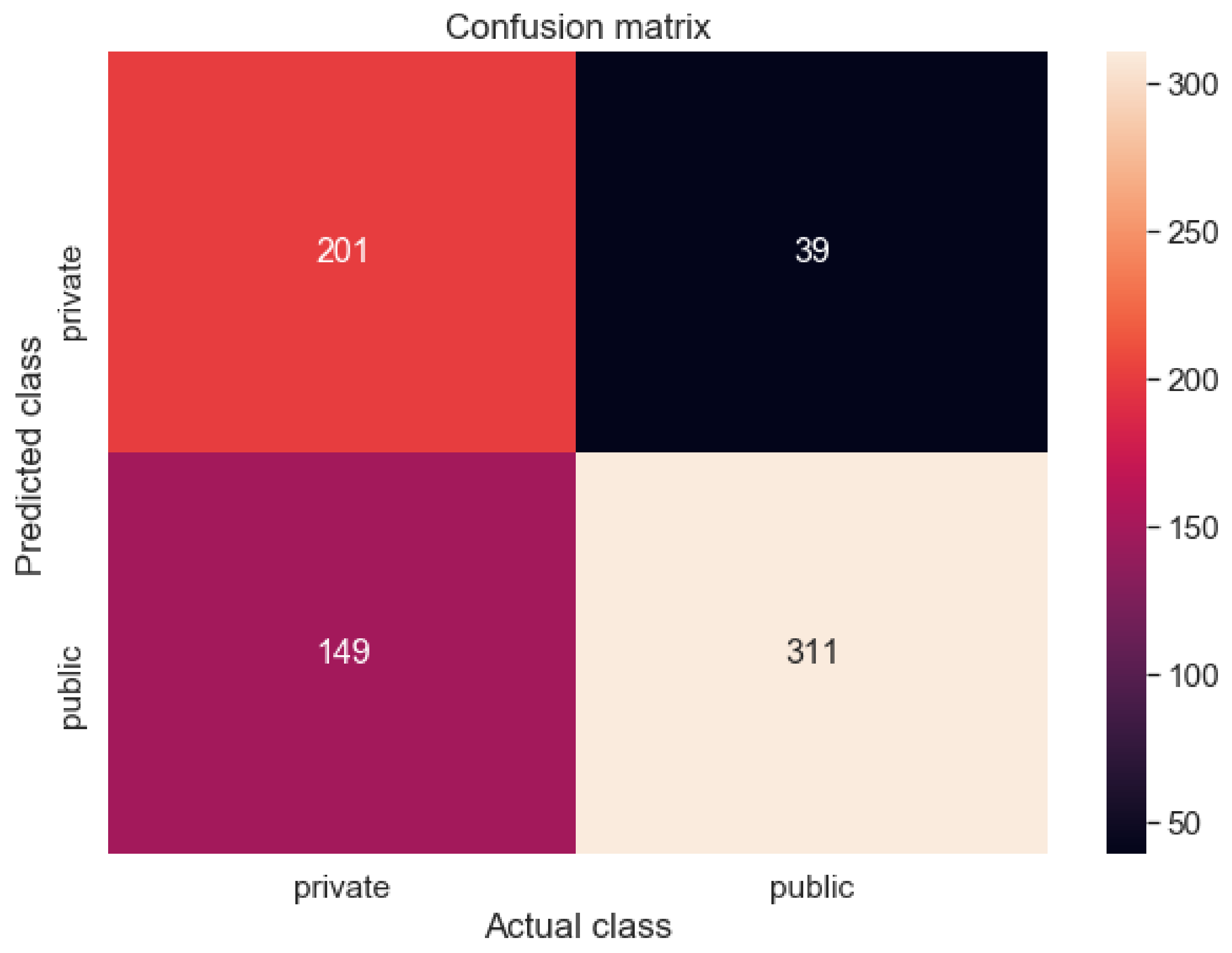
firma/signature

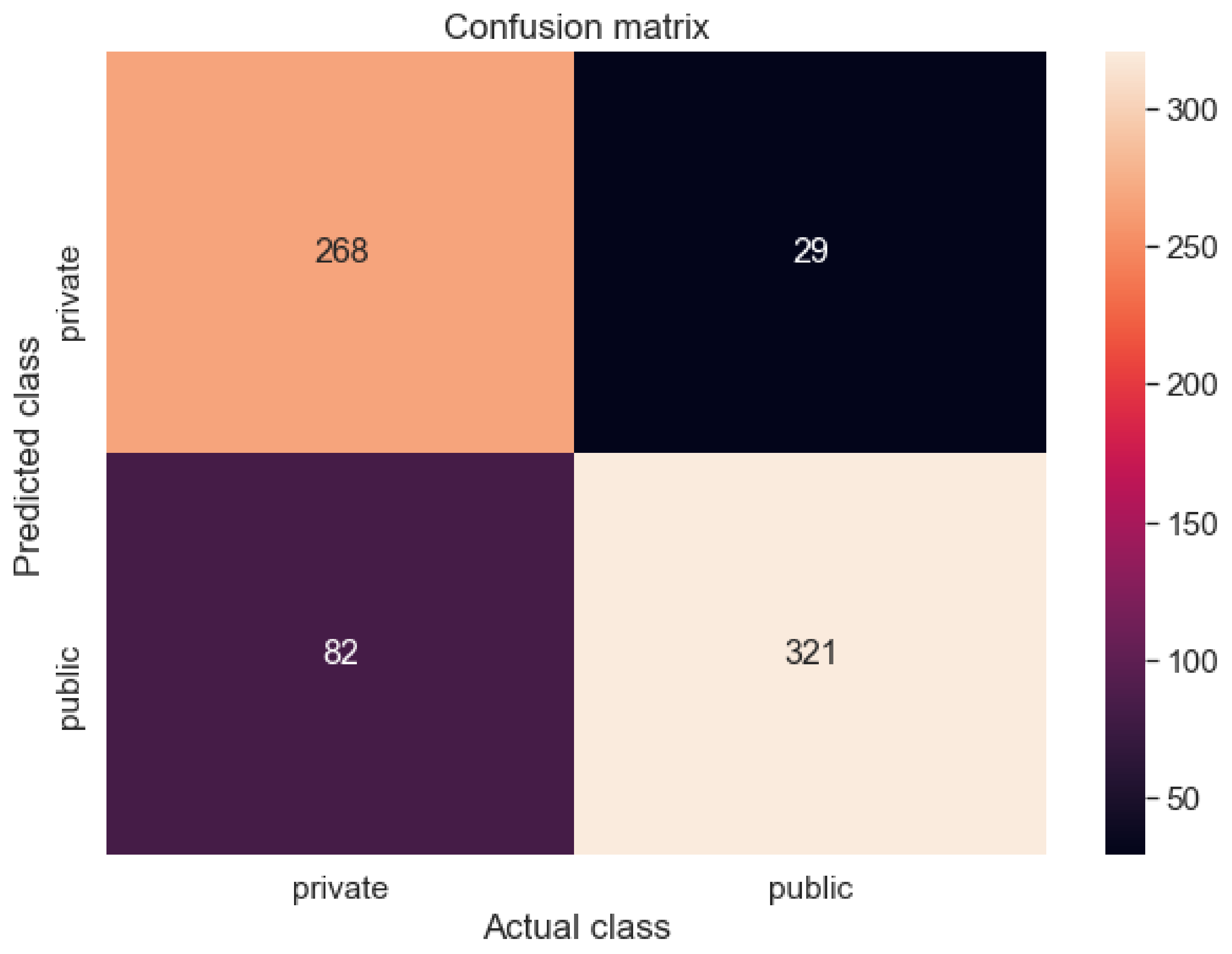
Monika Sogli

EXPERIMENTS AND RESULTS

KEYWORD SPOTTING WITH TWO VARIANTS OF FEATURE EXTRACTION

Metrics	Tesseract only	Tesseract+EAST detector
Accuracy, %	73.2	83.3
Precision, %	83.7	90.2
Recall, %	57.4	76.5
F1-score, %	67.6	82.8





EXPERIMENTS AND RESULTS

NEURAL NETWORKS

Fully-connected network		
	Tesseract only	**Tesseract +EAST text-detector**
1 hidden layer (16 hidden units)	95.4	95.1
2 hidden layers (16 hidden units)	94.9	95.7
3 hidden layers (16 hidden units)	95.7	96.2
2 hidden layers of 16 hidden units	94.9	95.7
2 hidden layers of 32 hidden units	94.9	96.2
2 hidden layers of 64 hidden units	94.5	95.7
2 hidden layers with ReLU activation	94.9	95.7
2 hidden layers with tanh activation	97.2	98.5

Recurrent network with LSTM		
	Tesseract only	**Tesseract +EAST text-detector**
1 LSTM layer with Output space=64	92.8	94.3
1 LSTM layer with Output space=128	95.7	96.2
2 LSTM layers with Output space=128	96.4	97.1

Convolutional network		
	Tesseract only	**Tesseract +EAST text-detector**
Embedding output_dim=32, 2 conv1d layers	83.9	85.6
Embedding output_dim=64, 2 conv1d layer	83.7	85.2
Embedding output_dim=128 , 2 conv1d layers	82.1	84.3
Embedding output_dim=128 , 1 conv1d layer	80.4	82
Embedding output_dim=128 , 2 conv1d layers	82.1	84.3
Embedding output_dim=128 , 3 conv1d layers	84.2	85.6

CONCLUSION

- We proposed a novel approach for classification of private and public scanned documents using EAST text detection and text recognition in the detected region based on Tesseract OCR library.
- We showed that the preliminarily text detection with EAST improves the quality of classification with both keyword spotting and neural nets.
- It was shown that deep FCNN with bag of most frequently used words outperforms more complicated network architecture like CNN and a model with LSTM recurrent layers
- Neural-network based classification decreases the error rate of keyword spotting on more than 15%.

FUTURE WORK

- Detection of private photos on mobile platforms.
- As the vast majority of private documents contain personal photos, face identification and clustering techniques should be applied to extract photos of closed friend and relatives.

THANK YOU FOR ATTENTION!