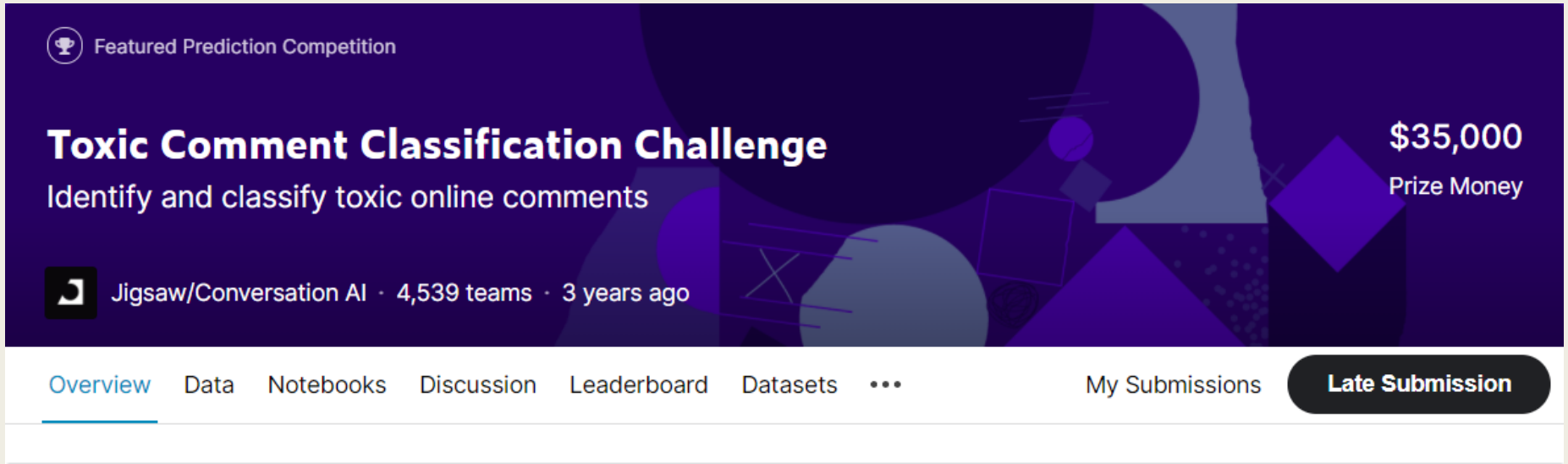
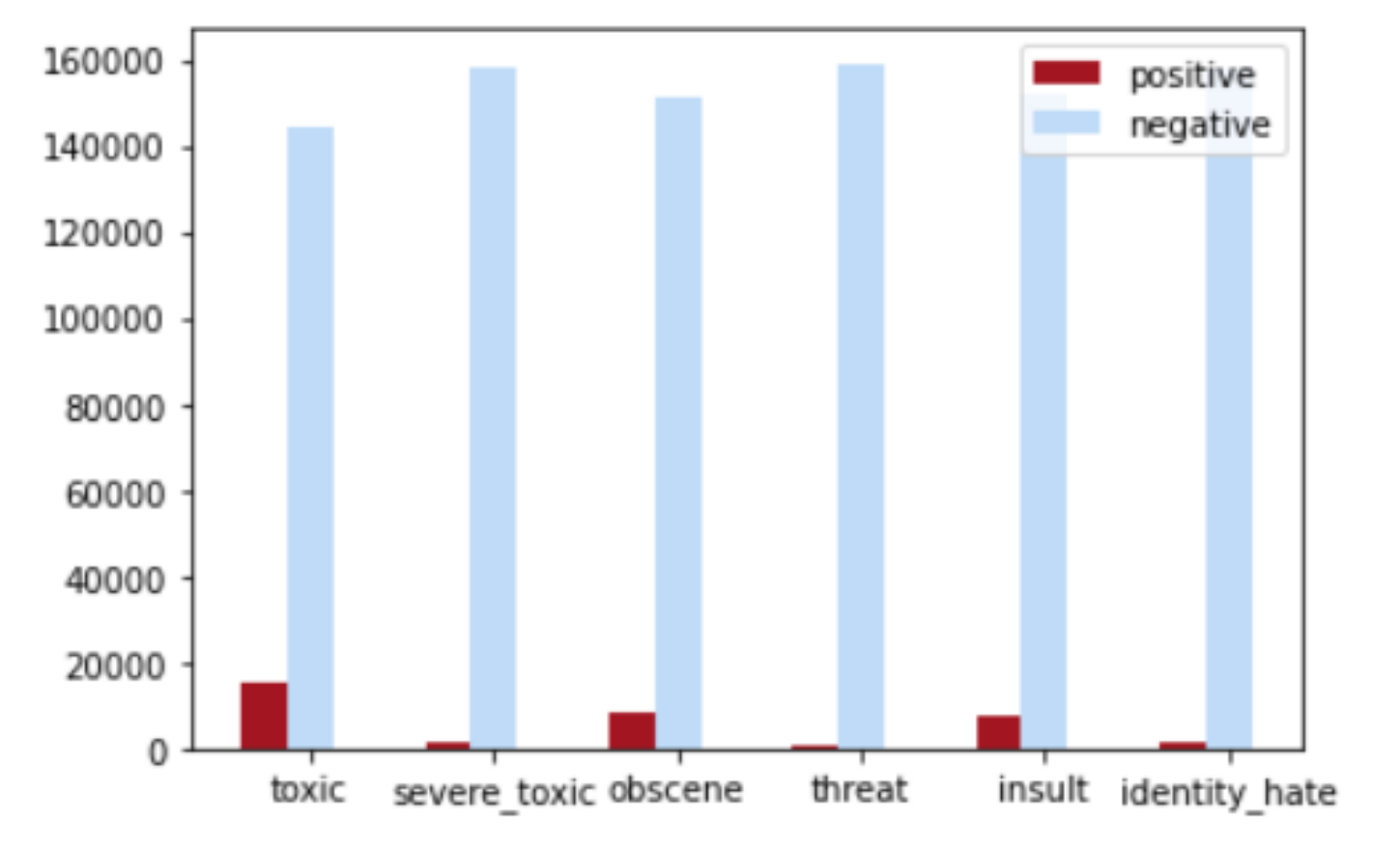
TOXIC COMMENT CLASSIFICATION

Simone Monti – 807994
Vittorio Maggio – 817034
University of Milano-Bicocca



https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge

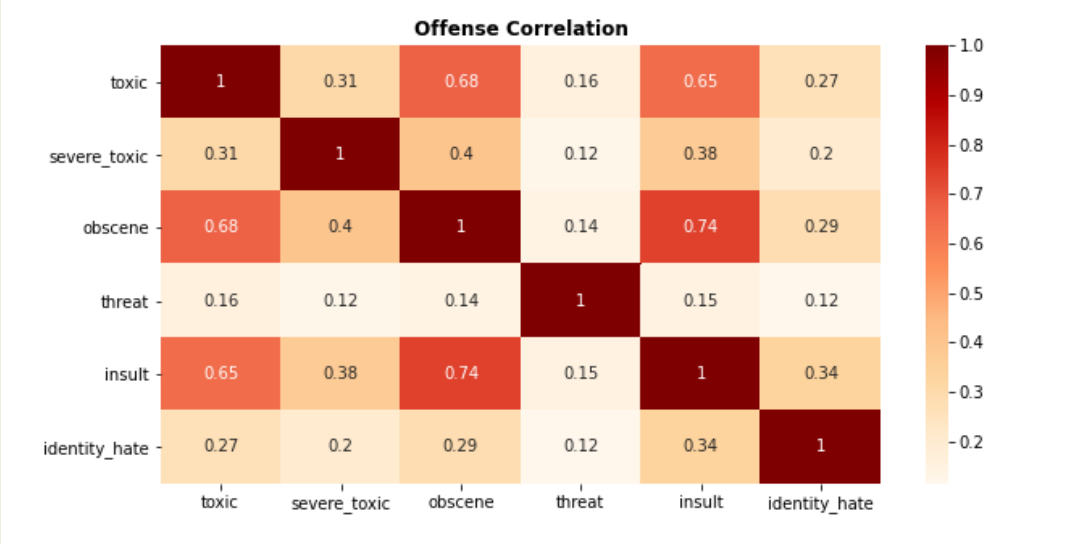
DEEP LEARNING APPROACH

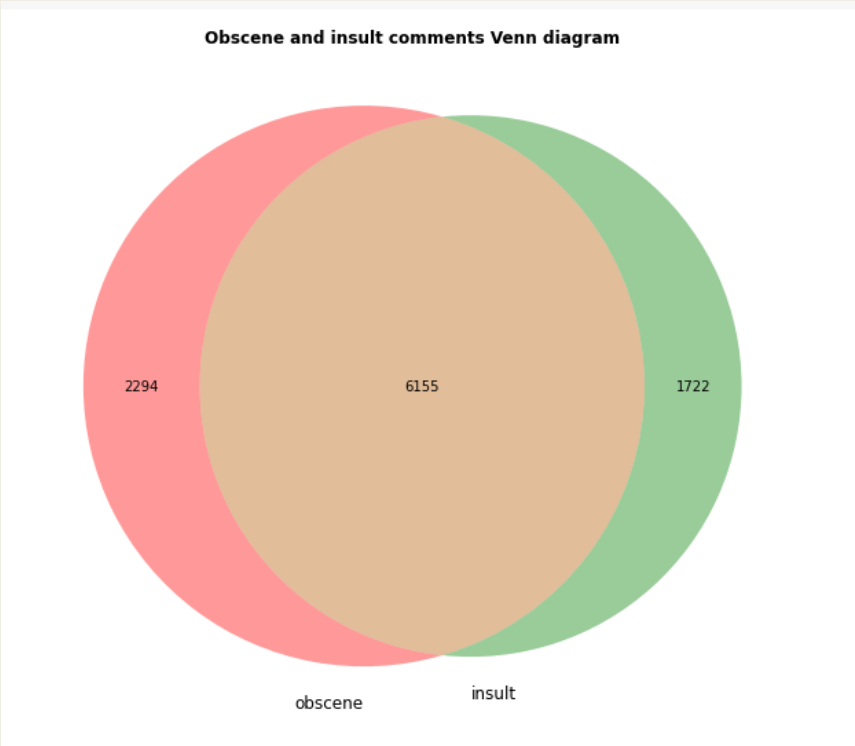


Total: 159571 comments

Dataset

- Toxic
- Severe toxic
- Obscene
- Threat
- Insult
- Identity hate





CORRELATION AMONG CATEGORIES

Identity hate

Severe toxic

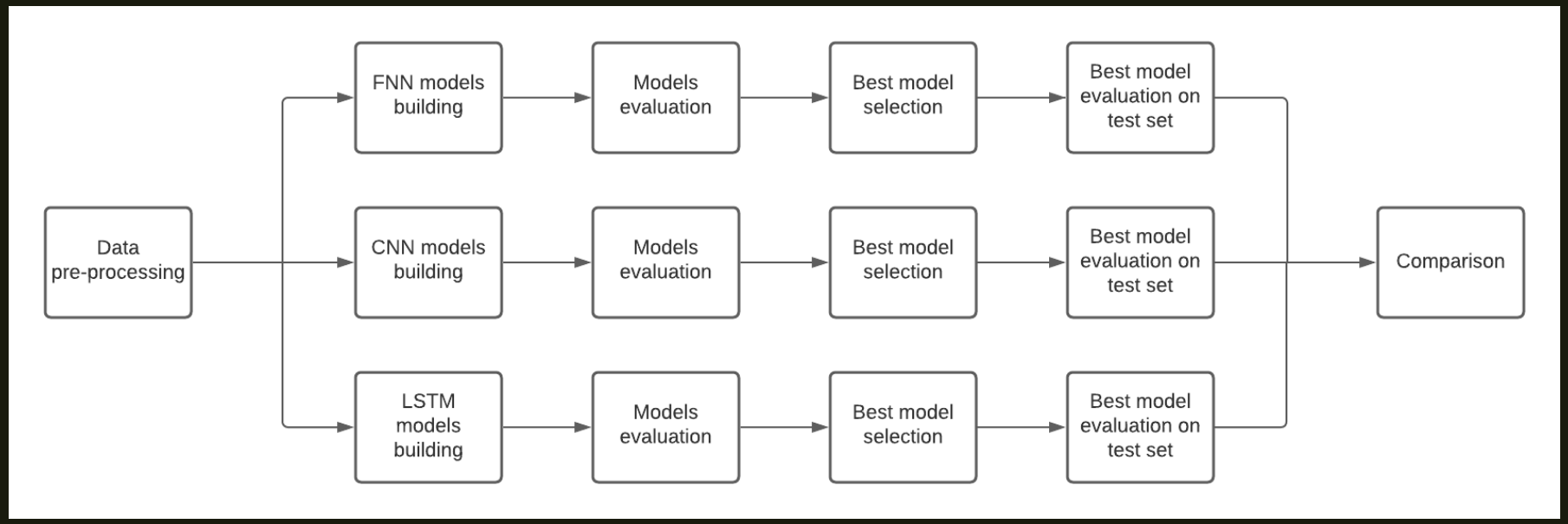
Toxic

Insult

Threat

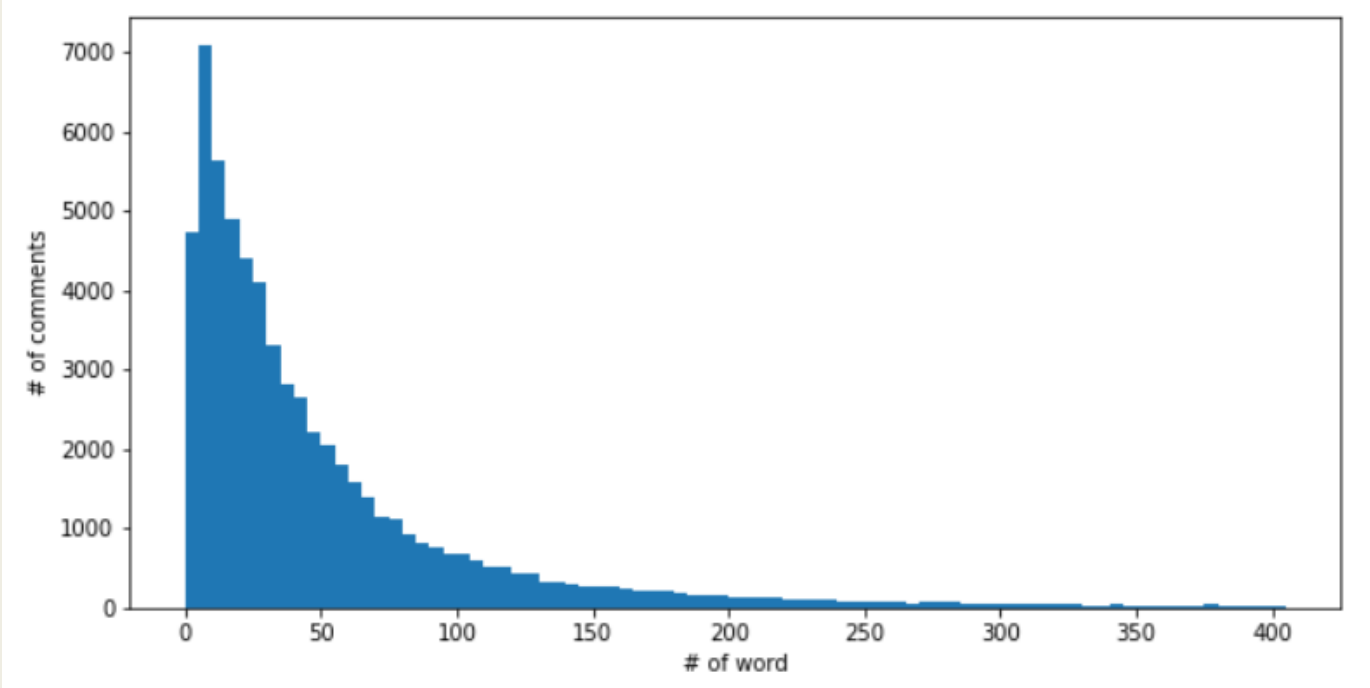
Obscene

PIPELINE


Data pre-processing
FNN models building
CNN models building
LSTM models building
Models evaluation
Models evaluation
Models evaluation
Best model selection
Best model selection
Best model selection
Best model evaluation on test set
Best model evaluation on test set
Best model evaluation on test set
Comparison

Pre-processing

- Text tokenization
- Sequence padding
  - *Length: 200*



FNN – Architectures

FNN

- An Embedding layer (input_dim: 200, output_dim=128),
- GlobalMaxPool1D,
- a Dense layer formed by X* neurons and with Relu as activation function,
- Dropout (rate: 0.3),
- a Dense layer formed by 6 neurons with Sigmoid function as activation function (the output is composed by 6 different binary values)

*X = (20 , 30, 40, 50)

FNN – additional layer

- An Embedding layer (input_dim: 200, output_dim=128),
- GlobalMaxPool1D,
- a Dense layer formed by 40 neurons and with Relu as activation function,
- Dropout (rate: 0.3),
- a Dense layer formed by 20 neurons and with Relu as activation function,
- Dropout (rate: 0.3),
- a Dense layer formed by 6 neurons with Sigmoid function as activation function (the output is composed by 6 different binary values).

CNN – Architectures

- An Embedding layer (input_dim: 200, output_dim=128),
- SpatialDropout1D (rate: 0.3),
- layer Conv1D with X* filters of dimension 4x1 and Relu as activation function,
- BatchNormalization,
- GlobalMaxPool1d,
- Dropout (rate: 0.3),
- a Dense layer formed by 20 neurons and with Relu as activation function,
- a Dense layer formed by 6 neurons with Sigmoid function as activation function (the output is composed by 6 different binary values).

* X = (20, 30, 40, 50, 60)

LSTM – Architectures

Bidirectional

- Embedding layer (input_dim: 200, output_dim=128)
- SpatialDropout1D (rate: 0.3),
- Bi-directional LSTM layer with X* neurons and with Tanh as the activation function, and Sigmoid as recurrent activation,
- BatchNormalization,
- GlobalMaxPool1d,
- Dropout (rate: 0.3),
- A Dense layer with 15 neurons and Relu as actiovation function,
- a Dense layer formed by 6 neurons with Sigmoid function as activation function (the output is composed by 6 different binary values)

* X = (25, 50)

Standard

- Embedding layer (input_dim: 200, output_dim=128)
- SpatialDropout1D (rate: 0.3),
- LSTM layer with 25 neurons with Tanh as activation function, and Sigmoid as recurrent activation,
- BatchNormalization,
- GlobalMaxPool1d,
- layer Dropout (rate: 0.3),
- A Dense layer with 15 neurons and Relu as actiovation function,
- a Dense layer formed by 6 neurons with Sigmoid function as activation function (the output is composed by 6 different binary values).

Training

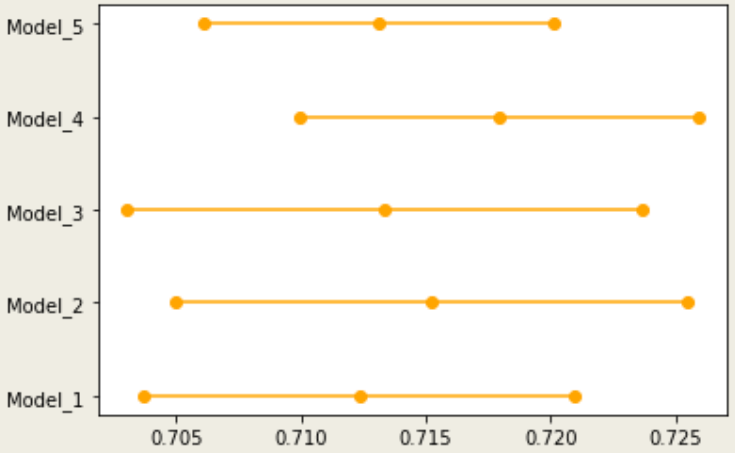
HyperParametes:

- Epochs: 10;
- Batch size: 256;
- Optimizer: Adam
  (learning rate: 0.01);
- Loss: Binary Cross Entropy;

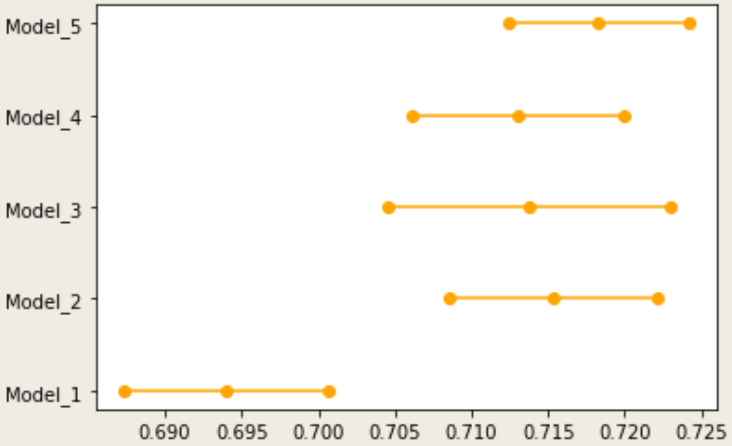
Early Stopping

- Monitor: validation loss;
- Patience: 2;

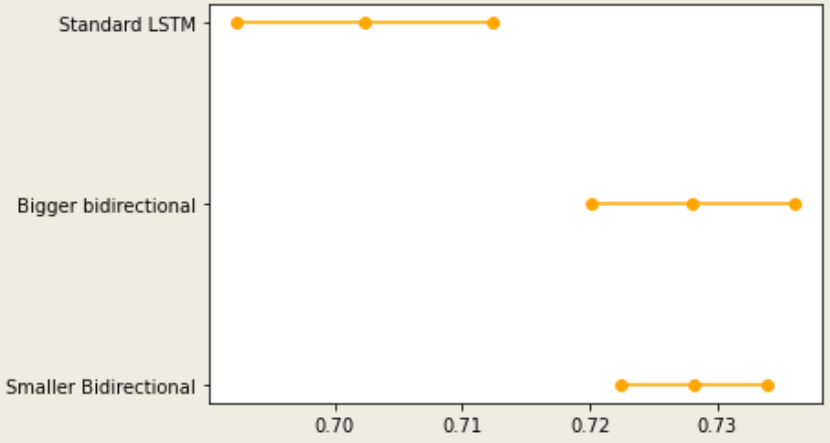
FNN



CNN



RNN



F1 SCORE – VALIDATION SET

95% confidence interval - 10 folds crossvalidation

Step

- The 3 best models selected:
  - *FNN: Model_4*
  - *CNN: Model_5*
  - *RNN: Smaller bidirectional*
- Training on the full training set (Train-Val 80%-20%)
- Evaluation on the test set

FINAL COMPARISON

- High accuracy due to high unbalanced dataset
- F1 score much lower
- No dominance among selected networks
- The models are not able to predict categories with few examples in the training

FNN

	Precision	Recall	F1-score	support
Toxic	0.59	0.79	0.67	6090
Severe_toxic	0.39	0.32	0.35	367
Obscene	0.67	0.72	0.69	3691
Threat	0.37	0.10	0.16	211
Insult	0.63	0.63	0.63	3427
Identity_hate	0.75	0.12	0.21	712
Micro avg	0.61	0.68	0.64	14498

Loss: 0.072
Accuracy: 0.995

CNN

	Precision	Recall	F1-score	support
Toxic	0.56	0.79	0.65	6090
Severe_toxic	0.29	0.38	0.33	367
Obscene	0.60	0.74	0.66	3691
Threat	1.00	0.00	0.00	211
Insult	0.56	0.64	0.60	3427
Identity_hate	0.50	0.01	0.03	712
Micro avg	0.56	0.68	0.62	14498

Loss: 0.079
Accuracy: 0.978

RNN

	Precision	Recall	F1-score	support
Toxic	0.56	0.83	0.67	6090
Severe_toxic	0.52	0.06	0.11	367
Obscene	0.69	0.70	0.69	3691
Threat	1.00	0.00	0.00	211
Insult	0.64	0.60	0.62	3427
Identity_hate	1.00	0.00	0.00	712
Micro avg	0.57	0.71	0.63	14498

Loss: 0.073
Accuracy: 0.998

Possible future developments

- Using of Data Augmentation technique
- Using of pre-trained embedding layers such as:
  - *Word2Vec*
  - *Glove*
  - *BERT*

THANK YOU FOR YOUR ATTENTION

Simone Monti – 807994
Vittorio Maggio – 817034
University of Milano-Bicocca