Advanced Machine Learning A.A. 2020-2021

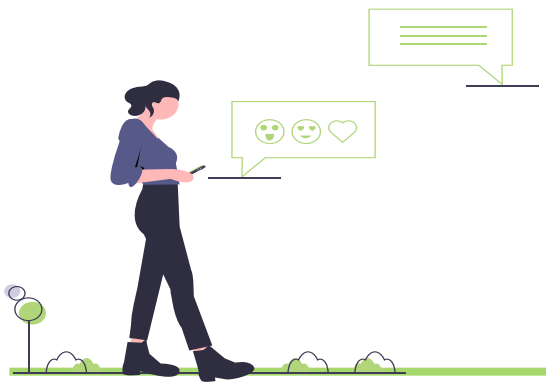
Toxic Comment Classification

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

Roberta Bensi - 861555
Martina Roberta Cecchetto - 852566
Giuseppe Monea - 850432

AGENDA

1. DATASET
2. DATA PREPROCESSING
3. METHODOLOGICAL APPROACH
4. RESULTS
5. CONCLUSIONS



DATASET DESCRIPTION

The train set contains a total of 159.571 comments, with 1 or 0 for each of these labels :

- **toxic** 16.225
- **severe_toxic** 1.595
- **obscene** 8.449
- **threat** 478
- **insult** 7.877
- **identity_hate** 1.405

DATA PREPROCESSING

- Replacement of “0” in *toxic* with “1” if the comment belongs to at least one of the five subcategory of *toxic*
- Lower case
- Transformation of contracted form in long form
- Removal of punctuation and stop words
- Tokenization

DATA PREPROCESSING

Class imbalance problem:

Data Augmentation

- **Synonym Replacement**
- **Random Swap**
- **Random Deletion (p = 0,2)**

Categoria	Dataset originale	Dataset aumentato
Toxic	16225	109108
Severe toxic	1595	27584
Obscene	8449	67573
Threat	478	21010
Insult	7877	65339
Identity hate	1405	24610

DATA PREPROCESSING

Class imbalance problem:

Subsampling

- Selection of all the toxic comments
- Random sampling of non toxic comments
- Dataset of 30.000 observation

FIRST METHODOLOGICAL APPROACH

Multi-label classification:

- **Bidirectional LSTM**
- **Bidirectional GRU**
- **CNN**

Common parameters:

- Loss function: binary crossentropy
- Optimizer: Adamax
- Early stopping for validation loss

FIRST METHODOLOGICAL APPROACH

Bidirectional LSTM

Embedding layer

- Input dim=60.000
- Output dim=128
- Input length=150

Bidirectional LSTM

- 128 units
- Dropout rate=0.2

Dense layer

- 6 units
- Sigmoid function

FIRST METHODOLOGICAL APPROACH

Bidirectional GRU

Embedding layer

- Input dim=60.000
- Output dim=64
- Input length=150

Bidirectional GRU

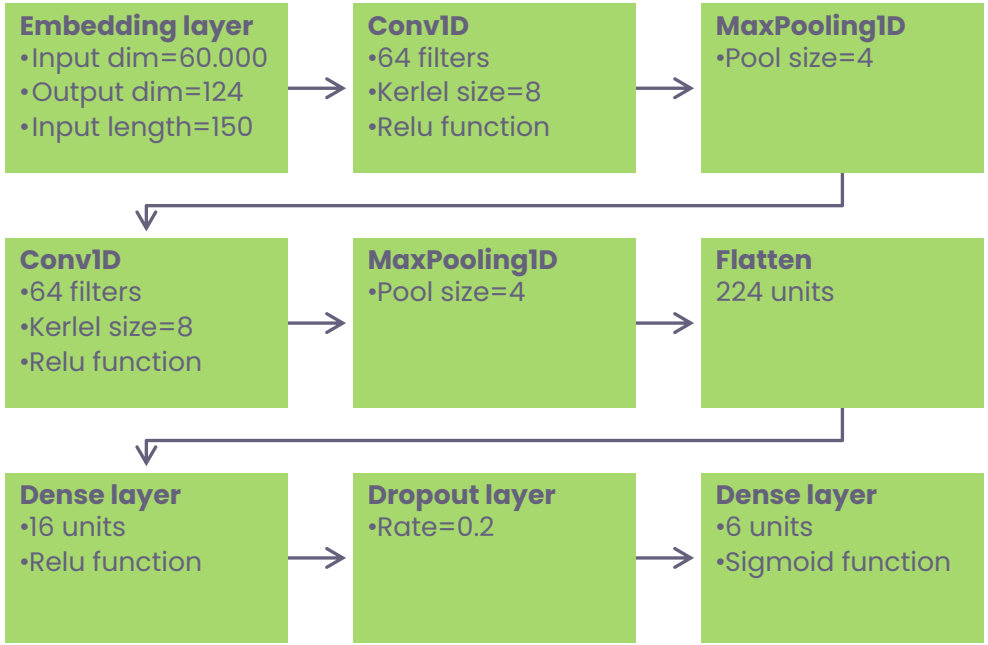
- 128 units
- Dropout rate=0.2

Dense layer

- 6 units
- Sigmoid function

FIRST METHODOLOGICAL APPROACH

CNN



SECOND METHODOLOGICAL APPROACH

- **Binary classification of *toxic/non toxic***
- **Multi-label classification of toxic category**

Common aspects:

- Bidirectional LSTM
- Loss function: binary crossentropy
- Optimizer: Adamax
- Early stopping for validation loss

SECOND METHODOLOGICAL APPROACH

Bidirectional LSTM – Binary

Embedding layer

- Input dim=60.000
- Output dim=128
- Input length=150

Bidirectional LSTM

- 128 units
- Dropout rate=0.2

Dense layer

- 1 unit
- Sigmoid function

SECOND METHODOLOGICAL APPROACH

Bidirectional LSTM – Multi-label

Embedding layer

- Input dim=60.000
- Output dim=128
- Input length=150

Bidirectional LSTM

- 128 units
- Dropout rate=0.3

Dense layer

- 5 units
- Sigmoid function

RESULTS

Given the imbalaced nature of the dataset, the following metrics have been used:

$$\text{Recall} = \frac{TP}{TP+FN}$$

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{F1-score} = \frac{TP}{TP+\frac{1}{2}(FN+FP)}$$

RESULTS OF FIRST APPROACH

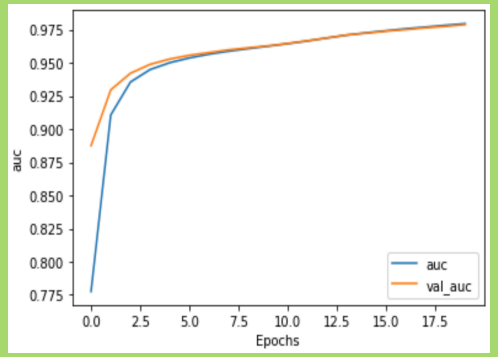
Multi-label model on the test set with 63978 comments

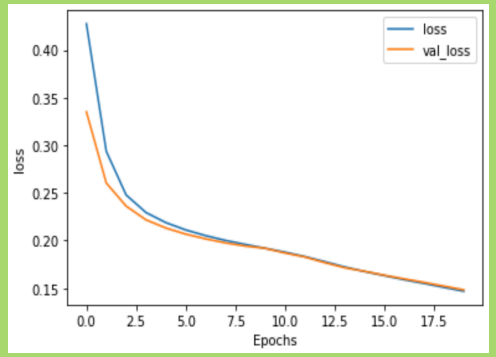
	LSTM			GRU			CNN			
	precision	recall	f1 score	precision	recall	F1 score	precision	recall	F1 score	support
toxic										6090
subsampling	0.39	0.93	0.55	0.38	0.93	0.54	0.25	0.83	0.39	
data augmentation	0.42	0.92	0.57	0.44	0.91	0.59	0.30	0.78	0.43	
severe toxic										367
subsampling	0.44	0.26	0.32	0.40	0.27	0.32	0.49	0.17	0.25	
data augmentation	0.27	0.68	0.39	0.29	0.62	0.40	0.29	0.53	0.38	
obscene										3691
subsampling	0.64	0.74	0.68	0.61	0.75	0.68	0.46	0.53	0.49	
data augmentation	0.63	0.77	0.69	0.60	0.78	0.68	0.52	0.50	0.51	
threat										211
subsampling	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
data augmentation	0.27	0.67	0.38	0.22	0.66	0.34	0.11	0.00	0.01	
insult										3427
subsampling	0.59	0.60	0.60	0.57	0.60	0.59	0.45	0.48	0.47	
data augmentation	0.55	0.68	0.61	0.52	0.67	0.58	0.48	0.46	0.47	
identity_hate										712
subsampling	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
data augmentation	0.59	0.24	0.34	0.45	0.11	0.17	0.48	0.02	0.04	

RESULTS OF FIRST APPROACH

During training the following metrics were evaluated:

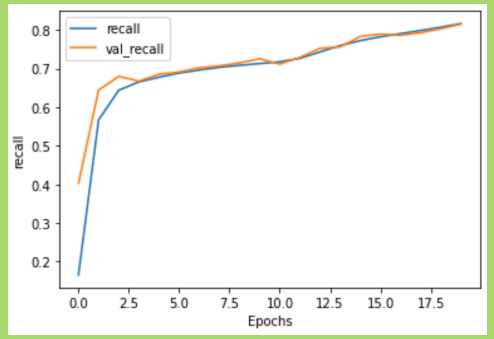
AUC

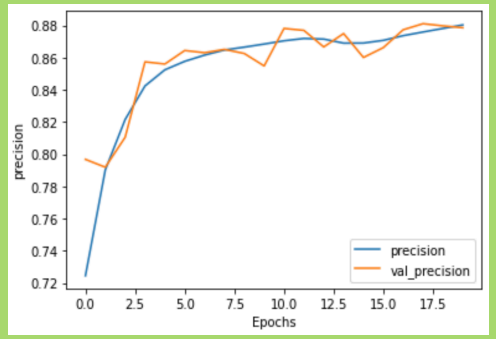




LOSS

RECALL





PRECISION

RESULTS OF SECOND APPROACH

Binary model on the test-set with 63978 comments

Label	precision	recall	F1 score	support
non toxic (0)	0.99	0.87	0.93	57888
toxic (1)	0.43	0.92	0.59	6090

Multi-label model on the test-set with 6090 comments

Label	precision	recall	F1 score	support
severe_toxic	0.23	0.44	0.30	367
obscene	0.78	0.76	0.77	3626
threat	0.46	0.48	0.47	205
insult	0.70	0.67	0.68	3342
identity_hate	0.60	0.51	0.55	693

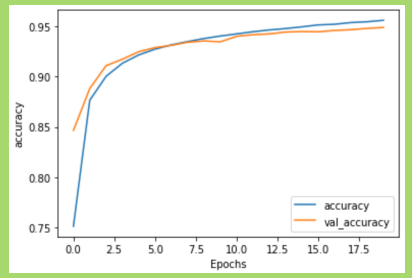
RESULTS OF SECOND APPROACH

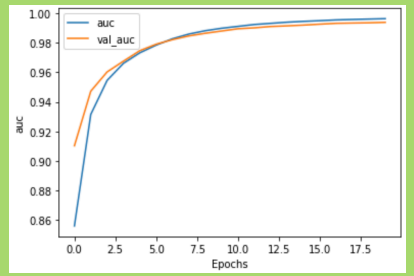
ACCURACY

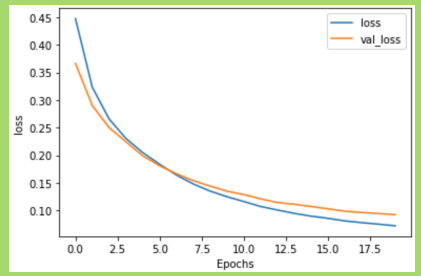
AUC

LOSS

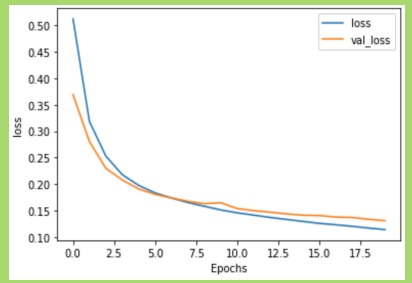
BINARY MODEL

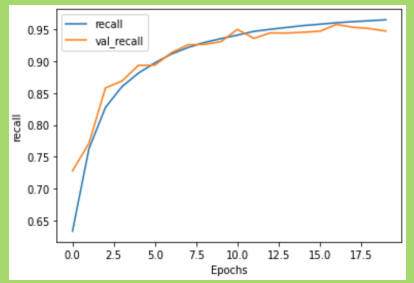


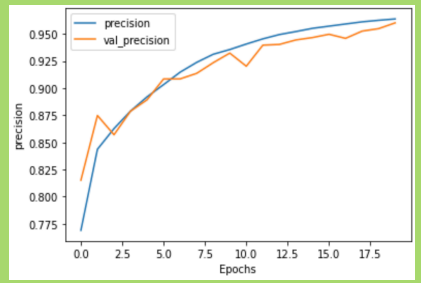




MULTI-LABEL MODEL







LOSS

RECALL

PRECISION

CONCLUSIONS

Evaluation criteria in order of importance:

1. F1-score
2. Recall
3. Precision

Binary + multi-label model

Future development:

- Prediction of level of toxicity
- Implementation in other languages

THANK YOU FOR THE ATTENTION!