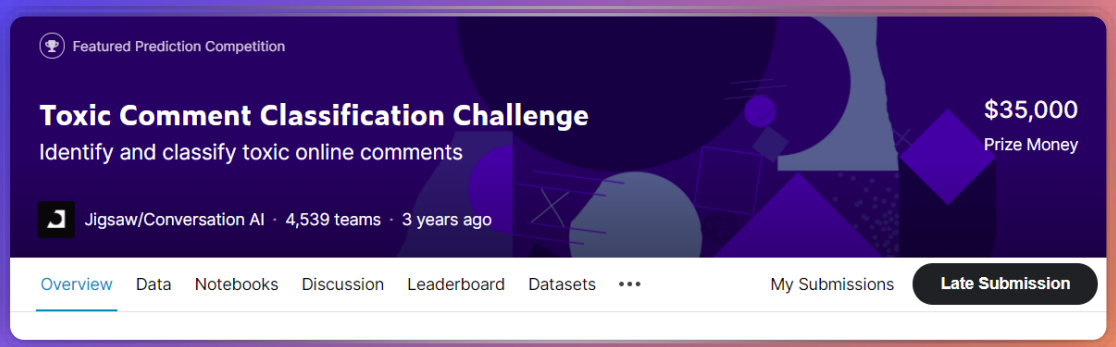
TOXIC COMMENT CLASSIFICATION

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

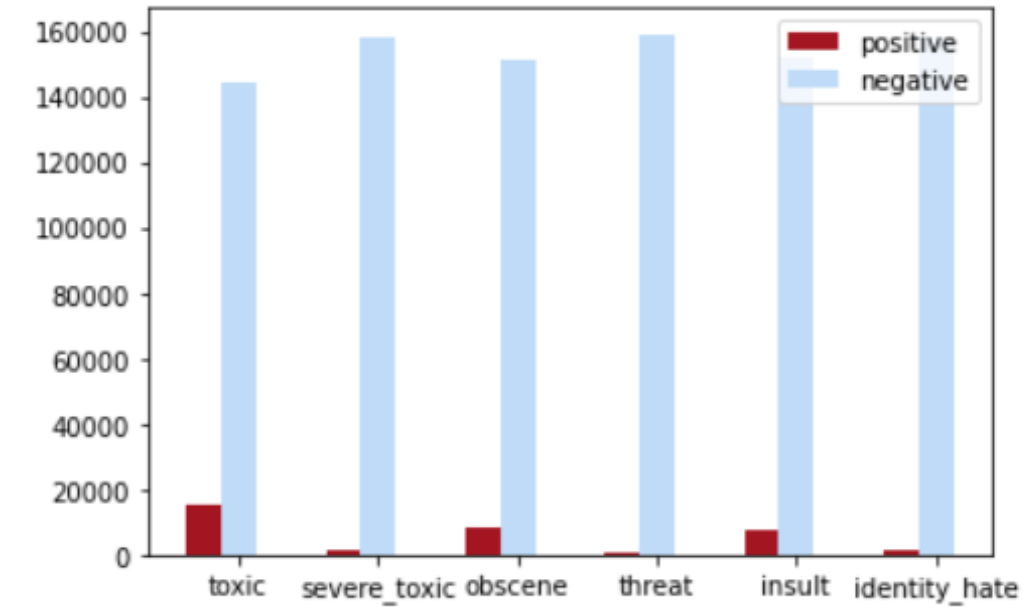


Deep learning approach

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

DATASET EXPLORATION

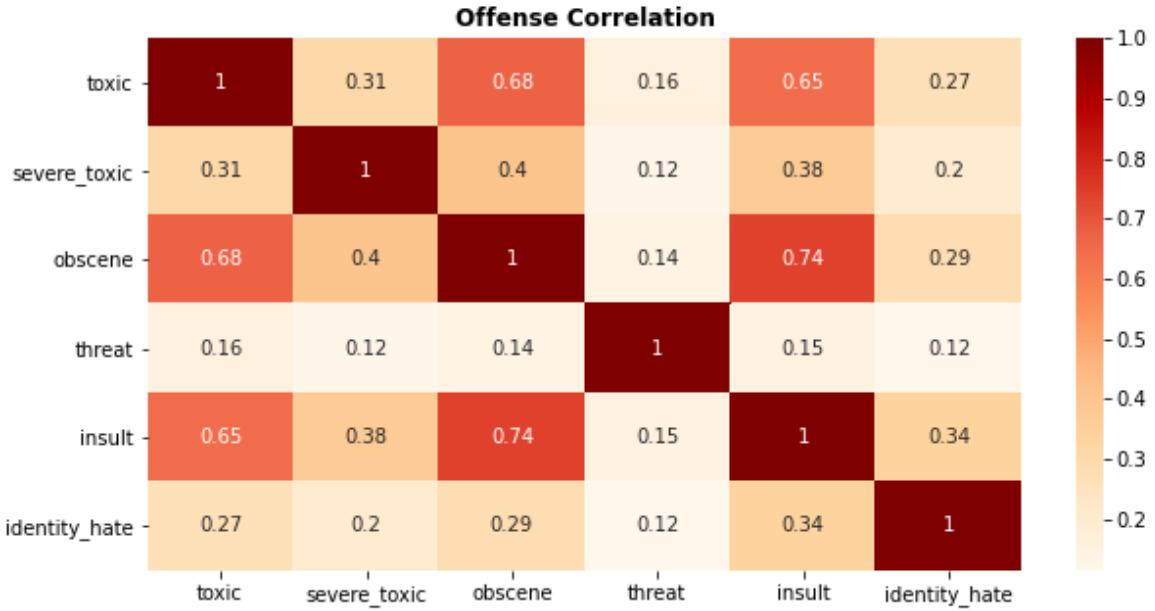
Dataset

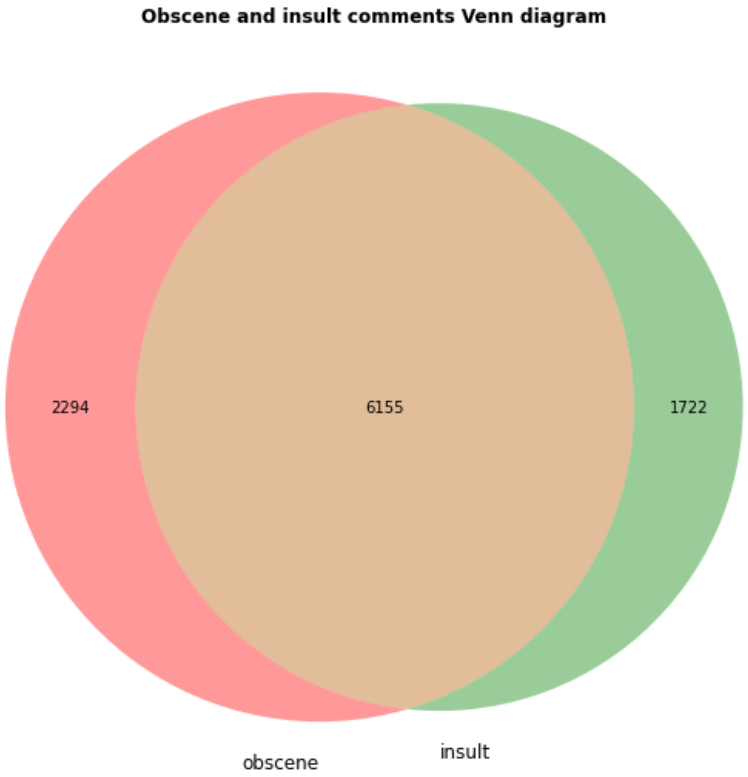


Total: 159571 comments

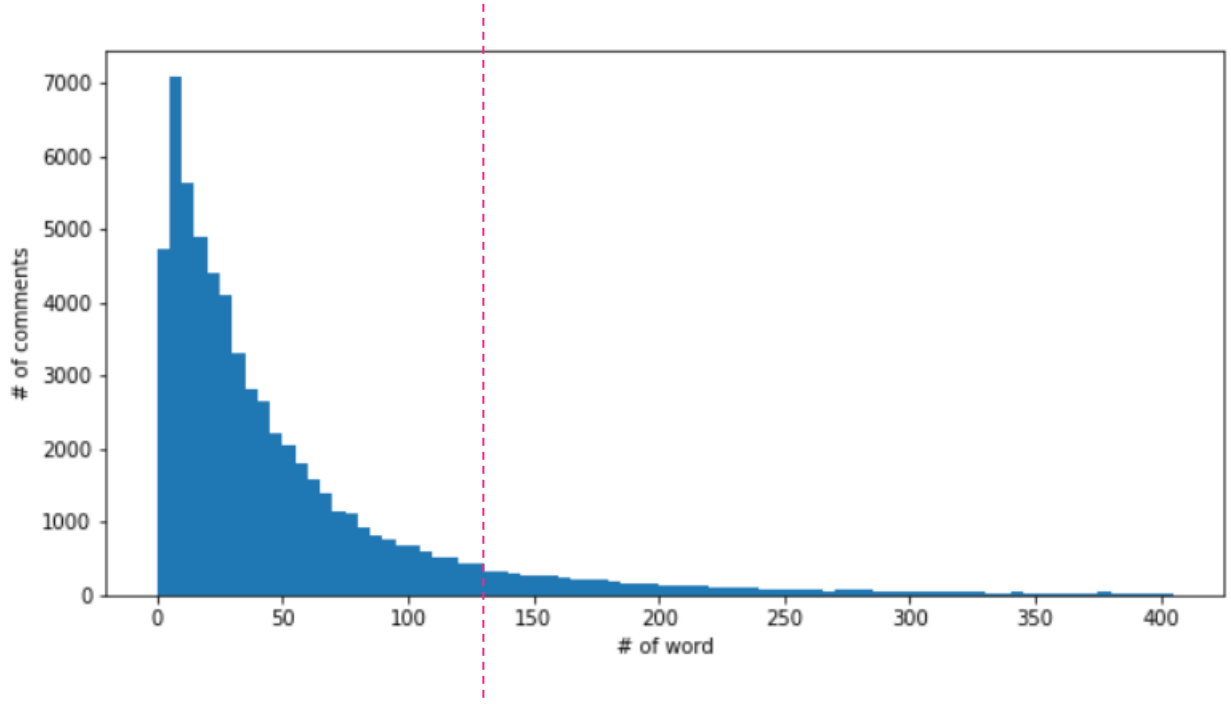
- Toxic
- Severe toxic
- Obscene
- Threat
- Insult
- Identity hate

Correlation among categories





Comments length



128
Optimal length

WordCloud


fuck
suck
shit
nig
lik
ass
hat
faggot
wikiped
edit
go
Toxic


fuck
suck
bitch
cocksuck
ass
shit
die
go
faggot
nig
cunt
wikiped
Severe toxic

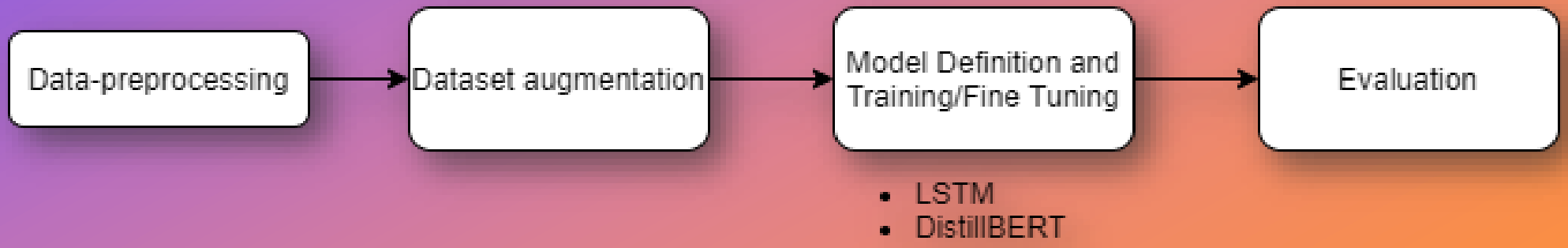

fuck
suck
nig
faggot
ass
shit
jew
go
edit
stupid
cunt
bitch
hat
idiot
Insult


nigg
fat
jew
fuck
die
gay
faggot
suck
shit
hug
mex
cunt
stupid
Identity hate


die
kill
ass
fuck
must
going
block
wikiped
jim
wal
Threat


fuck
suck
ass
nig
faggot
shit
bitch
go
cock
dick
Obscene

PIPELINE


Data-preprocessing
Dataset augmentation
Model Definition and Training/Fine Tuning
Evaluation
LSTM
DistillBERT

DATASET PREPARATION

Pre-processing

- Lowercase
- Punctuation and spacing removal

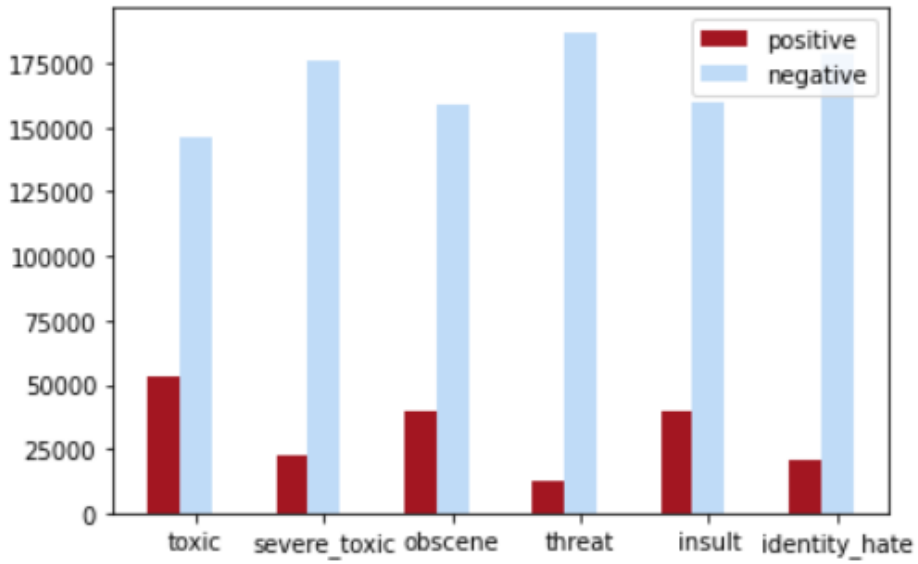
Explanation\nWhy the edits made under my usern...
D'aww! He matches this background colour I'm s...
Hey man, I'm really not trying to edit war. It...
"\nMore\nI can't make any real suggestions on ...
You, sir, are my hero. Any chance you remember...

explanation why the edits made under my userna...
daww he matches this background colour im seem...
hey man im really not trying to edit war its j...
more i cant make any real suggestions on impro...
you sir are my hero any chance you remember wh...

Dataset augmentation: EDA

https://github.com/jasonwei20/eda_nlp

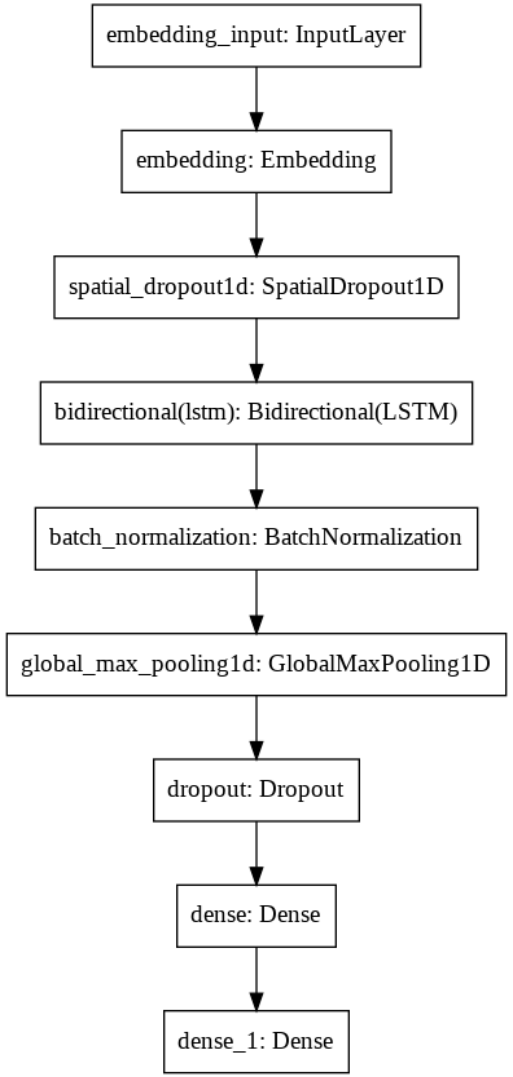
- **Synonym Replacement (SR):** Randomly choose *n* words from the sentence that are not stop words. Replace each of these words with one of its synonyms chosen at random.
- **Random Insertion (RI):** Find a random synonym of a random word in the sentence that is not a stop word. Insert that synonym into a random position in the sentence. Do this *n* times.
- **Random Swap (RS):** Randomly choose two words in the sentence and swap their positions. Do this *n* times.
- **Random Deletion (RD):** For each word in the sentence, randomly remove it with probability *p*.
- Used with *threat(x20), severe_toxic(x10), identity_hate(x10)*
- 199.131 elements



MODELS

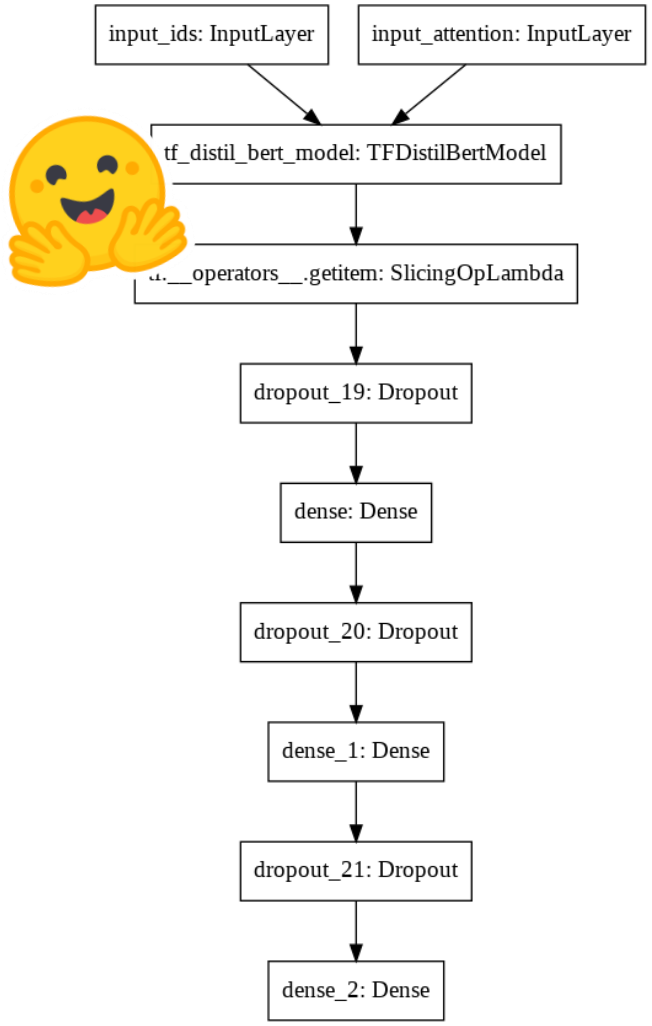
LSTM – Architecture

- Embedding layer (input_dim: 128, output_dim=128)
- SpatialDropout1D (rate: 0.3),
- Bi-directional LSTM layer with 42 neurons and with Tanh as activation function, and Sigmoid as recurrent activation,
- BatchNormalization,
- GlobalMaxPool1d,
- Dropout (rate: 0.3),
- Dense layer with 18 neurons and Relu as activation function,
- Dense layer formed by 6 neurons with Sigmoid function as activation function (the output is composed by 6 different binary values)



DistillBERT Fine-tuned – Architecture

- DistilBERT base uncased (6-layer, 768-hidden, 12-heads)
- Dropout (rate: 0.2)
- Dense layer with 256 neurons and Relu as activation function
- Dropout (rate: 0.2)
- Dense layer with 32 neurons and Relu as activation function
- Dopout (rate: 0.2)
- Dense layer formed by 6 neurons with Sigmoid function as activation function (the output is composed by 6 different binary values)



Training

LSTM

- **HyperParametes:**
  - Epochs: 50;
  - Batch size: 256;
  - Validation split: 0.2
  - Optimizer: Adam (learning rate: 0.0001);
  - Loss: Binary Cross Entropy;
- **Early Stopping:**
  - Monitor: validation loss;
  - Patience: 3,
  - restore_best_weights: True;

DistillBERT Fine-tuned

- **HyperParametes:**
  - Epochs: 6;
  - Batch size: 64;
  - Validation split: 0.2
  - Optimizer: Adam (learning rate: 0.00005);
  - Loss: Binary Cross Entropy;
- **Early Stopping:**
  - Monitor: validation loss;
  - Patience: 2,
  - restore_best_weights: True;

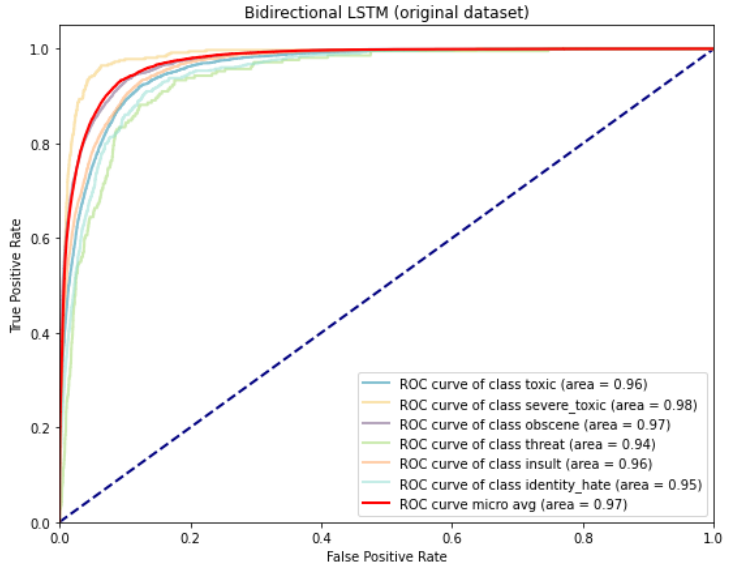
EVALUATION

Evaluated on the test dataset provided by the competition

LSTM

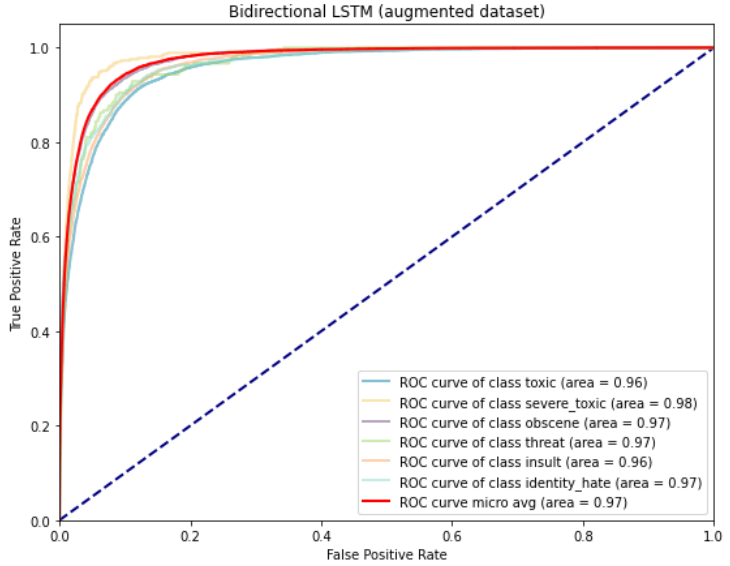
ORIGINAL DATASET

	Precision	Recall	F1-score	support
Toxic	0.61	0.75	0.67	6090
Severe_toxic	0.35	0.26	0.30	367
Obscene	0.66	0.73	0.69	3691
Threat	0.00	0.00	0.00	211
Insult	0.62	0.63	0.63	3427
Identity_hate	0.60	0.00	0.01	712
Micro avg	0.62	0.66	0.64	14498



AUGMENTED DATASET

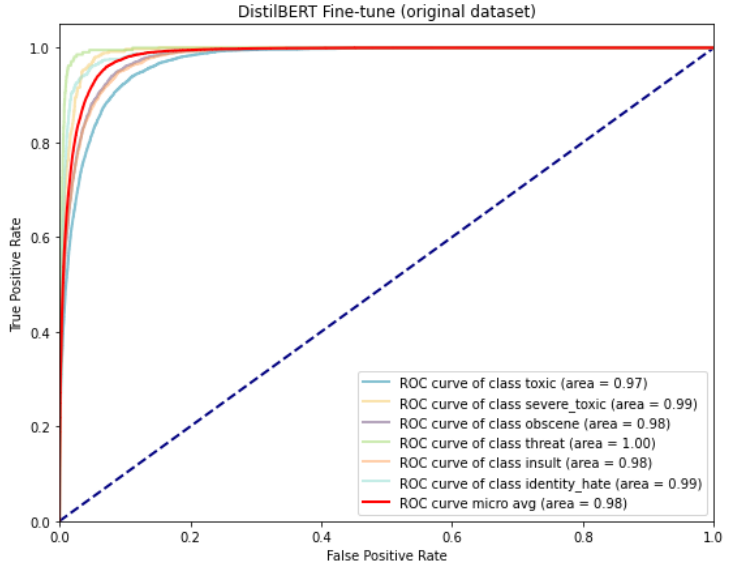
	Precision	Recall	F1-score	support
Toxic	0.52	0.85	0.65	6090
Severe_toxic	0.42	0.13	0.20	367
Obscene	0.62	0.77	0.69	3691
Threat	0.33	0.02	0.04	211
Insult	0.61	0.67	0.64	3427
Identity_hate	0.78	0.14	0.24	712
Micro avg	0.57	0.72	0.63	14498



DistilBERT Fine-tuned

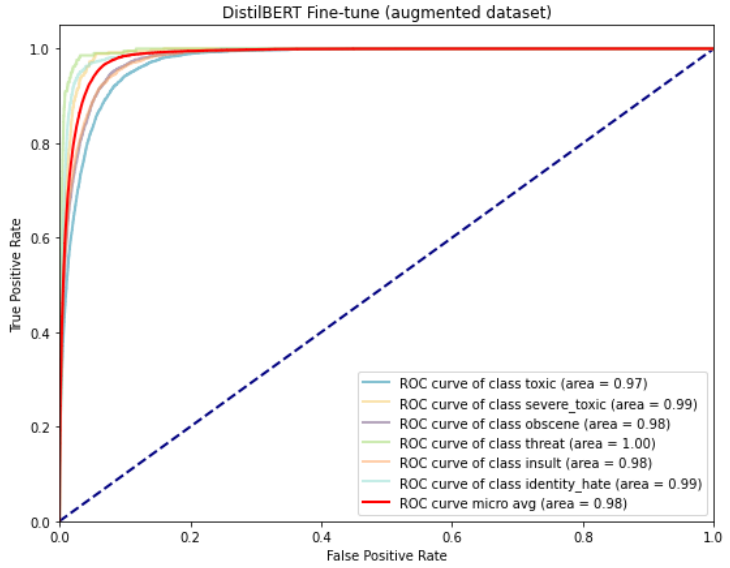
ORIGINAL DATASET

	Precision	Recall	F1-score	support
Toxic	0.50	0.92	0.65	6090
Severe_toxic	0.27	0.75	0.39	367
Obscene	0.58	0.83	0.68	3691
Threat	0.44	0.62	0.51	211
Insult	0.61	0.77	0.68	3427
Identity_hate	0.56	0.61	0.58	712
Micro avg	0.53	0.84	0.65	14498

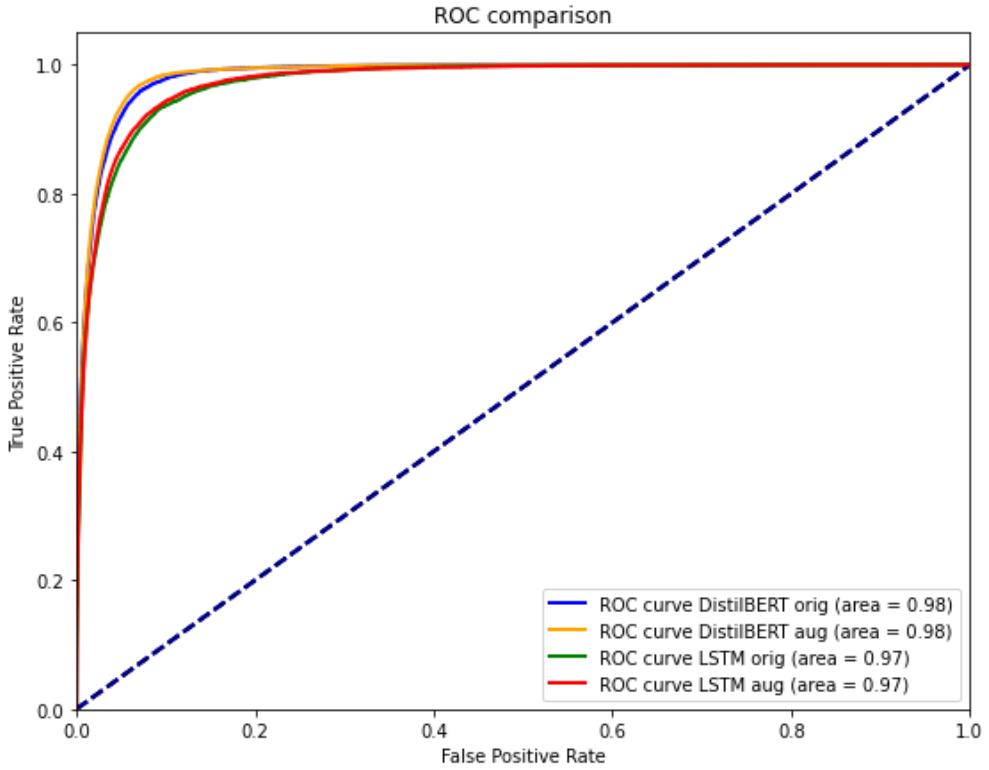


AUGMENTED DATASET

	Precision	Recall	F1-score	support
Toxic	0.50	0.94	0.65	6090
Severe_toxic	0.26	0.72	0.39	367
Obscene	0.52	0.89	0.65	3691
Threat	0.46	0.57	0.51	211
Insult	0.51	0.87	0.65	3427
Identity_hate	0.51	0.97	0.58	712
Micro avg	0.50	0.89	0.64	14498



Comparison



	Dataset originale		Dataset aumentato	
	LSTM	DistilBERT Fine tune	LSTM	DistilBERT Fine tune
AUC	0.97	0.98	0.97	0.98
F1	0.64	0.65	0.63	0.64

THANK YOU FOR YOUR ATTENTION

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