Addressing the lack of pragmatic knowledge in multimodal models

Victor Callejas Fuentes[1], C.S.Bahushruth[2]

[1] https://www.linkedin.com/in/victor-callejas-fuentes/
[2] https://www.linkedin.com/in/c-s-bahushruth-4b8449150/

Contents

Problem Statement as we understand

[1]A direct or indirect attack on people based on characteristics, including ethnicity, race, nationality, immigration status, religion, caste, sex, gender identity, sexual orientation, and disability or disease

What we need to detect:

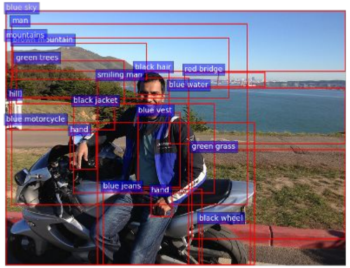
Hate sentiment + Protected Entities

(Multimodal and holistically)

[1] https://arxiv.org/pdf/2005.04790.pdf

Object detection

[1]Faster R-CNN with ResNet-101 backbone trained on Visual Genome



Object representation

Feature vector

Position encoding

Confidence

Text-like data

blue sky, smiling man,
black jacket,
blue motorcycle...

[1] Anderson et al Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering - https://arxiv.org/abs/1707.07998

Adding Internet knowledge

Object representation

white woman, white table, blue wave...

[1]Vision API Entities

Bethany Hamilton

[2]Duck Duck Go API Topic

American disabled sportspeople

Shark Attack Victims

With Post-Processing (cleaning, summarizing, translating)

[1] https://cloud.google.com/vision/docs/detecting-web
[2] https://duckduckgo.com/api

Slur auxiliary tokens based on hate speech policy

Slur	Targets
Cotton Picker	Blacks
Cow kisser	Indians
Vegetable	Disabled

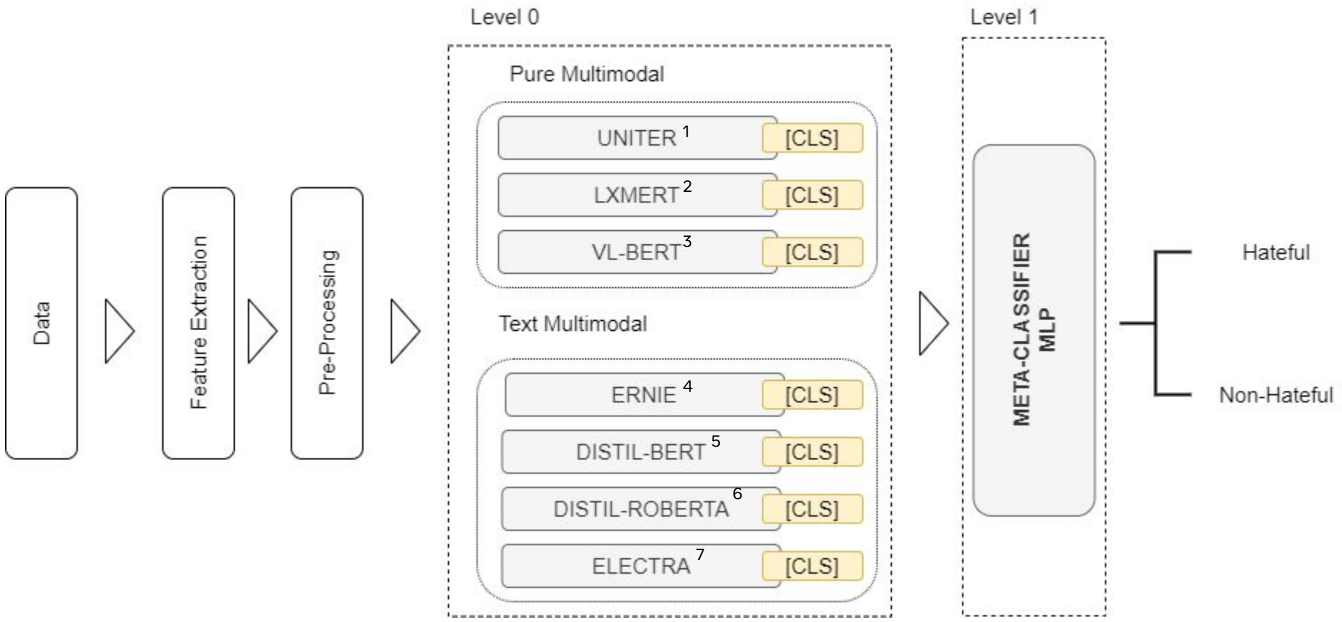
Slur database

Input Text (Meme caption + Additional text information)

Information Retrieval Techniques
(Text matching, Context matching, Input text type)

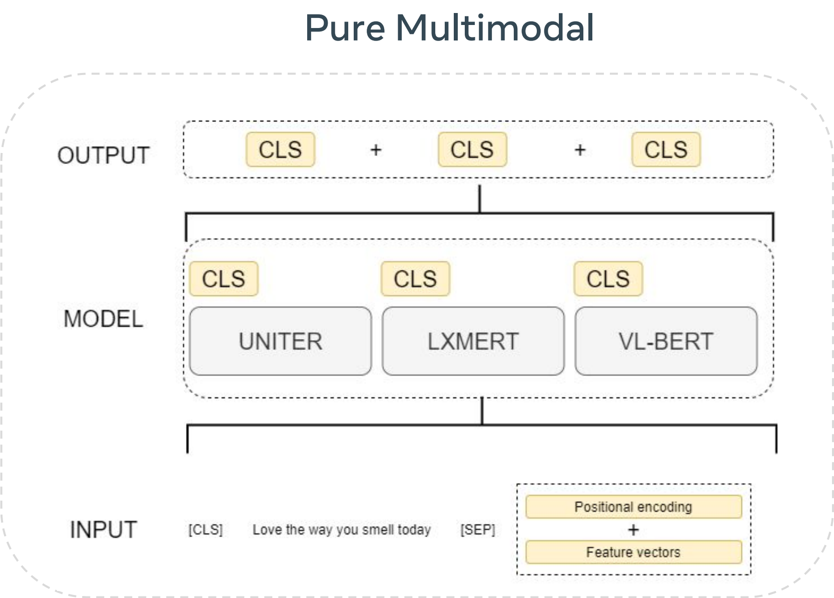
Auxiliary Tokens: [BLACKS], [INDIANS], [DISABLED] etc

Architecture

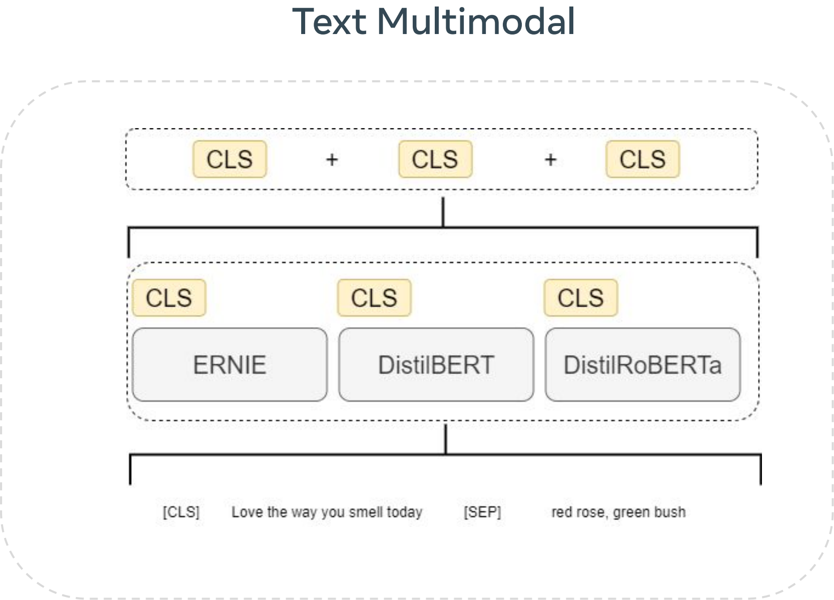


[1] https://arxiv.org/abs/1909.11740 [2] https://arxiv.org/abs/1908.07490 [3] https://arxiv.org/abs/1908.08530
[4] https://arxiv.org/abs/1907.12412 [5] https://arxiv.org/abs/1910.01108 [6] https://arxiv.org/abs/1907.11692 [7] https://arxiv.org/abs/2003.10555

Level-0



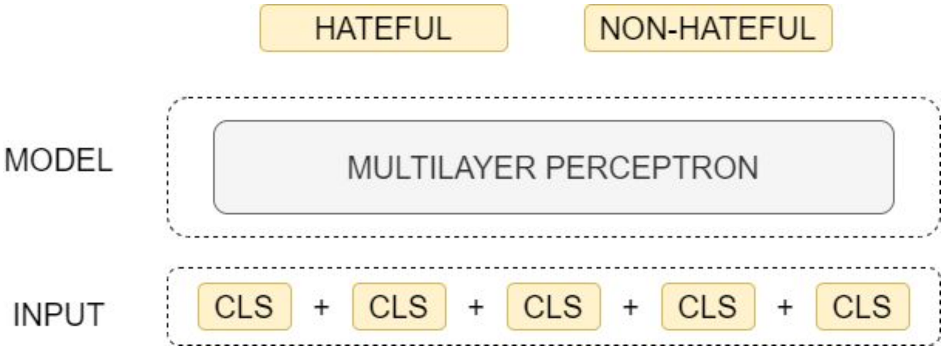
Pure Multimodal
OUTPUT
CLS + CLS + CLS
MODEL
CLS
UNITER
CLS
LXMERT
CLS
VL-BERT
INPUT
[CLS]
Love the way you smell today
[SEP]
Positional encoding
+
Feature vectors



Text Multimodal
CLS + CLS + CLS
CLS
ERNIE
CLS
DistilBERT
CLS
DistilRoBERTa
[CLS]
Love the way you smell today
[SEP]
red rose, green bush

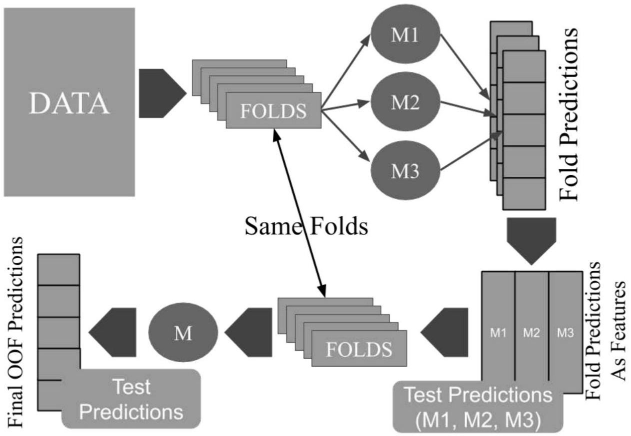
Love the way
you smell today

Level-1 Meta-Classifier



Training

Cross Validation Scheme



Techniques

Stochastic Weight Averaging (SWA)[1]

- Maintains an average of the model weights across multiple epochs
- Allow us to get a stable cross validation score

FP16[2]

- Twice faster training and bigger batches with same score

Optimizing AUC through loss function[3]

[1] PyTorch 1.6 - https://pytorch.org/blog/pytorch-1.6-now-includes-stochastic-weight-averaging/
[2] PyTorch 1.6 - https://pytorch.org/blog/accelerating-training-on-nvidia-gpus-with-pytorch-automatic-mixed-precision/
[3] https://github.com/iridiumblue/roc-star

Results & Further work

Scores

Mean 10-Fold Cross Validation

Accuracy	0.813	
AUC Roc	0.882	

Test Unseen

Accuracy	0.745	*2th position*
AUC Roc	0.788	*7nd position*

Findings

Using directly object labels

- Image and text tokens share the same representation at the cost of lost information and bias

Web entities

- Historic or internet knowledge improves score

Combining different models

- Different models have different tokenizers and pre-training methods, so each one of them can extract information that the others can not and vice versa, so the combination of them achieves best results

Further work

- [1]ERNIE-ViL, current state-of-the-art multimodal model. (adds relationships between objects reconstructing a scene graph)
- [2]Conterfactual training may help avoid bias

[1] https://arxiv.org/abs/2006.16934
[2] https://arxiv.org/abs/2006.04315