Mitigating Negative Transfer with Task Awareness for Sexism, Hate Speech, and Toxic Language Detection

Angel Felipe Magnossão de Paula, Paolo Rosso and Damiano Spina
Department of Computer Systems and Computation, Universitat Politècnica de València
School of Computing Technologies, RMIT University

BIG PICTURE

In the field of machine learning, the common strategy is to apply the Single-Task Learning approach in order to train a supervised model to solve a specific task. Training a robust model requires a lot of data and a significant amount of computational resources, making this solution unfeasible in cases where data are unavailable or expensive to gather. Therefore another solution, based on the sharing of information between tasks, has been developed: Multi-task Learning (MTL). Despite the recent developments regarding MTL, the problem of negative transfer has still to be solved. Hence, we propose a new approach to mitigate the negative transfer problem based on the task awareness concept. The proposed approach results in diminishing the negative transfer together with an improvement of performance over classic MTL solution.

PROBLEM

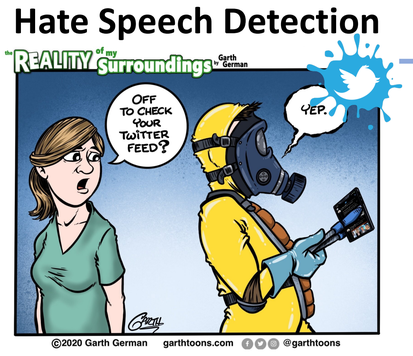
Sexism Detection

Example:
[YES] "Woman behind the wheel, be careful!"
[NO] "not all white women are like that"

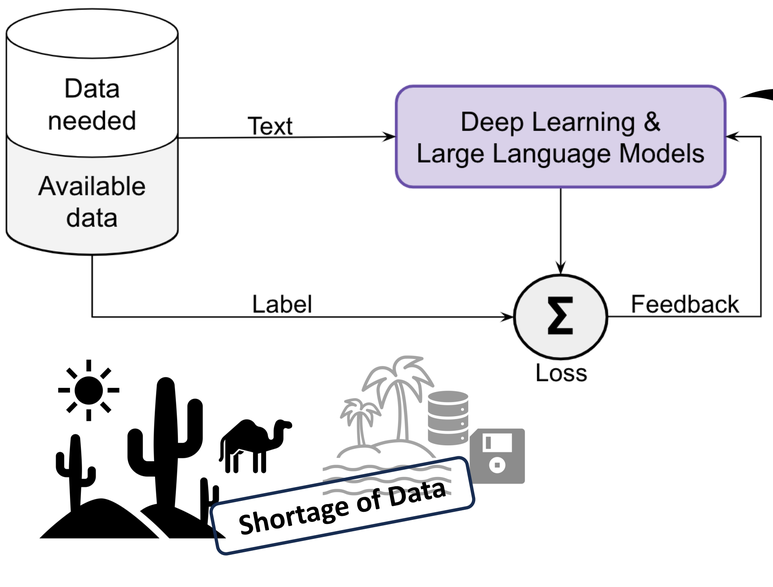
Toxic Language Detection

Example:
[YES] "To these fucking mangy animals lock them up and throw away the key"
[NO] "Against misinformation data"

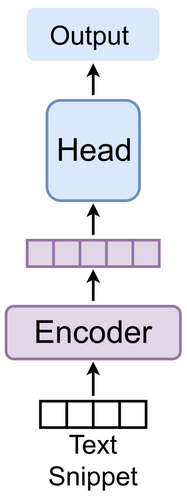
Hate Speech Detection



Example:
[YES] "Shut up you motherfucker. What did you do for politics... nothing"
[NO] "This is inhumane Karma is a bitch she get around these brainless heartless assholes! "

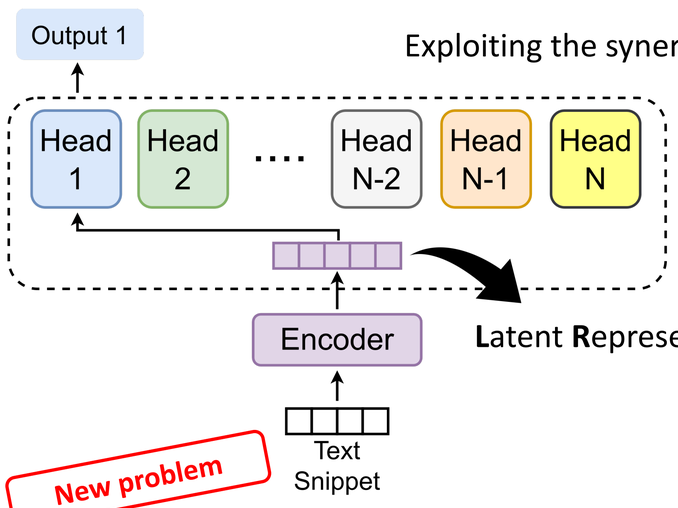


Single-Task Learning (STL)



CURRENT SOLUTION

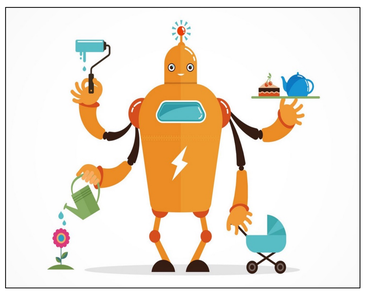
Multi-**T**ask **L**earning (**MTL**)



Exploiting the synergy between different tasks to reduce:

- **Training Data**
- Computational Resources

Latent **R**epresentation (**LT**)



Negative Transfer: occurs when noisy information is shared between tasks, resulting in a drop in performance.

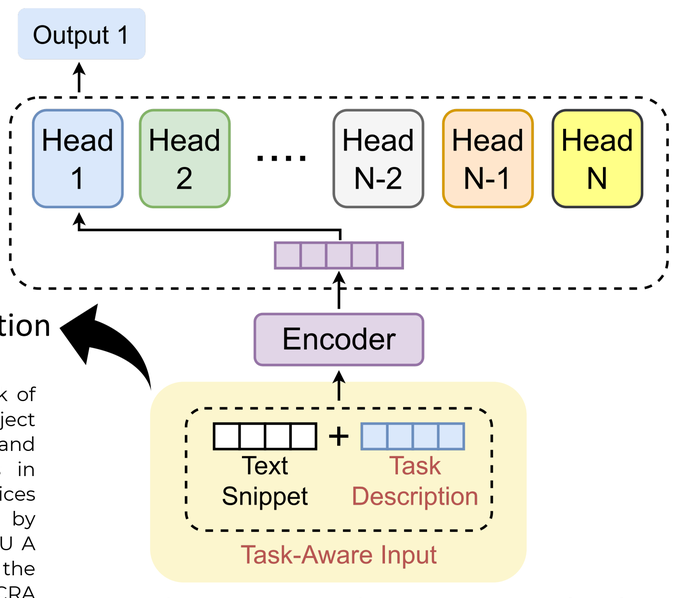
OUR SOLUTION

Task **A**wareness (**TA**)
- Model aware of task
- **LT** based on the task

Example:

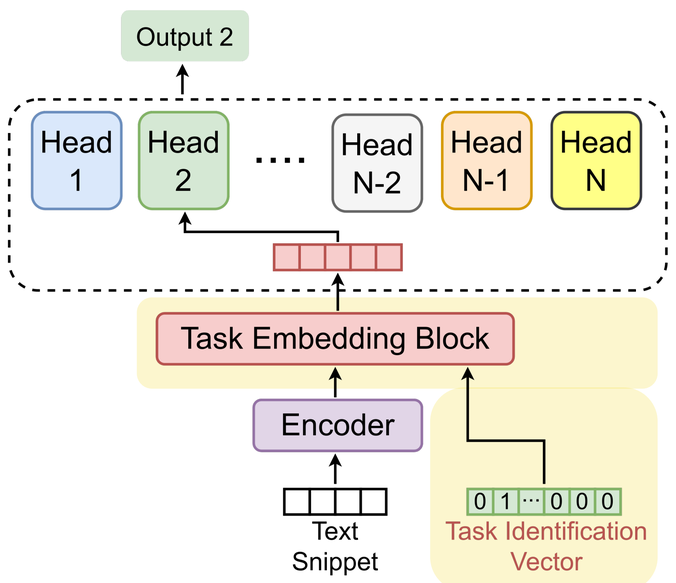
<CL> Tweet **<SEP>** Sexism Detection

Model 1: **MTL T**ask-**A**ware **I**nput (**MTL-TAI**)



Mechanism 1: **T**ask-**A**ware **I**nput (**TAI**)

Model 2: **MTL T**ask **E**mbedding (**MTL-TE**)



Example:
Sexism: 1 0 0
Toxic Language: 0 1 0
Hate Speech: 0 0 1

Mechanism 2: **T**ask **E**mbedding (**TE**)

EXPERIMENTS

- Cross-validation
- Train-test split

Data
EXIST-2021
DETOXIS-2021
HatEval-2019

State-Of-The-Art Models
IA-UPV
SINAI
Atalaya

TABLE IV
RESULTS OF THE CROSS-VALIDATION EXPERIMENT WITH 95% CONFIDENCE INTERVALS

Model	Task Heads	EXIST-2021	DETOXIS-2021	HatEval-2019
		Accuracy	F1-score	F1-macro
STL	Sexism	0.789 ± 0.011	–	–
	Toxic-language	–	0.640 ± 0.014	–
	Hate-speech	–	–	0.846 ± 0.009
MTL	Sexism + Toxic-language	0.788 ± 0.011	0.628 ± 0.014	–
	Sexism + Hate-speech	0.791 ± 0.011	–	0.843 ± 0.009
	Toxic-language + Hate-speech	–	0.632 ± 0.014	0.841 ± 0.009
	Toxic-language + Hate-speech + Sexism	0.799 ± 0.010	0.634 ± 0.014	0.842 ± 0.009
MTL-TAI	Sexism + Toxic-language	0.799 ± 0.010	0.649 ± 0.014	–
	Sexism + Hate-speech	0.805 ± 0.010	–	0.984 ± 0.003
	Toxic-language + Hate-speech	–	0.649 ± 0.014	0.988 ± 0.003
	Toxic-language + Hate-speech + Sexism	0.800 ± 0.010	0.650 ± 0.014	0.980 ± 0.003
MTL-TE	Sexism + Toxic-language	0.797 ± 0.011	0.653 ± 0.014	–
	Sexism + Hate-speech	**0.806** ± 0.010	–	**0.992** ± 0.002
	Toxic-language + Hate-speech	–	0.653 ± 0.014	0.980 ± 0.003
	Toxic-language + Hate-speech + Sexism	0.801 ± 0.010	**0.659** ± 0.014	0.988 ± 0.003

TABLE V
RESULTS OF THE TRAINING-TEST EXPERIMENT WITH 95% CONFIDENCE INTERVALS

Model	Task Heads	EXIST-2021	DETOXIS-2021	HatEval-2019
		Accuracy	F1-score	F1-macro
AI-UPV [23]	–	0.790 ± 0.018	–	–
SINAI [30]	–	–	**0.646** ± 0.031	–
Atalaya [28]	–	–	–	0.730 ± 0.022
STL	Sexism	0.790 ± 0.017	–	–
	Toxic-language	–	0.620 ± 0.032	–
	Hate-speech	–	–	0.764 ± 0.021
MTL	Sexism + Toxic-language	0.776 ± 0.018	0.639 ± 0.032	–
	Sexism + Hate-speech	0.785 ± 0.017	–	0.778 ± 0.020
	Toxic-language + Hate-speech	–	0.593 ± 0.032	0.777 ± 0.020
	Toxic-language + Hate-speech + Sexism	0.775 ± 0.018	0.629 ± 0.032	0.773 ± 0.021
MTL-TAI	Sexism + Toxic-language	0.797 ± 0.017	0.633 ± 0.032	-
	Sexism + Hate-speech	**0.809** ± 0.017	-	0.789 ± 0.020
	Toxic-language + Hate-speech	–	0.628 ± 0.032	**0.790** ± 0.020
	Toxic-language + Hate-speech + Sexism	0.792 ± 0.017	0.629 ± 0.032	0.782 ± 0.020
MTL-TE	Sexism + Toxic-language	0.804 ± 0.017	0.626 ± 0.032	–
	Sexism + Hate-speech	0.804 ± 0.017	–	0.786 ± 0.020
	Toxic-language + Hate-speech	–	0.623 ± 0.032	0.786 ± 0.020
	Toxic-language + Hate-speech + Sexism	0.802 ± 0.017	0.633 ± 0.032	0.789 ± 0.020

CONTRIBUTION

(1) We propose the use of the Task Awareness (TA) concept to mitigate the negative transfer problem during MTL training. **(2)** Design of the Task-Aware Input (TAI) mechanism to grant the MTL models with task awareness ability to mitigate negative transfer and even improve results compared with traditional MTL models. **(3)** Design of the Task Embedding (TE) mechanism to give MTL models task recognition capability to diminish negative transfer and improve the results over classic MTL solutions. **(4)** Creation and validation of two unified architectures to detect Sexism, Hate Speech, and Toxic Language in text comments. **(5)** Our proposed method outperforms the State-Of-The-Art on two public benchmarks for Sexism and Hate Speech detection: (i) EXIST-2021 and (ii) HatEval-2019 datasets.

The work of Paolo Rosso was in the framework of the FairTransNLP-Stereotypes research project (PID2021-124361OB-C31) on Fairness and Transparency for equitable NLP applications in social media: Identifying stereotypes and prejudices and developing equitable systems, funded by MCIN/AEI/10.13039/501100011033 and by ERDF, EU A way of making Europe. Damiano Spina is the recipient of an Australian Research Council DECRA Research Fellowship (DE200100064).

Pattern Recognition and Human Language Technology Research Center

Paper

Code Repo