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 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

Toxic Language Detection

Hate Speech Detection


data

available

Deep Learning & Large Language Models


data

available

Exploiting the synergy between different tasks to reduce:
- Training Data
- Computational Resources

CURRENT SOLUTION

Multi-Task Learning (MTL)

Output 1

Head 1 Head 2 Head N-2 Head N-1 Head N

Head 1 Head 2 Head N-2 Head N-1 Head N

Encoder

Latent Representation (LT)

New problem

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

OUR SOLUTION

Task Awareness (TA)
- Model aware of task LT based on the task

Example:
<CL>Tweet <SEP> Sexism Detection

The work of Paolo Rosso was in the framework of the FairHandNLP-Discrimination research project [102012/IA34310/06-CT]. The FairHandNLP-Discrimination framework is part of the European Union’s Horizon 2020 (H2020) program. The European Commission has supported the project with Grant Agreement No. 731923. This work is part of the project HATEFIN (866082) in the frame of Open Call Digital Europe 2020 (DE/2020/00004).

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.

EXPERIMENTS

- Cross-validation
- Train-test split

Model: MTL Task Embedding (MTL-TE)

Model 1: MTL Task-Aware Input (MTL-TA)

Model 2: MTL Task Embedding (MTL-TE)

Example:
- Sexism: [NO] [YES]
- Toxic Language: [NO] [YES]
- Hate Speech: [NO] [YES]

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.

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.