

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, València, Spain 46022

†School of Computing Technologies, RMIT University, Melbourne, Australia 3000

Email: {adepau@doctor, proso@dsic}.upv.es, damiano.spina@rmit.edu.au

Abstract—This paper proposes a novelty approach to mitigate the negative transfer problem. 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. Negative transfer is a phenomenon that occurs when noisy information is shared between tasks, resulting in a drop in performance. This paper proposes 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. Moreover, the proposed approach has been implemented in two unified architectures to detect Sexism, Hate Speech, and Toxic Language in text comments. The proposed architectures set a new state-of-the-art both in EXIST-2021 and HatEval-2019 benchmarks.

Index Terms—Multi-task Learning, Negative Transfer, Natural Language Processing, Deep Learning

I. INTRODUCTION

Machine Learning has numerous applications in fields as diverse as Natural Language Processing (NLP) (e.g., named entity recognition and hate speech detection) [19], [26] or Computer Vision (CV) (e.g., object detection and object classification) [41]. Generally, a single model or an ensemble of models is trained to address all the desired tasks. These models are then fine-tuned and tweaked on the chosen task until they specialize, and their performance no longer increases. Despite producing satisfactory results, a Single-Task Learning (STL) strategy ignores knowledge that may be gathered from datasets of related tasks, allowing our model to generalize better on our original task. Furthermore, in many cases, more than the available data is needed to train a model robustly. Therefore, several strategies to transfer knowledge from one task to another have been developed [18].

Multi-Task Learning (MTL) [33], [49] is a new area of study that aims at exploiting the synergy between different tasks to reduce the amount of data or computational resources required for these activities. This approach aims at improving generalization by learning multiple tasks simultaneously. The

soft [43], [47] or *hard parameter-sharing* [13], [14] strategies are two of the most commonly used methods for MTL employing neural networks. In soft parameter-sharing, task-specific networks are implemented, while feature-sharing methods handle cross-task communication to encourage the parameters to be similar. Since the size of the multi-task network grows linearly with respect to the number of tasks, an issue with soft parameter-sharing systems is given by scalability. In hard parameter-sharing, the parameter set is split into shared and task-specific operations. It is commonly implemented with a shared encoder and numerous task-specific decoding heads [49]. One of the benefits of this method is the minimization of overfitting [33].

Multilinear relationship networks [20] enhanced this architecture by imposing tensor normal priors on the fully connected layers' parameter set. The branching sites in the network are set ad-hoc in these works, which can result in inefficient job groupings. To address this limitation, tree-based approaches [22], [38] have been proposed. Despite the improvement introduced by those works, jointly learning multiple tasks might lead to *negative transfer* [39], [46] if noisy information is shared among the tasks. During training, the hard parameter-sharing encoder learns to construct a generic representation that focuses on extracting specific features from the input data. Nevertheless, a subset of these features may provide critical information for a given decoder head but introduces noise to another decoder to solve its respective task. Hence, negative transfer refers to situations in which the transfer of information results in a decrease in the overall model performance.

In this work, we propose a new approach to overcome the negative transfer problem based on the concept of Task Awareness (TA). This approach enables the MTL model to take advantage of the information regarding the addressed task. The overarching goal is for the model to handle its internal weight for its own task prioritization. Unlike the State-Of-The-Art (SOTA) approaches (later presented in Section II), the proposed solution does not require a recursive structure, saving time and resources. Moreover, we designed two mechanisms based on the TA approach and implemented them in the creation of two Multi-Task Learning TA (MTL-TA) architectures

to address SOTA challenges: Sexism, Hate Speech, and Toxic Language detection. The source code is publicly available.¹

The main contributions of our work are as follows:

- We propose the use of the TA concept to mitigate the negative transfer problem during MTL training.
- 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.
- 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.
- Creation and validation of two unified architectures to detect Sexism, Hate Speech, and Toxic Language in text comments.
- Our proposed method outperforms the SOTA on two public benchmarks for Sexism and Hate Speech detection: (i) EXIST-2021 and (ii) HatEval-2019 datasets.

The rest of the paper is structured as follows. Section II presents the related works of transfer learning and MTL. Section III describes the details of our proposed method. Section IV illustrates the experiment setup. Section V discusses and evaluates the experimental results. Section VI presents the limitation of our approach. Finally, conclusions and future work are drawn in Section VII.

II. RELATED WORK

Transfer learning is a widespread technique in machine learning based on the idea that a model created for one task can be improved by transferring information from another task [27], [44]. Training a model from scratch requires a large quantity of data and resources, but there are some circumstances where gathering training data is prohibitively expensive or impossible. As a result, there is the need to construct high-performance learners trained with more easily accessible data from different tasks. Transfer learning techniques allow us to improve the results of target tasks through information extracted from related tasks. These techniques have been effectively used for a variety of machine learning applications, including NLP [31], [34], [42], [43] and CV [11], [18]. The MTL framework [33], [49], which seeks to learn many tasks at once even when they are distinct, is a closely related learning technique to transfer learning. This approach works well and can take advantage of sharing information among tasks. Still, if the tasks are not sufficiently related, it can lead to negative transfer. The problem of negative transfer consists of performance degradation caused by noisy information being shared between tasks.

To solve this issue, several approaches for balancing learning between different tasks have been proposed based on a re-weighting of the losses (for instance, via Homoscedastic uncertainty [17], Gradient normalization [9] and Adversarial training [36]) or task prioritization [15], [35], [52]. Further

recent approaches [48], [50], [51] make use of the initial predictions obtained through multi-task networks to improve, once or repeatedly, each task output, overcoming a characteristic of the previously mentioned methods that computed all the task outputs for a given input at once. Those last approaches culminate to be very time-consuming and require a lot of computational resources due to their recursive nature.

This paper proposes two unified architectures to detect Sexism, Hate Speech, and Toxic Language in text comments. Abburi, Parikh, Chhaya, *et al.* [1] represents the first semi-supervised multi-task approach for sexism classification. The authors addressed three tasks based on labels achieved through unsupervised learning or weak labeling. The neural multi-task architecture they proposed allows shared learning across multiple tasks via common weight and a combined loss function. The method outperforms several SOTA baselines. Wu, Fei, and Ji [47] proposed an MTL innovative approach to solve Aggressive Language Detection (ALD) together with text normalization. The authors propose a shared encoder to learn the common features between the two tasks and a single encoder dedicated to learning the task-relevant features. The proposed model achieved a significant improvement in performance concerning the ALD task.

Those last approaches inspired the mechanism we propose in this paper. The main commonality is to have additional mechanisms added to the MTL models to improve the representation sent to the task heads. The main difference with respect to the TA approach we propose is that we enrich the model with the ability to discover by itself which task it will perform. It allows the MTL-TA models to create a suitable representation for each task head. In addition, the MTL-TA models do not need to learn an auxiliary task, resulting in more efficiency. In fact, the TA approach allows the MTL models, at each step, to try to optimize over the task at hand. The key idea is to learn a task-relevant latent representation of the data, efficiently solving many NLP tasks [16], [43]. The resulting mechanisms are proposed in the following section.

III. PROPOSED APPROACH

This section describes the details of the MTL-TA models. We first introduce the notion of TA and explain how it can be beneficial in diminishing the negative transfer [39], [46] for multi-task joint training [33]. Secondly, two different TA mechanisms are proposed in order to incorporate the task self-awareness capability into MTL models.

The mainstream approach to supervised multi-task is the hard parameter-sharing method [49]. The model is composed of an encoder and N decoders or task heads, where N corresponds to the number of tasks the model is simultaneously trained [45]. During execution, the encoder receives input and creates a task-agnostic latent representation that is sent to a certain task head, which is in charge of producing the final prediction.

The lack of a closer relationship between the latent representation generated by the encoder and the tasks degrades the overall MTL model performance [39]. For the same input,

¹<https://github.com/AngelFelipeMP/Mitigating-Negative-Transfer-with-TA>

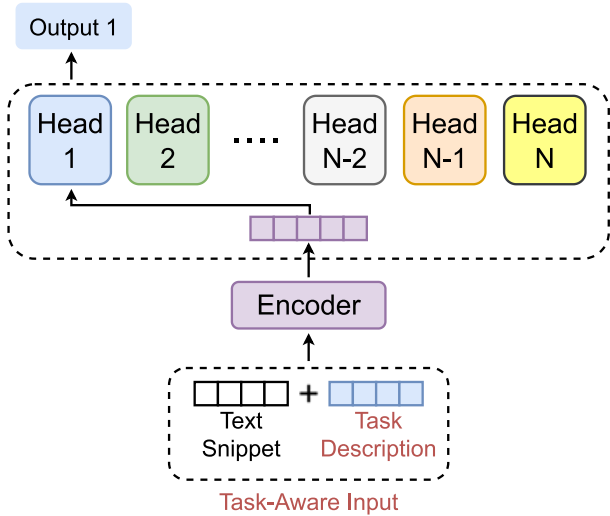


Fig. 1. Multi-Task Learning (MTL) model including Task-Aware Input (TAI) mechanism (MTL-TAI).

the optimal latent representation for task heads are likely to be different [14]. Furthermore, the encoder representation can get prone to more demanding tasks or with a larger data volume during training [33]. These model performance deteriorations are the reflex of the negative transfer phenomenon [39], [46], where a task head receives an inaccurate input representation to solve its respective task.

We propose two TA mechanisms to mitigate negative transfer when solving multiple NLP tasks by applying the MTL approach [49]. These mechanisms tailor, depending on the specific task that is addressed, the input representation that is sent to its respective head. In addition, our proposed MTL model still takes advantage of the generalization improvements the multi-task joint training provided. Hence, the encoder and other MTL model parts located before the task heads are updated during training for every task. It should be noted that all our proposed MTL models belong to the MTL-TA class, and they follow the conventional MTL paradigm. Therefore, only the specific task head attached to the input data is considered during the task parameter updating.

A. Task-Aware Input

The first mechanism we designed to introduce task awareness into MTL models is Task-Aware Input (TAI). To compel the encoder to generate a suitable representation for each task head, we proposed to modify the MTL conventional input for NLP tasks.

The TAI includes a Text Snippet (TS) plus a Task Description (TD), as shown in Fig. 1. The TS is a text chunk whose length varies according to the task. It is usually the integral input for the MTL encoders. The TD is a piece of text describing what a specific head is dealing with, such as ‘Sexism Detection’ and ‘Hate Speech Detection’. The new modified input provides context for the encoder to generate a

task-centered representation. The MTL model endowed with the TAI mechanism is referred as MTL Task-Aware Input (MTL-TAI).

B. Task Embedding

The second mechanism we designed to convey MTL models with the TA capability was named Task Embedding (TE). We proposed to insert an additional building block between the encoder and the task heads, which we call Task Embedding Block (TEB), as displayed in Fig. 2. It receives two inputs: (i) the Task Identification Vector (TIV) and (ii) the latent encoder representation. The TIV is a unidimensional one-hot vector whose size is proportional to the number of task heads. Each TIV location is related to one of the task heads.

The TEB is composed of Learning Units (LU) that encompass a linear layer followed by a ReLU layer. The number of LUs is a hyperparameter that depends on the task and data, among other factors. The TEB objective is to generate a suitable representation for the task the MTL model is solving at a specific time. Hence, depending on the task, the TEB will retrieve a different output for the same exact encoder representation. It relies on the TIV to indicate for which task the TEB will generate a representation. The TIV has the number one in the location that corresponds to the task the model is about to solve. The remainder of the vector is populated with zeros, as Fig. 2 reflects. The MTL model equipped with the TE mechanism is referred as MTL Task Embedding (MTL-TE).

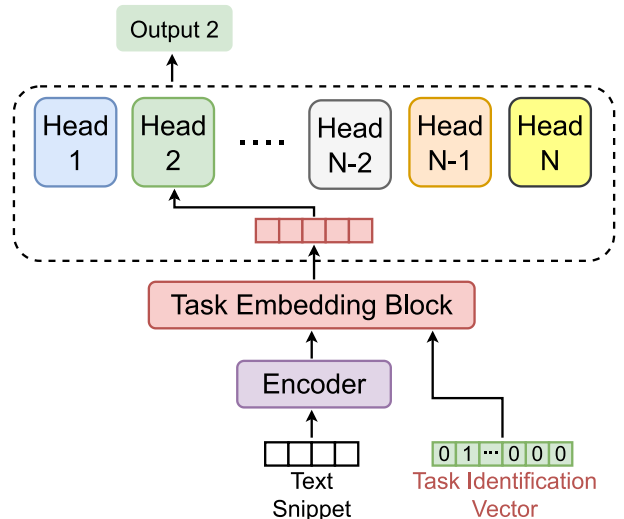


Fig. 2. Multi-Task Learning (MTL) model including Task Embedding (TE) mechanism (MTL-TE).

IV. EXPERIMENTAL SETUP

This section first describes the tasks and the datasets used to evaluate our approach. It then presents the implementation details and models for reference. Finally, we share the settings for the experiments.

TABLE I
EXIST-2021 DATA DISTRIBUTION

	Training		Test			
	Spanish	English	Spanish		English	
	Twitter	Twitter	Twitter	Gab	Twitter	Gab
Sexist	1,741	1,636	858	265	858	300
Not-Sexist	1,800	1,800	812	225	858	192

TABLE II
DETOXIS-2021 DATA DISTRIBUTION

	Training	Test
Toxic	1,147	239
Not-Toxic	2,316	652

TABLE III
HATEVAL-2019 DATA DISTRIBUTION

	Training		Development		Test	
	Spanish	English	Spanish	English	Spanish	English
Hate	1,741	1,636	1,741	1,636	858	300
Not-Hate	1,800	1,800	1,800	1,800	812	192

A. Data

Our approach for selecting the datasets for Sexism, Hate Speech, and Toxic Language detection was based on two requirements: (i) being publicly available; (ii) having been used to evaluate a high number of ML models. We use three datasets – EXIST-2021 [32], DETOXIS-2021 [37], and HateEval-2019 [2] – which we describe below.

EXIST-2021 [32]: The dataset was created for the sExism Identification in Social neTworks (EXIST) shared task at Iberian Languages Evaluation Forum (IberLEF) 2021. The dataset consists of 11345 annotated social media text posts in English and Spanish from Twitter and Gab.com (Gab), an uncensored social media platform. The dataset development was supervised and monitored by experts in gender issues. The EXIST was the first challenge on Sexism detection in social media, whose objective was to identify sexism in a wide sense, from explicit misogyny to more implicit sexist behaviors. The challenge received 70 official runs for the Sexism identification task. It is a binary classification where the samples belong to the Sexist class or the Not-Sexist class. The official evaluation metric was accuracy, and data was split into training and test sets. Table I shows the data distribution.

DETOXIS-2021 [37]: The dataset was collected for the DEtection of TOxicity in comments In Spanish (DETOXIS) shared task at IberLEF 2021. The objective of the shared task was toxic language detection in comments to various online

news articles regarding immigration. The proposed annotation methodology focused on diminishing the subjectivity of toxicity labeling considering contextual information (e.g., linguistic features and conversational threads). The team that worked on the data annotation was composed of trained annotators and expert linguists. The dataset consists of 4354 text comments from Twitter in Spanish and provides labels for Toxic Language detection. The task is characterized as a binary classification where the samples are divided between the Toxic and Not-Toxic classes. More than 30 teams evaluated their machine learning model in the collected dataset in the participation for DETOXIS shared task. The official data evaluation metric was F1-score in the Toxic class, and the data were divided into training and test sets. Table II shows the data distribution.

HatEval-2019 [2]: The dataset was constructed for the Detection of Hate Speech Against Immigrants and Women in Twitter (HatEval) shared task, which was part of the International Workshop on Semantic Evaluation (SemEval) 2019. The dataset comprises 19600 tweets published in English and Spanish and supplies labels for Hate Speech detection. The data collection methodology employed different gathering strategies: (i) monitoring likely victims of hate accounts; (ii) downloading the records of recognized haters; (iii) filtering Twitter streams with keywords. The annotation was performed by experts and crowdsourced contributors tested for reliable annotation. The task was defined as a binary classification where the samples are associated with the Hateful class or the Not-Hateful class. The data is composed of training, development, and test sets, and the official evaluation metric was the F1-macro, which is the unweighted mean of the F1-score calculated for the two classes. HatEval was one of the most popular shared tasks in SemEval 2019, with more than 100 submitted runs for Hate Speech detection. We can see the dataset distribution in Table III.

B. Implementation Details

The encoder was constructed using a popular BERT [10] version for Spanish called BETO [7], followed by max and mean pooling calculation over its output. BETO has 12 self-attention layers, each with 12 attention-heads, using 768 as the hidden size with around 110 million parameters. BETO receives a text sequence and returns a hidden representation dimensionally equivalent to its hidden size for each token that belongs to the sequence. The latent encoder representation is created by a concatenation of max pooling and mean pooling calculation on the entire 768-dimensional sequence of tokens returned by BETO. Regarding the TE approach, the TEB preserves the same dimension of the latent encoder representation.

The task heads are linear classifiers whose input dimension corresponds to the latent encoder representation, and the output depends on the task. In the case of binary classification, the linear classifier returns two values, and the higher value corresponds to the predicted class. Furthermore, the TDs for

the EXIST-2021 [32], DETOXIS-2021 [37], and HatEval-2019 [2] datasets are, respectively, the following pieces of text: ‘Sexism detection’, ‘Toxic Language detection’, and ‘Hate Speech detection’.

The models were trained using the optimization algorithm AdamW [21] with a linear decay learning rate schedule and a learning rate varying from 5e-6 to 1e-4. In the learning process, we trained our model for 15 epochs with a dropout of 0.3 and batch size of 64. Additionally, we experimented with 1 up to 3 LUs. Similar to the early stopping strategy [8], we adopted the model with the best performance within the epochs based on the task’s official metric.

C. Comparison Models

We compare our approach with two types of models: (i) Baselines and (ii) SOTA models. The baselines are the two models that we implemented:

- **MTL** is the classic MTL model. It is constructed with the same architecture as the MTL-TA model (described in Section III), but it does not include the TAI mechanism. Therefore, the MTL model receives only the TS as input.
- **STL** is the classic STL model. It has the same architecture as the MTL model, yet it encompasses only one task head. Hence, to compare this model type with the MTL models, it is necessary to train one model for each one of the addressed tasks.

The SOTA are the models which currently achieved the best performance on the datasets considered in our experiments:

- **AI-UPV** [23]: is a deep learning architecture based on the combination of different Transformers models [40]. It takes advantage of ensemble methods and, during training, applies data augmentation mechanisms. It is the SOTA for EXIST-2021 [32].
- **SINAI** [30]: is a BERT base model [10] trained using the MTL hard parameter-sharing method. In spite of addressing five tasks and six datasets, the model was focused on Toxic Language detection, while the other tasks were used as auxiliary tasks. It is the SOTA for DETOXIS-2021 [37].
- **Atalaya** [28]: is a model based on Support Vector Machines [6]. It was trained on several representations computed from FastText [5] sentiment-oriented word vectors, such as tweet embeddings [24], bag-of-characters [5], and bag-of-words [4]. It is the SOTA for HatEval-2019 [2].

D. Experimental Settings

We conducted two experiments to evaluate our TA approach for mitigating negative transfer [39], [46], as described below.

Cross-Validation Experiment: To assess whether the TAI and TE mechanisms were capable of reducing the negative transfer during MTL training, we performed a cross-validation experiment. Therefore, for each one of the datasets described in Subsection IV-A, we aggregate the different sets that compose the dataset in a unique set. Then, we run 5-fold cross-validation on the STL, MTL, MTL-TAI, and MTL-TE models.

Official Training-Test Split: In order to compare our approach to the SOTA models [23], [28], [30] in the utilized datasets, we carried out an experiment using the official training-test split of the respective datasets. We trained our models with the training set or a combination of the training and development sets when the last was available. After that, we evaluated the models in the test partitions.

In both experiments, we use only the data samples in the Spanish language and evaluate the models employing the dataset’s respective official metrics (described in Section IV-A). We explored versions that combined two and three tasks for the MTL models. Furthermore, models whose results were the highest regarding the evaluation metrics were selected. Finally, we applied the t-test to calculate the 95% confidence interval for the experiments results.

V. RESULTS AND ANALYSIS

This section presents the experiment’s results and the comparison among the evaluated models described in Section IV.

A. Cross-Validation Experiment

Table IV shows the cross-validation results. It is organized into three parts in the following order: model type, model’s task heads, and model’s performance. Regarding the Baseline models (described in Section IV-C), results show that the MTL training approach suffered negative transfer on nearly all occasions. The MTL model showed improvement over the STL model only for the Sexism detection task when the model was trained for Sexism and Hate Speech detection and when it was trained on the three tasks. Apart from that, the STL model achieved superior performance in the rest of the explored combinations. It probably happened because the negative transfer restrained the learning process of the MTL model on all the other occasions.

According to our results, the TA mechanisms worked well to diminish negative transfer. The MTL-TAI model equipped with the TA mechanism and the MTL-TE model equipped with the TE mechanism on all occasions achieved superior performance than the classic MTL model, as shown in Table IV. The MTL-TAI and MTL-TE models also overcame results obtained by the STL model for the three evaluated tasks. In general, the MTL-TE model performs better than the MTL-TAI model.

B. Official Training-Test Split

Table V, following the same organization as Table IV, presents the experiment carried out on the three datasets using their respective official training-test split. We see in Table V that the MTL training was not beneficial for the classic MTL model when addressing the sexism detection task. The model achieved lower accuracy compared with the STL model. We believe it was again due to the negative transfer phenomenon. Nevertheless, because of the TA mechanisms, the MTL-TA and MTL-TE models mitigated the negative transfer presented in the classic MTL training, achieving higher accuracy than the STL model and the EXIST-2021 SOTA (AI-UPV [23]).

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

The MTL training improves the result for Toxic Language detection over the STL baseline for the training-test experiment. In general, the MTL, MTL-TAI, and MTL-TE models achieved similar results, meaning there were low negative transfer levels for this task during the formal MTL training.

We see in Table V that for the training and test experiment, the MTL training improved the result of Hate Speech detection. The MTL model obtained a higher F1-macro than the HatEval-2019 SOTA (Atalaya [28]) and the STL Baseline. The MTL models with the TA mechanisms improved the results even more. They mitigate the negative transfer in the

traditional MTL training, and both models achieved superior F1-macro than the conventional MTL model.

C. Overall Analysis

Analyzing Tables IV and V, we see evidence that the STL model was a competitive baseline to compare our TA approach. Therefore, the STL models achieved close or better results than the SOTA models for the training-test experiment. The STL achieved the same results as the EXIST-2021 SOTA (AI-UPV [23]) and comparable results to the DETOXIS-2021

SOTA (SINAI [30]). Furthermore, the STL obtained better results than the HatEval-2019 SOTA (Atalaya [28]).

Summarizing the results of the two experiments, the MTL-TA models (MTL-TAI & MTL-TEB) outperformed both the STL and the classic MTL models. It shows that our proposed TA approach could mitigate the negative transfer presented in the conventional MTL training.

VI. LIMITATIONS

In this section, we mention the main limitations of our MTL-TA models. First, the two models depending on a powerful encoder to achieve good performance. It could be a problem for low-resource computation systems that cannot afford to use deep learning architectures such as Transformers [40] for the encoder. Secondly, dealing with a higher number of tasks means having more task heads – increasing the number of model parameters. Therefore, MTL-TA models will require more computational power to be fine-tuned. Finally, we wonder if the MTL-TA models have their ability to adapt to unseen tasks (e.g., few-shot learning and instruction-based prompts) reduced due to the fine-tuning process utilizing information about the tasks.

VII. CONCLUSION AND FUTURE WORK

We proposed the TA strategy to address the negative transfer [39] problem during MTL training. The proposed method has been translated into two mechanisms: TAI and TE. The TAI mechanism is the inclusion of the TD information to enrich the input of the MTL model encoder. The TE mechanism is the introduction of the TEB, an extra component that receives the representation generated by the encoder plus a TIV representation. The TD and the TIV provide information regarding the task the MTL model will perform at that precise moment. The objective of the TAI and TE is to enable the MTL model to construct task-dependent representations for the task heads to diminish negative transfer during MTL training and improve the MTL model performance. We proposed two MTL models, the MTL-TAI equipped with the TAI mechanisms and the MTL-TE that includes the TE mechanism.

Our two experiments show that the TA capability reduces negative transfer during traditional MTL training and improves performance over standard MTL solutions. We achieved competitive results compared with SOTA for the two proposed MTL-TA models for the addressed tasks: Sexism, Hate Speech, and Toxic Language detection. In particular, the proposed models set a new SOTA on two public benchmarks: (i) EXIST-2021 [32] and (ii) HatEval-2019 [2] datasets, demonstrating a general performance improvement of the proposed approach with respect to both the STL and classic MTL model. The TA mechanisms proved to be a valid approach to mitigate the negative transfer [46] problem in the MTL training.

This research demonstrated how an MTL approach equipped with TA mechanism leads to performance improvement in several NLP tasks. This approach has been demonstrated to be feasible in cases where we have a scarcity of labeled

data. In future studies, it would be interesting to deepen the analyses to find out how many labeled samples or volumes of information it is worth applying MTL rather than using STL. Further analyses regarding the enrichment of the MTL model input with low-level task supervision are worth it. In this scenario, the decoder receives all or a subgroup of the encoder’s hidden representations instead of just the last one. It would be interesting to analyze the impact of different encoder representations in an MTL model. We also plan to apply MTL with TA to other scenarios, such as sexism identification under the learning with disagreement regime [29], where it is necessary to learn from all the labels provided by the annotators rather than the aggregated gold label. This new paradigm is gaining importance in NLP, especially for tasks where often there is not only one correct label. Finally, we would like to research unsupervised techniques to improve the suggested models and tackle the same problems (detecting Hate Speech, Toxic Language, and Sexism). For instance, Latent Dirichlet Allocation [3], Self-Organizing Maps [25], and K-Means Clustering [12] could be considered.

ACKNOWLEDGMENTS

Angel Felipe Magnossão de Paula has received a mobility grant for doctoral students by the Universitat Politècnica de València. 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).

REFERENCES

- [1] H. Abburi, P. Parikh, N. Chhaya, and V. Varma, “Semi-supervised Multi-task Learning for Multi-label Fine-grained Sexism Classification,” in *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 5810–5820.
- [2] V. Basile, C. Bosco, E. Fersini, *et al.*, “SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter,” in *Proceedings of the 13th International Workshop on Semantic Evaluation*, 2019, pp. 54–63.
- [3] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet Allocation,” *Journal of Machine Learning Research*, vol. 3, no. Jan, pp. 993–1022, 2003.
- [4] W. D. Blizard, “Multiset Theory,” *Notre Dame Journal of Formal Logic*, vol. 30, no. 1, pp. 36–66, 1988.
- [5] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, “Enriching Word Vectors with Subword Information,” *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017.
- [6] B. E. Boser, I. M. Guyon, and V. N. Vapnik, “A Training Algorithm for Optimal Margin Classifiers,” in *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, 1992, pp. 144–152.
- [7] J. Canete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, and J. Pérez, “Spanish Pre-trained Bert Model and Evaluation Data,” *Practical Machine Learning for Developing Countries (PMLADC) at Eleventh International Conference on Learning Representations (ICLR)*, vol. 2020, pp. 1–10, 2020.
- [8] R. Caruana, S. Lawrence, and C. Giles, “Overfitting in Neural Nets: Backpropagation, Conjugate Gradient, and Early Stopping,” *Advances in Neural Information Processing Systems*, vol. 13, pp. 381–387, 2000.

- [9] Z. Chen, V. Badrinarayanan, C.-Y. Lee, and A. Rabinovich, "Grad-Norm: Gradient Normalization for Adaptive Loss Balancing in Deep Multitask Networks," in *Proc. ICML*, PMLR, 2018, pp. 794–803.
- [10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019, pp. 4171–4186.
- [11] L. Duan, D. Xu, and I. W. Tsang, "Learning with Augmented Features for Heterogeneous Domain Adaptation," in *Proceedings of the 29th International Conference on International Conference on Machine Learning (ICML '12)*, Omnipress, 2012, pp. 667–674.
- [12] A. E. Ezugwu, A. M. Ikotun, O. O. Oyelade, et al., "A Comprehensive Survey of Clustering Algorithms: State-of-the-art Machine Learning Applications, Taxonomy, Challenges, and Future Research Prospects," *Engineering Applications of Artificial Intelligence*, vol. 110, p. 104 743, 2022.
- [13] L. Fang, G. Liu, and R. Zhang, "Sense-aware BERT and Multi-task Fine-tuning for Multimodal Sentiment Analysis," in *Proc. IJCNN*, 2022, pp. 1–8.
- [14] J. M. de Freitas, S. Berg, B. C. Geiger, and M. Mücke, "Compressed Hierarchical Representations for Multi-task Learning and Task Clustering," in *Proc. IJCNN*, 2022, pp. 01–08.
- [15] M. Guo, A. Haque, D.-A. Huang, S. Yeung, and L. Fei-Fei, "Dynamic Task Prioritization for Multitask Learning," in *Proc. ECCV*, 2018, pp. 270–287.
- [16] S. Indurthi, M. A. Zaidi, N. Kumar Lakumarapu, et al., "Task Aware Multi-Task Learning for Speech to Text Tasks," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 7723–7727.
- [17] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7482–7491.
- [18] B. Kulis, K. Saenko, and T. Darrell, "What You Saw is Not What You Get: Domain Adaptation Using Asymmetric Kernel Transforms," in *CVPR 2011*, IEEE, 2011, pp. 1785–1792.
- [19] I. Lauriola, A. Lavelli, and F. Aielli, "An Introduction to Deep Learning in Natural Language Processing: Models, Techniques, and Tools," *Neurocomputing*, vol. 470, pp. 443–456, 2022.
- [20] M. Long, Z. Cao, J. Wang, and P. S. Yu, "Learning Multiple Tasks with Multilinear Relationship Networks," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 1593–1602.
- [21] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," in *Proc. ICLR*, 2019.
- [22] Y. Lu, A. Kumar, S. Zhai, Y. Cheng, T. Javidi, and R. Feris, "Fully-adaptive Feature Sharing in Multi-task Networks with Applications in Person Attribute Classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5334–5343.
- [23] A. F. Magnossão de Paula, R. F. da Silva, and I. B. Schlicht, "Sexism Prediction in Spanish and English Tweets Using Monolingual and Multilingual BERT and Ensemble Models," in *Proc. IberLEF'21*, 2021, pp. 356–373.
- [24] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," in *Proc. ICLR*, 2013.
- [25] D. Miljković, "Brief Review of Self-organizing Maps," in *2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 2017, pp. 1061–1066.
- [26] D. W. Otter, J. R. Medina, and J. K. Kalita, "A Survey of the Usages of Deep Learning for Natural Language Processing," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 2, pp. 604–624, 2020.
- [27] S. J. Pan and Q. Yang, "A Survey on Transfer Learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.
- [28] J. M. Pérez and F. M. Luque, "Atalaya at SemEval 2019 Task 5: Robust Embeddings for Tweet Classification," in *Proceedings of the 13th International Workshop on Semantic Evaluation*, Association for Computational Linguistics, 2019, pp. 64–69.
- [29] L. Plaza, J. Carrillo-de-Albornoz, R. Morante, et al., "Overview of EXIST 2023: sEXism Identification in Social NeTworks," in *Proc. ECIR*, Springer Nature Switzerland, 2023, pp. 593–599.
- [30] F. M. Plaza-del-Arco, M. D. Molina-González, and L. Alfonso, "SINAI at IberLEF-2021 DETOXIS Task: Exploring Features as Tasks in a Multi-task Learning Approach to Detecting Toxic Comments," in *Proc. IberLEF'21*, 2021, pp. 580–590.
- [31] P. Prettnerhofer and B. Stein, "Cross-language Text Classification Using Structural Correspondence Learning," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 2010, pp. 1118–1127.
- [32] F. Rodríguez-Sánchez, J. Carrillo-de-Albornoz, L. Plaza, et al., "Overview of EXIST 2021: sEXism Identification in Social neTworks," *Procesamiento del Lenguaje Natural*, vol. 67, pp. 195–207, 2021.
- [33] S. Ruder, "An Overview of Multi-Task Learning in Deep Neural Networks," *CoRR*, vol. abs/1706.05098, 2017. arXiv: 1706.05098.
- [34] S. Ruder, M. E. Peters, S. Swayamdipta, and T. Wolf, "Transfer Learning in Natural Language Processing," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, 2019, pp. 15–18.
- [35] O. Sener and V. Koltun, "Multi-Task Learning as Multi-Objective Optimization," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018, pp. 525–536.
- [36] A. T. Sinha, A. Rabinovich, Z. Chen, and V. Badrinarayanan, *Gradient adversarial training of neural networks*, US Patent App. 17/051,982, 2021.
- [37] M. Taulé, A. Ariza, M. Nofre, E. Amigó, and P. Rosso, "Overview of DETOXIS at IberLEF 2021: DEtection of TOxicity in comments In Spanish," *Procesamiento del Lenguaje Natural*, vol. 67, pp. 209–221, 2021.
- [38] S. Vandenhende, S. Georgoulis, L. V. Gool, and B. D. Brabandere, "Branched Multi-task Networks: Deciding What Layers to Share," in *Proceedings of the 31st British Machine Vision Conference (BMVC '20)*, BMVA Press, 2020.
- [39] S. Vandenhende, S. Georgoulis, W. Van Gansbeke, M. Proesmans, D. Dai, and L. Van Gool, "Multi-task Learning for Dense Prediction Tasks: A Survey," *IEEE TPAMI*, vol. 44, no. 7, pp. 3614–3633, 2022.
- [40] A. Vaswani, N. Shazeer, N. Parmar, et al., "Attention is All You Need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 6000–6010.
- [41] A. Vouloimos, N. Doulamis, A. Doulamis, E. Protopapadakis, and D. Andina, "Deep Learning for Computer Vision: A Brief Review," *Computational Intelligence and Neuroscience*, vol. 2018, 2018.
- [42] C. Wang and S. Mahadevan, "Heterogeneous Domain Adaptation Using Manifold Alignment," in *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Two*, 2011, pp. 1541–1546.
- [43] Y. Wang, M. Xu, Y. Yan, T. Zhao, Y. Chen, and J. Yang, "Exploring Topic Supervision with BERT for Text Matching," in *Proc. IJCNN*, 2022, pp. 1–7.
- [44] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A Survey of Transfer Learning," *Journal of Big Data*, vol. 3, no. 1, pp. 1–40, 2016.
- [45] J. Worsham and J. Kalita, "Multi-task Learning for Natural Language Processing in the 2020s: Where are We Going?" *Pattern Recognition Letters*, vol. 136, pp. 120–126, 2020.
- [46] S. Wu, H. R. Zhang, and C. Ré, "Understanding and Improving Information Transfer in Multi-task Learning," in *Proc. ICLR*, 2020.
- [47] S. Wu, H. Fei, and D. Ji, "Aggressive Language Detection with Joint Text Normalization via Adversarial Multi-task Learning," in *CCF International Conference on Natural Language Processing and Chinese Computing*, 2020, pp. 683–696.
- [48] D. Xu, W. Ouyang, X. Wang, and N. Sebe, "Pad-net: Multi-tasks Guided Prediction-and-distillation Network for Simultaneous Depth Estimation and Scene Parsing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 675–684.
- [49] Y. Zhang and Q. Yang, "A Survey on Multi-task Learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 12, pp. 5586–5609, 2022.
- [50] Z. Zhang, Z. Cui, C. Xu, Z. Jie, X. Li, and J. Yang, "Joint Task-recursive Learning for Semantic Segmentation and Depth Estimation," in *Proc. ECCV*, 2018, pp. 235–251.
- [51] Z. Zhang, Z. Cui, C. Xu, Y. Yan, N. Sebe, and J. Yang, "Pattern-affinitive Propagation Across Depth, Surface Normal and Semantic Segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4106–4115.
- [52] X. Zhao, H. Li, X. Shen, X. Liang, and Y. Wu, "A Modulation Module for Multi-task Learning with Applications in Image Retrieval," in *Proc. ECCV*, 2018, pp. 401–416.