Expert Systems

WILEY

Expert Systems

ORIGINAL ARTICLE OPEN ACCESS

Mitigating the Negative Transfer in Multi-Task Learning for Harmful Language Detection in Spanish and Arabic

Angel Felipe Magnossão de Paula[1,2] | Imene Bensalem[3,4] | Damiano Spina[2] | Paolo Rosso[1,5]

[1]Department of Computer Systems and Computation, Universitat Politècnica de València, València, Spain | [2]School of Computing Technologies, RMIT University, Melbourne, Victoria, Australia | [3]MISC-Lab, Constantine 2 University, Constantine, Algeria | [4]ESCF de Constantine, Constantine, Algeria | [5]ValgrAI, Valencian Graduate School and Research Network of Artificial Intelligence, València, Spain

Correspondence: Angel Felipe Magnossão de Paula (adepau@doctor.upv.es)

Received: 27 September 2025 | **Revised:** 14 November 2025 | **Accepted:** 28 November 2025

Keywords: hate speech | multi-task learning | negative transfer | offensive language | sexism | toxic language

ABSTRACT

Negative transfer continues to limit the benefits of multi-task learning (MTL) in harmful language detection, where related tasks must share representations without diluting task-specific nuances. We introduce task awareness (TA), a methodological framework that explicitly conditions MTL models on the task they must solve. TA is instantiated through two complementary mechanisms: Task-aware input (TAI), which augments textual inputs with natural-language task descriptions, and task embedding (TE), which learns task-specific transformations guided by a task identification vector. Together they enable the encoder to disentangle shared and task-dependent signals, reducing interference during joint optimisation. We integrate TA with BETO and AraBERT encoders and evaluate on six Spanish and Arabic datasets covering sexism, toxicity, offensive language, and hate speech. Across cross-validation and official train-test splits, TA consistently mitigates negative transfer, surpasses single-task and conventional MTL baselines, and yields new state-of-the-art scores on EXIST-2021, HatEval-2019, and HSArabic-2023. The proposed methodology therefore combines a principled architectural innovation with demonstrated practical gains for multilingual harmful language detection. The resources to reproduce our experiments are publicly available at https://github.com/AngelFelipeMP/Arabic-MultiTask-Learning.

1 | Introduction

Machine learning applications are widespread, covering areas from natural language processing (NLP)—which includes tasks like named-entity recognition and automated hate speech detection—to computer vision (CV), enabling systems for object detection and classification (Otter et al. 2020; Lauriola et al. 2022; Voulodimos et al. 2018; Jamil et al. 2023). Standard practice often involves training a dedicated model or ensemble for each task, refining it until further performance gains are negligible. While this single-task learning (STL) approach frequently produces acceptable outcomes, it fails to leverage potential knowledge sharing from related tasks, which might otherwise improve model generalisation. Furthermore, insufficient data can impede the development of robust models. To overcome these limitations, several methods have been proposed for transferring knowledge across different tasks (Kulis et al. 2011; Zhu et al. 2023).

An emerging area, multi-task learning (MTL) (Ruder 2017; Aguilar et al. 2017; Plaza-del-Arco et al. 2021, 2021a, 2021b; Zhang and Yang 2022; Chen et al. 2024), aims to exploit synergies between various tasks, potentially lowering requirements for data and computational power. MTL endeavours to enhance generalisation by concurrently training on several tasks. Within MTL using neural networks, two prevalent techniques are *soft* (Wu, Fei, and Ji 2020; Wang et al. 2022) and *hard parameter-sharing* (Fang et al. 2022; De Freitas et al. 2022).

In soft parameter-sharing, task-specific networks are employed, while cross-task communication is facilitated through feature-sharing methods to encourage parameter similarity. However, scalability can be an issue, as the size of the multi-task network increases linearly with the number of tasks. In contrast, hard parameter-sharing strategy divides the parameter set into shared and task-specific components, often implemented using a shared encoder with multiple task-specific decoding heads (Zhang and Yang 2022; Chen et al. 2024). This approach has the added benefit of reducing overfitting (Ruder 2017).

The hard parameter-sharing framework has been augmented by multilinear relationship networks (Long et al. 2017), which apply tensor normal priors to the parameters of fully connected layers. Nevertheless, selecting branching points arbitrarily in these networks can result in suboptimal task arrangements. Tree-based structures (Lu et al. 2017; Vandenhende et al. 2020) have been suggested to address this issue.

Despite these developments, learning multiple tasks concurrently can sometimes result in *negative transfer* (Vandenhende et al. 2022; Wu, Zhang, and Ré 2020). This occurs when shared noisy information between tasks impairs the model's performance. Negative transfer signifies a reduction in the model's effectiveness on target tasks due to knowledge transfer (Wu, Zhang, and Ré 2020; Vandenhende et al. 2022).

This work introduces a novel method to address the negative transfer challenge, utilising the concept of task awareness (TA) (Magnossão de Paula et al. 2023). Our technique allows MTL models to use information about the particular task being processed, enabling the model to prioritise its internal weights appropriately for each task. In contrast to state-of-the-art (SOTA) approaches (see Section 2), our method avoids recursive structures, thus conserving computational resources and time.

Employing the TA concept, we devised two mechanisms integrated into two distinct MTL TA (MTL-TA) architectures. The goal of these architectures is to tackle SOTA difficulties in identifying sexism, toxic language, and hate speech within Spanish text, as well as sexism, offensive language, and hate speech in Arabic comments.

Examples illustrating each task in its original language, accompanied by English translations, are provided in Table 1. This table also specifies the source dataset for each text sample. Section 4.1 offers a detailed account of the datasets used.

Although hate speech, sexism, offensive language, and toxic language represent related concepts (Poletto et al. 2021; Alkomah and Ma 2022; Pachinger et al. 2023; Bensalem et al. 2024), they each possess unique characteristics and societal consequences. Hate Speech typically targets specific demographics (Plaza-del-Arco et al. 2021, 2021a, 2021b), whereas Sexism relates to gender-based discrimination (Frenda et al. 2019). Offensive and Toxic Language are broader terms encompassing expressions that promote hostility or negativity (Derczynski et al. 2024; Magnossão de Paula and Schlicht 2021). The conceptual overlaps between these tasks are depicted in the Venn diagram (Figure 1). Given their interrelations, creating MTL models capable of identifying these various forms of harmful language presents a valuable opportunity.

Among the languages most frequently used on social media platforms like Twitter, Facebook, and TikTok are Spanish and Arabic. For instance, data indicates Spanish ranks as the third most common language on Twitter, with Arabic fourth Alshaabi et al. (2021). Regrettably, the prevalence of a language often correlates with the volume of hostile and harmful content generated in it. We believe this paper is the first to put forward effective strategies for reducing negative transfer across numerous sensitive tasks in both Spanish and Arabic. The source code developed for this study is openly accessible.[1]

The primary contributions of this research include:

- Introduction of task awareness (TA): Our paper introduces the concept of TA and proposes two unified architectures equipped with TA mechanisms (MTL-TAI & MTL-TE) that can mitigate the negative transfer phenomenon during MTL training.
- Development of TA mechanisms: To equip MTL models with task recognition capabilities, we developed the Task-aware input (TAI) and task embedding (TE) mechanisms, aimed at alleviating negative transfer and improving performance compared to traditional MTL approaches.
- Validation of MTL-TA models: We evaluated the effectiveness of the two TA-equipped architectures in detecting sexism, toxic language, and hate speech in Spanish comments, as well as sexism, offensive language, and hate speech in Arabic textual comments. The results demonstrate that both MTL-TAI and MTL-TE mitigate negative transfer in these two languages.2
- Achieving SOTA performance: Our approach exceeds SOTA results on established public benchmarks for detecting sexism (EXIST-2021) and hate speech (HatEval-2019). Furthermore, it sets a new SOTA benchmark for the HSArabic-2023 dataset concerning offensive language identification, marking considerable advancements over prior techniques.

The novelty of this work lies in its focus on methods and practical applications. We present effective architectural mechanisms that implement the TA principle within MTL systems and demonstrate how these mechanisms can be integrated into real moderation workflows. While the concept of using task cues to guide shared representations is rooted in the theory of negative transfer, our key advances are in the practical design, integration, and empirical validation of TAI and TE for detecting harmful language in Spanish and Arabic.

The rest of this paper is structured as follows. Section 2 reviews related work on transfer learning, MTL and its application to Harmful Language detection. Section 3 describes our proposed method in detail. Section 4 outlines the experimental setup. Section 5 presents and analyzes the experimental outcomes.

TABLE 1 | Examples of tweets containing sexism, hate speech, toxic language, and offensive language in Spanish and Arabic.

Task	Dataset	Original language	English translation
Sexism	EXIST-2021	@USER Que. rica putita obediente, afortunado tu marido de tener una mujer como tú, saludos	@USER What a nice obedient little whore, your husband is lucky to have a woman like you, greetings
	ArMI-2021	مستخدما مستخدما مستخدما الزوجة تهتم بالمطبخ و نظافة و ترتيب البيت و تجلس في بيتها و الزوج يعمل من أجل	USER@ USER@ USER@ The wife takes care of the kitchen, cleanliness and organisation of the house and sits at home, while the husband works for his home.
Hate speech	HatEval-2019	Hay varias paginas de feministas a las que deberia darles verguenza exponer sus ideas ya que no tienen ni pies ni cabeza	There are several feminist pages that should be ashamed of themselves for expressing their ideas because they have no head or tai
	HSArabic-2023	RT @USER: #حزبالله سوف نغلق نباحك ايها الارهابي الدجال النتن الى الابد https://t.co/nbSrVJdVlm	RT @USER: #Hezbollah We will shut down your barking, you stinking, liar terrorist, forever https://t.co/nbSrVJdVlm
Toxic language	DETOXIS-2021	Está claro que vienen los mejores. Haced que pase putos rojos de mierda.	It's clear that the best are coming. Make it happen you fucking Reds
Offensive language	OSACT-2022	RT @USER: اصلاً لو العربية ما جابت هالخبر كان استغربنا .. متعودين على اخبارها السخيفة اللي تظهر للعالم ابيخ صورة للشعب السعودي URL	RT @USER If Al Arabiya had not reported this news, we would have been surprised. We are used to its ridiculous news that shows the world the worst image of the Saudi people URL

Note: The original texts are provided alongside their English translations, which were generated using Google Translate.

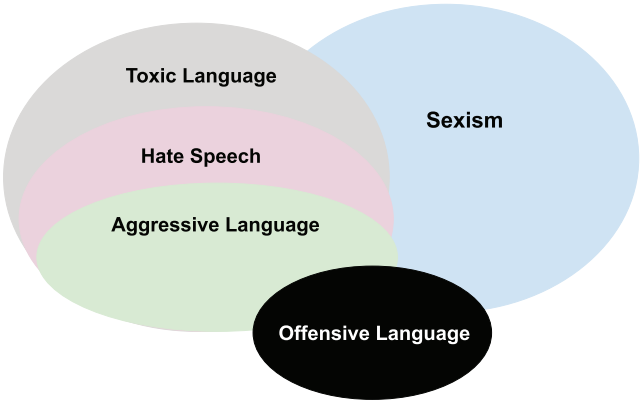

FIGURE 1 | Adaptation of the (Poletto et al. 2021; Alkomah and Ma 2022) Venn diagrams showing the relationships among sexism, hate speech, toxic language, and offensive language.

Potential limitations of our method are discussed in Section 6. Finally, Section 7 offers concluding remarks and outlines potential avenues for future investigation.

2 | Related Work

2.1 | Transfer Learning and Multi-Task Learning

Transfer learning represents a common machine learning method, founded on the idea that a model developed for one task can be improved by integrating knowledge from a related task (Pan and Yang 2009; Weiss et al. 2016; Zhu et al. 2023). Training models entirely from scratch often demands significant data and computational power; however, situations arise where acquiring enough training data is excessively costly or infeasible. This necessitates creating high-performing learning systems using more accessible data from alternate tasks. Knowledge transfer methods facilitate enhancing target task performance by utilising information derived from associated tasks. Such methods have seen successful deployment in diverse machine learning domains, notably NLP (Ruder et al. 2019; Wang and Mahadevan 2011; Prettenhofer and Stein 2010; Wang et al. 2022) and CV (Duan et al. 2012; Kulis et al. 2011). Closely associated with transfer learning is the MTL framework (Ruder 2017; Zhang and Yang 2022), which

seeks to learn multiple, potentially different, tasks in parallel. The effectiveness of this paradigm stems from its capacity to leverage shared information across tasks. Nevertheless, if the tasks lack sufficient relatedness, negative transfer can occur. This term describes performance decline resulting from sharing noisy or inappropriate information between tasks (Wu, Zhang, and Ré 2020; Vandenhende et al. 2022).

Recent studies have focussed on identifying and mitigating negative transfer in MTL. Li et al. (2023) introduced a surrogate modelling approach to predict and partially prevent negative transfer by estimating relevance scores for each task. This method significantly improves the accuracy of MTL by selecting optimal task subsets.

Hierarchical Prompt Learning (HiPro), as proposed by Liu et al. (2023), demonstrates how a hierarchical task-sharing approach can reduce the risks of negative transfer. By organising tasks into more granular groups based on their relatedness, HiPro constructs a task tree that allows the model to learn both shared and individual task prompts, balancing generalisation with task-specific adaptation.

Several approaches have been proposed to address negative transfer and balance learning across different tasks. These include re-weighting of losses through methods like Homoscedastic uncertainty (Cipolla et al. 2018), Gradient normalisation (Chen et al. 2018), and Adversarial training (Sinha et al. 2021), as well as task prioritisation (Guo et al. 2018; Zhao et al. 2018; Sener and Koltun 2018). Additionally, other approaches (Xu et al. 2018; Zhang et al. 2018, 2019) utilise initial predictions from multi-task networks to iteratively refine each task's output, thereby overcoming the limitations of methods that compute all task outputs simultaneously. However, these approaches are often time-consuming and require substantial computational resources due to their recursive nature.

Knight and Duan (2023) introduced an innovative framework that uses summary statistics to address the challenges of MTL in data-sharing-constrained environments, such as healthcare settings. This approach enables efficient model training without the need for access to individual-level data, preserving privacy while still benefiting from shared information across tasks.

In the domain of NLP, Chen et al. (2024) provide an overview that underscores the importance of MTL in mitigating overfitting and addressing data scarcity. The authors review MTL architectures and optimisation techniques, demonstrating how MTL can leverage related tasks to enhance overall performance across NLP applications.

2.2 | MTL In Harmful Language Detection

The initial semi-supervised multi-task method for Sexism classification was introduced by Abburi et al. (2020). Their work addressed three tasks utilising labels derived from unsupervised learning or weak labelling processes. The neural multi-task architecture they designed facilitates shared learning among tasks through common weights and an aggregated loss function, surpassing several SOTA baselines.

Wu, Fei, and Ji (2020) put forward a novel MTL strategy to concurrently manage Aggressive Language Detection (ALD) and text normalisation. They employed a shared encoder for learning common inter-task features and a task-specific encoder for task-relevant features. This configuration led to considerable performance gains in ALD.

Abu Farha and Magdy (2020) proposed CNN-BiLSTM-based models trained for three tasks: Hate speech detection, offensive language detection, and sentiment analysis. The authors evaluated their models using the OSACT2020 (Mubarak et al. 2020) Arabic dataset, demonstrating that their multi-task architecture outperformed traditional monotask models.

In this paper, we introduce two unified architectures designed to identify sexism, toxic language, and hate speech in comments written in Spanish, as well as to detect sexism, offensive language, and hate speech in Arabic textual comments. The designed architectures intend to lessen the negative transfer effect in MTL training, consequently enhancing the identification rate for harmful content.

The methodology detailed here draws inspiration from mechanisms suggested by (Abburi et al. 2020; Wu, Fei, and Ji 2020). While those methods focus on refining the representations passed to task heads to improve MTL models, our TA technique distinguishes itself. It empowers the model to independently ascertain the task it needs to execute. Consequently, MTL-TA models can generate suitable representations for every task head without needing an auxiliary learning task, enhancing efficiency. The core concept involves learning a task-pertinent latent data representation capable of effectively addressing multiple NLP tasks (Wang et al. 2022; Indurthi et al. 2021). The specific mechanisms developed are elaborated upon in the subsequent section.

3 | Proposed Approach

This section details the MTL-TA models introduced in (Magnossão de Paula et al. 2023). We begin by introducing the concept of TA and explaining its potential in reducing the effects of negative transfer (Vandenhende et al. 2022; Wu, Zhang, and Ré 2020) in multi-task joint training (Ruder 2017). Following this, we introduce two specific TA mechanisms designed for incorporating task self-awareness into MTL models.

Our methodological innovation is anchored in three design principles. First, we expose the encoder to explicit task descriptors so that the shared representation can be shaped by both linguistic content and the downstream objective. Second, we interpose a lightweight task-conditional transformation that can reconfigure the shared representation before it reaches each task head, thereby curbing interference from unrelated gradients. Third, we ensure both mechanisms can be trained end-to-end within standard hard-parameter-sharing pipelines,

allowing practitioners to retrofit TA into existing moderation systems without extensive re-engineering. The remainder of this section formalises these ideas and details how they differ from prior MTL formulations.

The most common technique for supervised MTL utilises the hard parameter-sharing strategy (Zhang and Yang 2022). In this configuration, the model comprises an encoder alongside N decoders (or task heads), with N representing the count of tasks the model trains on concurrently (Worsham and Kalita 2020). During operation, the encoder takes an input and produces a task-neutral latent representation, subsequently passed to the designated task head for the final prediction.

However, a weaker connection between the encoder's generated latent representation and the specific tasks can impair overall MTL model efficacy (Vandenhende et al. 2022). It is probable that the ideal latent representations for identical inputs will differ across various task heads (De Freitas et al. 2022). Furthermore, during the training phase, the encoder's representation might develop a bias towards tasks that are more complex or possess larger datasets (Ruder 2017). Such reductions in performance exemplify the negative transfer issue (Vandenhende et al. 2022; Wu, Zhang, and Ré 2020), where a task head gets an unsuitable input representation, hindering its capacity to effectively address its assigned task.

To mitigate negative transfer when tackling multiple NLP tasks using the MTL approach (Zhang and Yang 2022), we propose two TA mechanisms. These mechanisms tailor the task heard input representation based on the specific task being addressed, ensuring that the representation sent to each respective head is optimised for that task. Furthermore, our proposed MTL model continues to benefit from the generalisation improvements provided by multi-task joint training. Updates during training apply to the encoder and other MTL model components preceding the task heads for every task. It is crucial to recognise that all our suggested MTL models belong to the MTL-TA category and follow the standard MTL paradigm. Consequently, parameter updates only involve the specific task head corresponding to the current input data.

This holistic formulation means that TA augments, rather than replaces, classic hard-parameter sharing: practitioners can reuse established optimisation recipes while equipping the model with explicit mechanisms to preserve task-specific signals. In Section 4 we describe how this design readily scales to heterogeneous datasets without bespoke tuning for each task pairing.

3.1 | TAI

The initial mechanism formulated to embed TA within MTL models is the TAI. To help the encoder produce appropriate representations for every task head, we suggest altering the standard MTL input structure for NLP applications.

Concretely, for each sample we concatenate a short natural-language task description (TD) to the original text snippet and rely on the encoder's positional embeddings to disentangle the segments. This approach aligns the latent space with task semantics from the very first layer, steering the encoder to highlight lexical and syntactic cues that are predictive for the specified task. Unlike prompt-based conditioning that often requires task-specific templates or additional pre-training, TAI uses a uniform schema that can be populated automatically from dataset metadata, which makes it robust across languages and label distributions.

The TAI comprises a text snippet (TS) paired with a TD, illustrated in Figure 2. The TS represents a segment of text, variable in length based on the task, and usually serves as the primary input for MTL encoders. The TD is textual information specifying the particular task managed by a certain head, for instance, "sexism detection" or "hate speech detection". This adjusted input supplies context to the encoder, facilitating the creation of a task-focused representation. An MTL model incorporating the TAI mechanism is denoted as MTL task-aware input (MTL-TAI).

3.2 | TE

The second mechanism conceived to bestow TA capability upon MTL models is named TE. We propose inserting an extra component situated between the task heads and the encoder, designated as the task embedding block (TEB), depicted in Figure 3. This block takes two inputs: (i) the task identification vector (TIV), and (ii) the latent representation from the encoder. The TIV is constructed as a one-dimensional one-hot vector, its length matching the number of task heads. Every position within the TIV corresponds to a distinct task head.

The TEB is composed of learning units (LU), each containing a linear layer succeeded by a ReLU activation function. The quantity of LUs acts as a hyperparameter, influenced by factors like task type and data properties. The primary goal of the TEB is to craft a fitting representation for the task the MTL model is currently addressing. Consequently, given the same encoder output, the TEB yields varied outputs contingent on the task. It uses the TIV as a signal to determine for which task the representation should be generated. As shown in Figure 3, the TIV features a "1" at the index related to the task being processed, while all other positions are zero. An MTL model utilising the TE mechanism is identified as MTL task embedding (MTL-TE).

The TEB acts as a learned gating function that re-weights shared features according to the active task. During training, gradients flowing through the TIV-conditioned layers encourage the model to isolate features that consistently help a given task while suppressing features that trigger negative transfer. Because the same parameters are reused across tasks with different activations, TE mechanism promotes parameter efficiency while still enabling task-specific specialisation.

4 | Experimental Setup

This section commences by detailing the tasks and datasets employed for evaluating our method. Subsequently, it provides implementation specifics and reference models. Lastly, it outlines the experimental configurations.

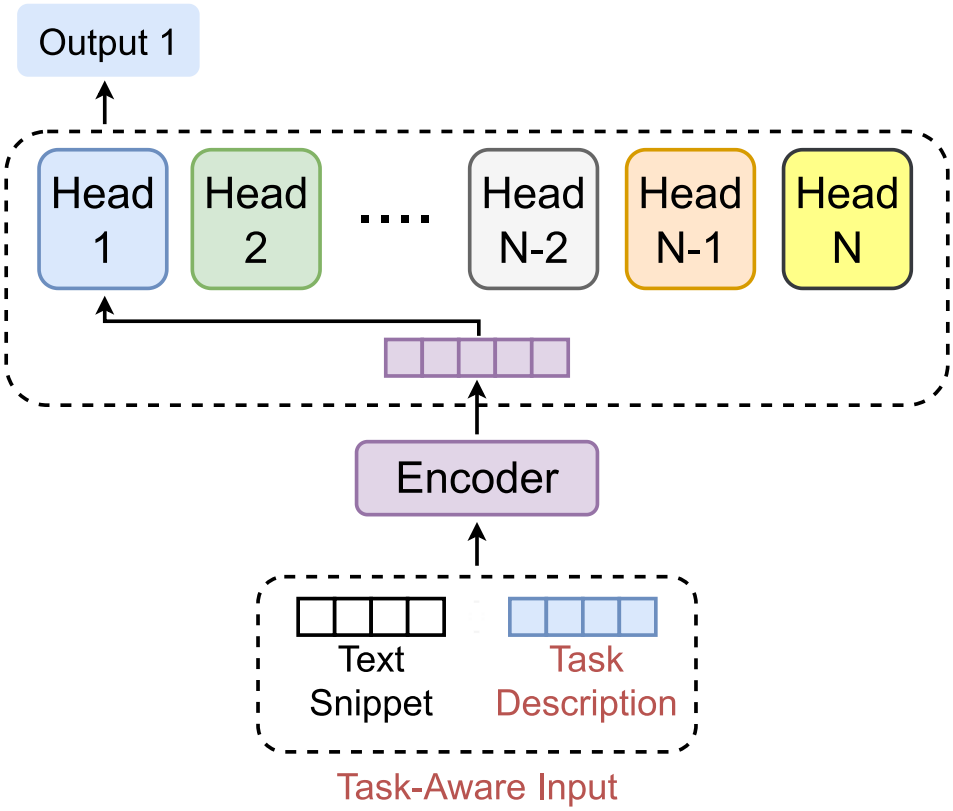

FIGURE 2 | Multi-task learning (MTL) model including task-aware input (TAI) mechanism (MTL-TAI).

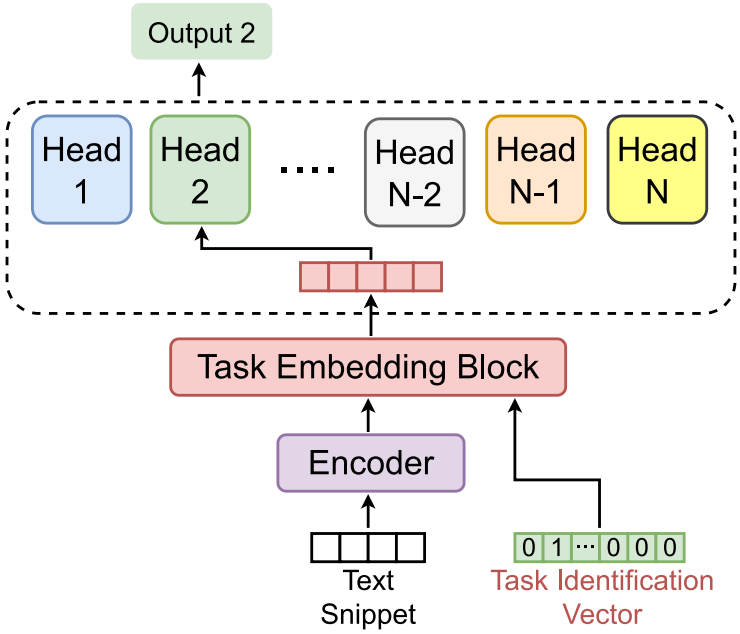

FIGURE 3 | Multi-task learning (MTL) model including task embedding (TE) mechanism (MTL-TE).

4.1 | Data

In addition to addressing the negative transfer problem, our research focussed on identifying and filtering harmful and abusive content in social media text. Our overarching goal is to promote a respectful and inclusive environment by preventing the spread of discriminatory or harmful language online and in other communication channels. We decided to focus on Spanish and Arabic, as these languages are among the most widely used on social media Alshaabi et al. (2021). However, the prevalence of these languages also corresponds with the production of hostile and harmful content. Therefore, it is crucial to make substantial efforts to tackle harmful language behaviour in Spanish and Arabic, rather than English (Plaza-del-Arco et al. 2021, 2021a, 2021b). For Spanish, we tackled tasks related to detecting Sexism, Toxic Language and Hate

TABLE 2 | EXIST-2021 data distribution (Rodríguez-Sánchez et al. 2021).

	Training		Test			
	Spanish	English	Spanish		English	
	Twitter	Twitter	Twitter	Gab	Twitter	Gab
Sexist	1741	1636	858	265	858	300
Not-Sexist	1800	1800	812	225	858	192

Speech. In Arabic, we focussed on detecting Sexism, Offensive Language, and Hate Speech. We utilised six datasets, each tailored to specific tasks within these languages. For Spanish, we used the EXIST-2021 (Rodríguez-Sánchez et al. 2021), DETOXIS-2021 (Taulé et al. 2021), and HateEval-2019 (Basile et al. 2019) datasets. For Arabic, we employed the ArMI-2021 (Mulki and Ghanem 2021), HSArabic-2023, and OSACT-2022 (Mubarak et al. 2022) datasets. Below is a detailed description of each dataset:

4.1.1 | EXIST-2021 (Rodríguez-Sánchez et al. 2021)

This dataset originated from the sExism Identification in Social neTworks (EXIST) shared task during the Iberian Languages Evaluation Forum (IberLEF) 2021. It comprises 11,345 annotated social media posts (in English and Spanish) sourced from Twitter and the uncensored platform Gab.com (Gab). Experts in gender issues supervised and monitored the dataset's creation. EXIST represented the inaugural challenge focused on social media Sexism detection, aiming to identify Sexism broadly, encompassing explicit misogyny to subtler sexist actions. The task related to Sexism identification attracted 70 official submissions. It involves binary classification, categorising samples as either Sexist or Not-Sexist. Accuracy served as the official evaluation standard, with data partitioned into training and test sets. Table 2 presents the data breakdown.

4.1.2 | DETOXIS-2021 (Taulé et al. 2021)

Data collection for this set occurred for the DEtection of TOxicity in comments In Spanish (DETOXIS) shared task at IberLEF 2021. The task's goal was detecting Toxic Language within comments responding to online news articles concerning immigration. The annotation method developed aimed at reducing subjectivity in toxicity labeling by considering context (like linguistic cues and conversation threads). The data annotation team comprised trained annotators and linguistics experts. The collection contains 4354 text comments responding to various articles from Spanish online news sources (e.g., ABC, elDiario.es, El Mundo, NIUS) and discussion platforms (like Menéame). The task involves a binary classification, assigning samples to either Toxic or Not-Toxic categories. Over 30 teams assessed their machine learning models using this dataset during the DETOXIS shared task participation. The F1-score for the Toxic class was the official evaluation measure, and the dataset was split into training and testing portions. The data distribution is shown in Table 3.

TABLE 3 | DETOXIS-2021 data distribution (Taulé et al. 2021).

	Training	Test
Toxic	1147	239
Not-toxic	2316	652

4.1.3 | HatEval-2019 (Basile et al. 2019)

This dataset was assembled for the Detection of Hate Speech Against Immigrants and Women in Twitter (HatEval) task, a component of the SemEval 2019 workshop. It includes 19,600 tweets (English and Spanish) with labels for hate speech detection. The collection process involved several gathering techniques: (i) observing accounts likely targeted by hate; (ii) obtaining records from known haters; (iii) applying keyword filters to Twitter feeds. Annotation involved experts and crowdsourced workers verified for annotation reliability. The task required binary classification, associating samples with hateful or not-hateful labels. The dataset contains training, development, and test partitions. The official metric was F1-macro (the unweighted average F1-score across both classes). HatEval ranked among SemEval 2019's most participated tasks, receiving over 100 submissions for hate speech detection. Table 4 details the dataset's composition.

4.1.4 | ArMI-2021 (Mulki and Ghanem 2021)

The dataset served the Arabic Misogyny Identification shared task (ArMI), part of the hate speech and offensive content detection (HASOC) track at FIRE-2021. It comprises 9833 tweets in formal Arabic and various dialects, including Levantine, Gulf, and Egyptian. These tweets were gathered using different expressions and hashtags related to anti-women topics, as well as from the accounts of seven female journalists who were active during the Lebanon protests in October 2019. The annotators labelled the tweets for sexism identification (a binary task) and sexism categorization (a multi-class task). The challenge received 15 official runs. The official evaluation metric was accuracy, and data was split into training and test sets. The dataset statistics are presented in Table 5, with our focus exclusively on the binary task.

4.1.5 | HSArabic-2023

This dataset was created by Hamad Bin Khalifa University and Carnegie Mellon University in Qatar as part of a research project. It contains more than 15,000 tweets in different Arabic dialects. Annotators from various Arabic-speaking countries in the Middle East and North Africa labelled the tweets. The

TABLE 4 | HatEval-2019 data distribution (Basile et al. 2019).

	Training		Development		Test	
	Spanish	English	Spanish	English	Spanish	English
Hate	1741	1636	1741	1636	858	300
Not-hate	1800	1800	1800	1800	812	192

TABLE 5 | ArMI-2021 data distribution (Mulki and Ghanem 2021).

	Training	Test
Sexist	4805	1201
Not-sexist	3061	766

TABLE 6 | HSArabic-2023 data distribution.

	Training	Test
Offensive	2234	559
Not-offensive	10,057	2514

TABLE 7 | OSACT-2022 data distribution (Mubarak et al. 2022).

	Training	Development	Test
Hate	959	109	271
Not-hate	7928	1161	2270

annotations cover different tasks, including Offensive Language detection, which we considered in our experiments. An updated version of the dataset is described in the paper by (Charfi et al. 2024). We divided the dataset into training and test subsets, as shown in Table 6 and adopted F1-macro as an official evaluation metric.

4.1.6 | OSACT-2022 (Mubarak et al. 2022)

This dataset was utilised in the OSACT 2022 shared task on Arabic offensive language and hate speech detection. It comprises 12,698 examples collected from Twitter using a predefined list of emojis frequently associated with offensive texts. The tweets were labelled via a crowdsourcing platform for three subtasks: Offensive language detection (binary task), Hate Speech detection (binary task), and fine-grained hate speech detection (multi-class task). The data is composed of training, development, and test sets. In total, 40 teams signed up to participate in the offensive language detection task, and the official evaluation metric was the F1-macro. Our analysis focuses on the binary annotations, with statistics detailed in Table 7.

4.2 | Implementation Details

The encoder employs a bidirectional encoder representation from transformers (BERT) (Devlin et al. 2019), pre-trained in the language of the applied task data. We utilised the most popular BERT versions for each language: BERT for Spanish transformers (BETO) (Canete et al. 2020) and Arabic BERT (AraBERT) (Antoun et al. 2020). Following the BERT encoding, we applied both max pooling and mean pooling calculations to its output. These BERT models consist of 12 self-attention layers, each with 12 attention heads, and a hidden size of 768 dimensions, totaling approximately 110 million parameters.

The respective encoder (BETO or AraBERT) handles a text sequence, yielding a hidden representation per token equivalent to the 768 hidden size dimensions. Concatenating the max and mean pooling results derived from the full sequence of encoder output tokens forms the latent encoder representation. Within the TE method, the TEB preserves the precise dimensionality of this latent encoder representation.

Task heads function as linear classifiers. Their input dimensions align with the latent encoder representation, while output dimensions vary by task. For binary classification tasks, the linear classifier outputs two values; the larger value determines the predicted class. Additionally, task descriptions (TDs) for each dataset were formulated by appending 'detection' to the task name; for example, for EXIST-2021 (Rodríguez-Sánchez et al. 2021), the TD used was "Sexism Detection".

Model training utilised the AdamW optimizer (Loshchilov and Hutter 2019), incorporating a linear decay learning rate schedule spanning from 5e-6 to 1e-4. Training involved 15 epochs, a dropout rate of 0.3, and a batch size of 64. We tested configurations using 1, 2, and 3 LUs. Adopting an approach akin to early stopping (Caruana et al. 2000), the model demonstrating best performance on the task's official metric was chosen.

4.3 | Comparison Models

Our method is compared against two model categories: (i) Baseline models and (ii) SOTA models. We specifically implemented two baseline types: STL and MTL models. These baselines are essential for determining whether negative transfer occurred during the training of the classic MTL model, as revealed by comparing the models' performance on the test data. Negative transfer is identified when the classic MTL model performs worse—according to the chosen evaluation metric—than the STL model. In such cases, we further evaluate the performance of the MTL Task-Aware (MTL-TA) models against the classic MTL model to determine whether our proposed solutions effectively address the negative transfer issue by achieving superior results. Below there is a detailed description of the two modes:

- **MTL** refers to the standard MTL model. Its construction mirrors the MTL-TA model architecture (see Section 3) but lacks the TAI mechanism. Consequently, the MTL model processes only the TS as input.
- **STL** designates the standard STL model. While sharing the MTL model's architecture, it features just one task head. Therefore, comparing this model type against MTL models necessitates training a separate model for every task addressed.

SOTA models signify the top-performing methods currently available for the datasets included in our experiments. Comparing the performance of classic MTL, SOTA, and MTL-TA models provides valuable insights into how effectively a simple MTL model can approach SOTA results when negative transfer is mitigated. Below is a comprehensive overview of the SOTA models:

- **AI-UPV** (Magnossão de Paula et al. 2021): A deep learning architecture leveraging a combination of different Transformer models (Vaswani et al. 2017). It capitalises on ensemble techniques and incorporates data augmentation during training. This model holds the SOTA position for EXIST-2021 (Rodríguez-Sánchez et al. 2021).
- **SINAI** (Plaza-del-Arco et al. 2021, 2021a, 2021b): A BERT base model (Devlin et al. 2019) trained via the MTL hard parameter-sharing approach. Despite covering five tasks and six datasets, its primary focus was Toxic Language detection, utilising other tasks as auxiliary support. It represents the SOTA for DETOXIS-2021 (Taulé et al. 2021).
- **Atalaya** (Pérez and Luque 2019): This model employs Support Vector Machines (Boser et al. 1992). Training involved multiple representations derived from FastText (Bojanowski et al. 2017) sentiment-focused word vectors, including tweet embeddings (Mikolov et al. 2013), bag-of-characters (Bojanowski et al. 2017), and bag-of-words (Blizard 1988). It is recognised as the SOTA for HatEval-2019 (Basile et al. 2019).
- **UM6P-NLP** (Mahdaouy et al. 2021): This approach utilises MARBERT (Abdul-Mageed and Elmadany 2021), a language model pre-trained on a corpus of 1 billion tweets spanning various Arabic dialects. It uses a multi-task methodology, formulated to address both binary and multiclass classification tasks within the ArMI-2021 shared task (Mulki and Ghanem 2021), where it secured SOTA status.
- **GOF** (Mostafa et al. 2022): This method uses an ensemble strategy based on majority voting among three pre-trained models: QARiB (Abdelali et al. 2021), MARBERT, and MERBERT v2 (Abdul-Mageed and Elmadany 2021). Optimization for each model involved a distinct loss function. It stands as the SOTA for OSACT-2022 (Mubarak et al. 2022).

4.4 | Evaluation Setup

Two experiments were performed to assess our TA method's efficacy in reducing negative transfer (Vandenhende et al. 2022; Wu, Zhang, and Ré 2020), detailed below.

4.4.1 | Cross-Validation Experiment

To determine if the TAI and TE mechanisms could decrease negative transfer in MTL training contexts, a cross-validation procedure was executed. For every dataset mentioned in Subsection 4.1, the constituent sets were merged into one consolidated set. Subsequently, 5-fold cross-validation was applied to the STL, MTL, MTL-TAI, and MTL-TE models.

4.4.2 | Official Training-Test Split

For comparison against SOTA models (Magnossão de Paula et al. 2021; Plaza-del-Arco et al. 2021, 2021a, 2021b; Pérez and Luque 2019; Mahdaouy et al. 2021; Mostafa et al. 2022) relevant to the datasets, an experiment was run using the official train-test partitions provided with these datasets. HSArabic-2023 was the sole exception due to being a single partition; for this, we performed a stratified 80/20 split into training and test sets. Model training utilised the designated training set, or a combination of training and development sets if available. Post-training, model evaluation occurred on the test partitions.

For both experimental setups, only Spanish or Arabic data samples were employed. Models were assessed using the official metrics specific to each dataset (as described in Section 4.1). In practice, three evaluation metrics were required. Accuracy—used for EXIST-2021 and ArMI-2021—measures the proportion of correctly classified texts across both labels, which is appropriate for the moderately balanced sexism datasets. The DETOXIS-2021 and HSArabic-2023 benchmark specifies the F1-score for the Toxic class. We therefore compute precision and recall for that class and report their harmonic mean to reflect performance on the minority label. HatEval-2019 and OSACT-2022 rely on F1-macro, defined as the average of the per-class F1-scores, so that the score weights positive and negative classes equally, despite dataset imbalance. When aggregating results across tasks we keep the metric required by each dataset, ensuring that comparisons remain faithful to the official evaluation protocols.

All metrics are computed on the corresponding validation or test splits for every fold or official partition. During cross-validation, we calculate the metric on each fold before averaging them to obtain the reported values. For the training-test experiments, the metric is derived once on the held-out test portion. This consistent procedure allows us to attribute performance differences directly to the presence or absence of the TA mechanisms. We investigated MTL model versions combining two tasks and versions combining three tasks. The 95% confidence interval for results was computed via the formula:

$$\text{Margin of Error} = Z \times \sqrt{\frac{\text{value} \times (1 - \text{value})}{n_{\text{sample}}}}$$

where value denotes the obtained evaluation metric score (like accuracy or F1-score), n_{sample} signifies the test set size, and $Z = 1.96$ relates to the 95% confidence level. This calculation quantifies the uncertainty associated with the reported performance figures by supplying confidence intervals.

5 | Results and Analysis

This section reports the outcomes of the experiments and contrasts the performance of the models evaluated (detailed in Section 4). The results are displayed in two table formats. The experimental results (large tables) are organised into three parts: model type, model's task heads, and model's performance. The aggregated experimental results (small tables) are organised into two sections: model type and aggregated task heads' performance. Bold values indicate the highest values among all analysed models (column), while underlined values indicate the highest values among the MTL models. We also include bar charts showcasing the best result of each model for every dataset.

5.1 | Cross-Validation Experiment

5.1.1 | Spanish

The outcomes from the Spanish cross-validation experiment are presented in Table 8. Analysis of the Baseline models (outlined in Section 4.3) indicates the conventional MTL model experienced negative transfer in almost all scenarios. Relative to the STL model, the MTL model demonstrated improvements only in the Sexism detection task under two conditions: when trained jointly for Sexism and Hate-speech detection, and when trained across all three tasks. For every other configuration, the STL model yielded better performance. This suggests that negative transfer likely impeded the MTL model's learning process in those instances. Our findings suggest the TA mechanisms successfully reduced negative transfer. As detailed in Table 8, both the MTL-TAI model (using the TAI mechanism) and the MTL-TE model (using the TE mechanism) achieved consistently superior performance compared to the standard MTL model. Furthermore, the MTL-TAI and MTL-TE models surpassed the STL model's performance across the three assessed tasks. Between the two TA models, the MTL-TE generally performed better than the MTL-TAI model.

Figure 4 compares the top results of each model across the datasets in the Spanish cross-validation experiment. Negative transfer is evident in bar charts (b) DETOXIS-2021 and (c) HatEval-2019, where the STL model outperforms the classic MTL model. However, the MTL-TA models (MTL-TAI and MTL-TE) effectively mitigate this negative transfer and outperform the classic MTL model. Additionally, the MTL-TA models also surpass the classic MTL model in bar chart (a) EXIST-2021. This indicates that negative transfer may have influenced the learning process of the MTL model, but not to the extent that it performs worse than the STL model.

Table 9 presents the aggregated results of the MTL models for the Spanish cross-validation experiment. The classic MTL model performed poorly, achieving the lowest aggregated results across all task combinations. This outcome is likely attributable to the negative transfer effect hindering the model's learning capability. In contrast, the MTL-TAI model achieved the highest aggregated results for Toxic-language and Hate-speech detection. The MTL-TE model obtained the highest aggregated results for all other task combinations. Consistent with the results in Table 8, the TAI and TE mechanisms lessen the negative transfer effect during MTL training. As a result, the MTL-TAI and MTL-TE models outperform the traditional MTL model in all cases. These

TABLE 8 | Results of the Spanish cross-validation experiment with 95% confidence intervals.

		EXIST-2021	DETOXIS-2021	HatEval-2019
Model	**Task heads**	**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

Note: Evaluation metric values are shown with their 95% confidence intervals. Bold values indicate the highest scores across all analysed models, while underlined values denote the highest scores among the MTL models.

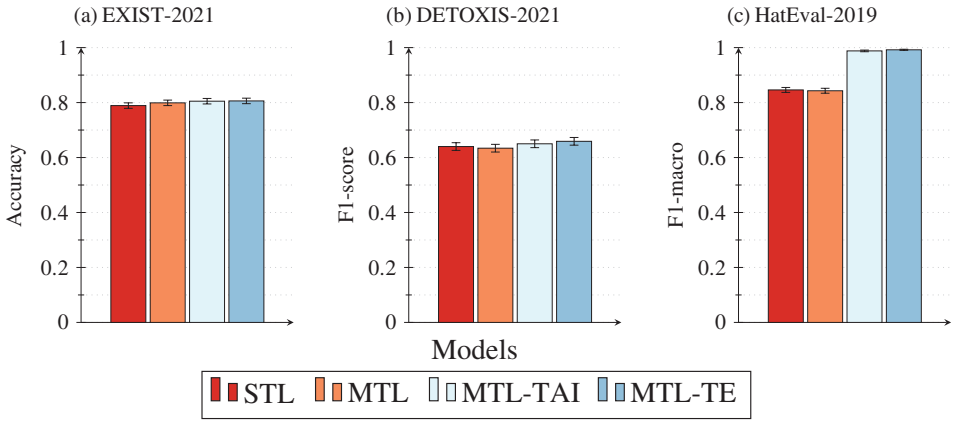

FIGURE 4 | Bar plots depicting the models' best performance based on the datasets' official evaluation metrics in the Spanish cross-validation experiment. The bars include 95% Confidence Intervals at the top. (a) Displays the models' best result for sexism detection on the EXIST-2021 dataset; (b) showcases the models' best result for toxic language detection on the DETOXIS-2021 dataset; (c) illustrates the models' best result for Hate Speech detection on the HatEval-2019 dataset.

TABLE 9 | Aggregated Spanish cross-validation results for the MTL models by model type. Bold values indicate the highest score across all analysed models within a column.

	Task Heads			
	Sexism	Sexism		Sexism
	Toxic-language		Toxic-language	Toxic-language
Models		Hate-speech	Hate-speech	Hate-speech
MTL	0.708	0.817	0.737	0.758
MTL-TAI	0.724	0.895	**0.819**	0.810
MTL-TE	**0.725**	**0.899**	0.817	**0.816**

results demonstrate the superiority of the MTL-TA approach over the traditional MTL model, owing to its ability to mitigate the negative transfer phenomenon.

5.1.2 | Arabic

Table 10 shows the results of the Arabic cross-validation experiment. Comparing the outcomes of the classic MTL and STL models, we observe that the MTL model performs worse due to the negative transfer phenomenon. The STL model outperforms the classic MTL model in Sexism and Offensive-language detection. The MTL-TA approach demonstrates superior performance compared to the traditional MTL approach in two of the three tasks. Among the MTL models, the MTL-TAI model achieves the best performance for Sexism detection, while the MTL-TE model obtains the highest F1-macro score for Offensive-language detection. The TA mechanisms effectively minimise negative transfer, leading to consistent improvements. For Hate-speech detection, the traditional MTL approach performs slightly better than the MTL-TA models. Overall, the results in Table 10 suggest that incorporating TAI and TE mechanisms in MTL models significantly enhances performance, reducing the adverse effects of negative transfer. The MTL-TE model, in particular, demonstrates the best overall performance across the evaluated tasks, making it the most effective model for the Arabic cross-validation experiment.

Figure 5 compares the top result of each model across the datasets in the Arabic cross-validation experiment. Negative transfer is present in bar charts (a) ArMI-2021 and (c) OSACT-2022, where the STL model outperforms the classic MTL model. In both cases, the MTL-TAI and MTL-TE models show slight improvement over the classic MTL model. This suggests that the task-aware capability helped to mitigate the negative transfer effect at least partially.

Table 11 presents the aggregated results of the MTL models for the Arabic cross-validation experiment. The classic MTL model performed poorly, achieving the lowest aggregated results in three out of the four task combinations. This is due to the negative transfer phenomenon that compromised MTL model's learning process. In contrast, the MTL-TAI model achieved the highest aggregated results for the Sexism, Toxic-language and Hate-speech task combination. The MTL-TE model obtained the highest aggregated results for the combinations of Sexism and Offensive-language tasks and Offensive-language and Hate-speech tasks. Consistent with the results in Table 10, the TAI and TE mechanisms mitigate the negative transfer effect during MTL training. As a result, the MTL-TAI and MTL-TE models outperform the traditional MTL model in three out of the four task combinations. These results demonstrate the superiority of the MTL-TA approach over the traditional MTL model, owing to its ability to mitigate the negative transfer phenomenon.

TABLE 10 | Results of the Arabic cross-validation experiment with 95% confidence intervals. Evaluation metric values are shown with their 95% confidence intervals.

		ArMI-2021	HSArabic-2023	OSACT-2022
Model	**Task Heads**	**Accuracy**	**F1-score**	**F1-macro**
STL	Sexism	**0.884** ± 0.006	—	—
	Offensive-language	—	0.608 ± 0.008	—
	Hate-speech	—	—	**0.778** ± 0.007
MTL	Sexism + offensive-language	0.874 ± 0.007	0.617 ± 0.008	—
	Sexism + hate-speech	0.876 ± 0.007	—	0.777 ± 0.007
	Offensive-language + hate-speech	—	0.603 ± 0.008	0.771 ± 0.007
	Offensive-language + hate-speech + sexism	0.878 ± 0.007	0.613 ± 0.008	0.770 ± 0.007
MTL-TAI	Sexism + offensive-language	0.883 ± 0.006	0.612 ± 0.008	—
	Sexism + hate-speech	0.880 ± 0.006	—	0.766 ± 0.007
	Offensive-language + hate-speech	—	0.603 ± 0.008	0.770 ± 0.007
	Offensive-language + hate-speech + sexism	0.883 ± 0.006	0.618 ± 0.008	0.776 ± 0.007
MTL-TE	Sexism + offensive-language	0.881 ± 0.006	**0.620** ± 0.008	—
	Sexism + hate-speech	0.878 ± 0.007	—	0.771 ± 0.007
	Offensive-language + Hate-speech	—	0.619 ± 0.008	0.772 ± 0.007
	Offensive-language + hate-speech + sexism	0.883 ± 0.006	0.618 ± 0.008	0.772 ± 0.007

Note: Bold values indicate the highest scores across all analysed models, while underlined values denote the highest scores among the MTL models.

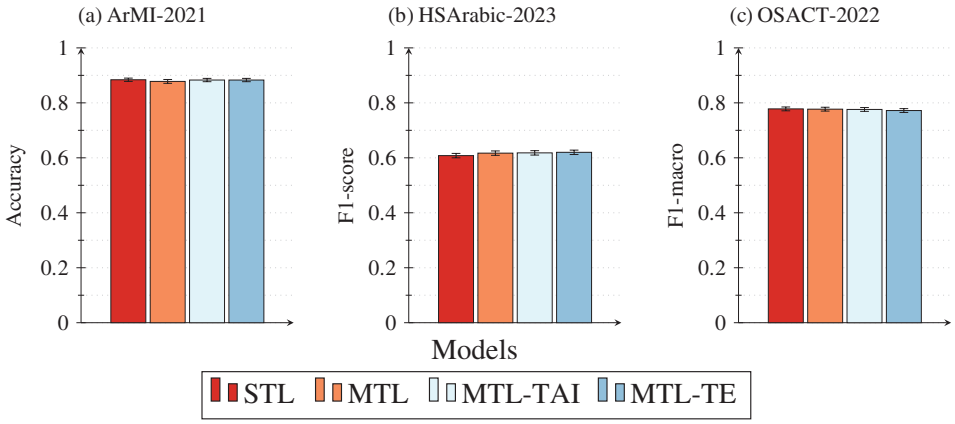

FIGURE 5 | Bar plots depicting the models' best performance based on the datasets' official evaluation metrics in the Arabic cross-validation experiment. The bars include 95% confidence intervals at the top. (a) Displays the models' best result for Sexism detection on the ArMI-2021 dataset; (b) showcases the models' best result for Toxic Language detection on the HSArabic-2023 dataset; (c) illustrates the models' best result for hate speech detection on the OSACT-2022 dataset.

5.2 | Official Training-Test Split

5.2.1 | Spanish

The experiment involving the three Spanish datasets using their official train-test partitions is detailed in Table 12. The results indicate that MTL training did not benefit the standard MTL model for the Sexism detection task, yielding lower accuracy than the STL model. This outcome is likely a consequence of the negative transfer phenomenon. Nevertheless, the MTL-TAI and MTL-TE models, incorporating TA mechanisms, counteracted the negative transfer observed during standard MTL training. They achieved superior accuracy compared to both the STL model and the SOTA model for EXIST-2021 (AI-UPV (Magnossão de Paula et al. 2021)). For Toxic-language detection, MTL training yielded better results than the STL baseline in the training-test setup. Broadly, the MTL, MTL-TAI, and MTL-TE models produced comparable outcomes, suggesting minimal negative transfer effects for this specific task during the standard MTL training process. Table 12 also demonstrates that MTL training enhanced performance for Hate-speech detection. The MTL model registered a higher F1-macro score than

TABLE 11 | Aggregated Arabic cross-validation results for the MTL models, shown by model type.

	Task Heads			
	Sexism	Sexism		Sexism
	Offensive-language		Offensive-language	Offensive-language
Models		Hate-speech	Hate-speech	Hate-speech
MTL	0.746	**0.826**	0.687	0.754
MTL-TAI	0.748	0.823	0.687	**0.759**
MTL-TE	**0.751**	0.825	**0.696**	0.758

Note: Bold values indicate the highest score across all analysed models within a column.

TABLE 12 | Results of the Spanish training-test experiment with 95% confidence intervals.

		EXIST-2021	DETOXIS-2021	HatEval-2019
Model	**Task heads**	**Accuracy**	**F1-score**	**F1-macro**
AI-UPV	—	0.790 ± 0.018	—	—
SINAI	—	—	**0.646** ± 0.031	—
Atalaya	—	—	—	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

Note: Evaluation metric values are shown with their 95% confidence intervals. Bold values indicate the highest scores across all analysed models, while underlined values denote the highest scores among the MTL models.

both the HatEval-2019 SOTA (Atalaya (Pérez and Luque 2019)) and the STL baseline. Models equipped with TA mechanisms further boosted these results, effectively lessening the negative transfer associated with conventional MTL training and yielding higher F1-macro scores than the standard MTL approach.

Figure 6 compares the top results of each model across the datasets in the Spanish training-test experiment. Bar chart (a) EXIST-2021 shows negative transfer, where the STL model outperforms the classic MTL model. However, the MTL-TA models (MTL-TAI and MTL-TE) effectively mitigate this negative transfer, outperforming the classic MTL model. In bar chart (c) HatEval-2019, the MTL-TA models also surpass the classic MTL model. This suggests negative transfer happened during training, even though the classic MTL's results were higher than the STL model's in this instance.

Table 13 displays the aggregated results of the MTL models for the Spanish official training-test split experiment. The traditional MTL model performed the worst, achieving the lowest aggregate results across all task combinations. This poor performance is attributed to the negative transfer phenomenon, which impaired the model's learning during training and resulted in subpar performance on the test. The MTL-TAI model achieved the best results for the task combinations of Sexism and Hate-speech, and Toxic-language and Hate-speech. Along with the

MTL-TE model, it also achieved top results for the combination of Sexism and Toxic-language. Additionally, the MTL-TE model achieved the highest results for the combination of all three tasks. The aggregated results demonstrate that the TAI and TE mechanisms alleviated the negative transfer phenomenon. In all cases, models equipped with these mechanisms outperformed the traditional MTL model, achieving superior aggregated results.

5.2.2 | Arabic

The study conducted on the three Arabic datasets using their specified training-test partitions is documented in Table 14. Observations from the table indicate that standard MTL training did not improve results for the Sexism detection task compared to STL. The peak accuracy achieved by the conventional MTL model matched that of the STL model. This lack of improvement is ascribed to the negative transfer effect constraining the MTL model's learning phase. Conversely, the MTL-TAI and MTL-TE models, utilising TA mechanisms, lessened the negative transfer observed in the standard MTL approach, yielding higher accuracy than both STL and standard MTL models. For the Sexism detection task, the SOTA model from ArMI-2021, UM6P-NLP (Mahdaouy et al. 2021), produced the best outcome. Regarding Offensive-language detection, MTL training showed enhanced results over the STL baseline within the training-test framework. Models incorporating TA mechanisms further amplified these gains, effectively reducing negative transfer from traditional MTL training and resulting in a better F1-score than the conventional MTL model. Notably, the MTL-TE model recorded the highest F1-score for Offensive-language detection, establishing a new SOTA for the HSArabic-2023 dataset. As seen in Table 14, MTL training also led to better outcomes for Hate-speech detection, with the standard MTL model scoring a higher F1-macro than the STL baseline. The MTL-TA models (MTL-TAI and MTL-TE) advanced these results further by mitigating the negative transfer effects present in standard MTL training, achieving superior F1-macro scores compared to the traditional MTL model. The top performance for Hate-speech detection was by the OSACT-2021 SOTA model, GOF (Mostafa et al. 2022), while among the MTL variants, the MTL-TE model achieved the best result for this task. Across all evaluated scenarios in the Arabic training-test split experiment, the MTL-TA models consistently delivered superior results relative to the classic MTL model.

Figure 7 compares the top results of each model across the datasets in the Arabic training-test experiment. Negative transfer is not clearly observed in any of the charts, as the classic MTL model consistently outperforms the STL model.

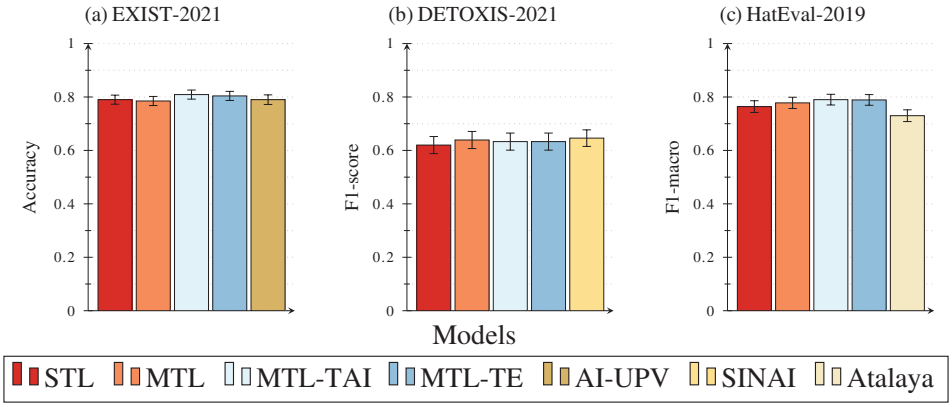

FIGURE 6 | Bar plots depicting the models' best performance based on the datasets' official evaluation metrics in the Spanish training-test experiment. The bars include 95% confidence intervals at the top. (a) Displays the models' best result for sexism detection on the EXIST-2021 dataset; (b) showcases the models' best result for Toxic language detection on the DETOXIS-2021 dataset; (c) illustrates the models' best result for hate speech detection on the HatEval-2019 dataset.

TABLE 13 | Aggregated Spanish training-test results for the MTL models by model type.

	Task Heads			
	Sexism	Sexism		Sexism
	Toxic-language		Toxic-language	Toxic-language
Models		Hate-speech	Hate-speech	Hate-speech
MTL	0.708	0.782	0.685	0.726
MTL-TAI	**0.715**	**0.799**	**0.709**	0.731
MTL-TE	**0.715**	0.795	0.702	**0.738**

Note: Bold values indicate the highest score across all analysed models within a column.

However, in all three charts, at least one of the MTL-TA models (MTL-TAI and MTL-TE) surpasses the classic MTL model, demonstrating their ability to mitigate negative transfer and improve performance.

Table 15 displays the aggregated results of the MTL models for the Arabic official training-test split experiment. The classic MTL model performed poorly, achieving the lowest aggregated results in three of the four task combinations. This likely stems from the negative transfer effect impeding the MTL model's learning progress. The MTL-TAI model obtained higher aggregated results than the MTL model in all cases except for the Sexism and Offensive-language task combination, where the difference was marginal. The MTL-TE model achieved the highest aggregated results for all task combinations. In line with the results from Table 14, the TAI and TE mechanisms reduce the

TABLE 14 | Results of the Arabic training-test experiment with 95% confidence intervals.

		ArMI-2021	HSArabic-2023	OSACT-2022
Model	**Task Heads**	**Accuracy**	**F1-score**	**F1-macro**
UM6P-NLP	—	**0.919** ±0.012	—	—
GOF	—	—	—	**0.852** ±0.014
STL	Sexism	0.892 ±0.014	—	—
	Offensive-language	—	0.605 ±0.017	—
	Hate-speech	—	—	0.775 ±0.016
MTL	Sexism + offensive-language	0.892 ±0.014	0.617 ±0.017	—
	Sexism + hate-speech	0.890 ±0.014	—	0.767 ±0.016
	Offensive-language + hate-speech	—	0.613 ±0.017	0.768 ±0.016
	Offensive-language + hate-speech + sexism	0.884 ±0.014	0.610 ±0.017	0.790 ±0.016
MTL-TAI	Sexism + offensive-language	0.888 ±0.014	0.617 ±0.017	—
	Sexism + hate-speech	<u>0.895</u> ±0.014	—	0.776 ±0.016
	Offensive-language + hate-speech	—	0.625 ±0.017	0.786 ±0.016
	Offensive-language + hate-speech + sexism	0.888 ±0.014	0.630 ±0.017	0.786 ±0.016
MTL-TE	Sexism + offensive-language	0.893 ±0.014	0.632 ±0.017	—
	Sexism + hate-speech	0.889 ±0.014	—	0.790 ±0.016
	Offensive-language + hate-speech	—	0.634 ±0.017	0.786 ±0.016
	Offensive-language + hate-speech + sexism	0.894 ±0.014	**<u>0.635</u>** ±0.017	<u>0.794</u> ±0.016

Note: Evaluation metric values are shown with their 95% confidence intervals. Bold values indicate the highest scores across all analysed models, while underlined values denote the highest scores among the MTL models.

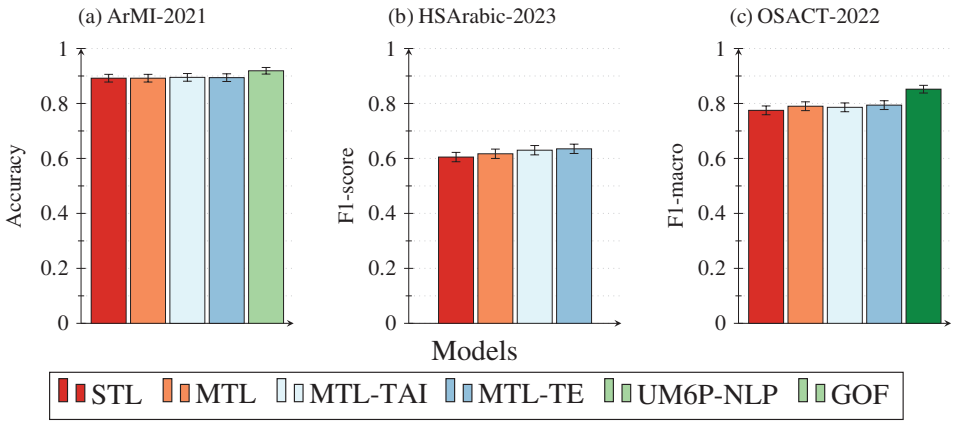

FIGURE 7 | Bar plots depicting the models' best performance based on the datasets' official evaluation metrics in the Arabic training-test experiment. The bars include 95% confidence intervals at the top. (a) Displays the models' best result for sexism detection on the ArMI-2021 dataset; (b) showcases the models' best result for toxic language detection on the HSArabic-2023 dataset; (c) illustrates the models' best result for hate speech detection on the OSACT-2022 dataset.

negative transfer impact during MTL training. As a result, the MTL-TAI and MTL-TE models outperform the traditional MTL model in all cases.

6 | Discussion and Limitations

6.1 | Discussion

The results from our experiments demonstrate the impact of negative transfer on the performance of traditional MTL models. Across both the Spanish and Arabic datasets, the classic MTL models consistently underperformed compared to STL models, particularly in tasks prone to negative transfer phenomena such as Sexism and Offensive Language detection. However, the introduction of TA mechanisms (TAI and TE) significantly mitigated the effects of negative transfer. Both MTL-TAI and MTL-TE models showed improvements over the traditional MTL models, achieving higher accuracy and F1 scores in almost all task combinations. In the Spanish cross-validation experiment, the MTL-TE model outperformed all others for the three evaluated tasks. The MTL-TAI model excelled for Sexism and Hate-speech detection in the Spanish official training-test split and the MTL-TE model became the new DETOXIS-2021 SOTA model for Toxic-language detection. These findings were consistent across the aggregated results evaluation for cross-validation and official training-test split, where the MTL-TA models continued to achieve superior results than the classic MTL model.

These gains align with the linguistic overlap among the Spanish datasets: sexism, toxic language, and hate speech corpora share recurring lexical markers (e.g., slurs and gendered stereotypes) but differ in their annotation scope. By conditioning the encoder on the task through TAI, the model learns to differentiate when a term should be interpreted as toxic rhetoric versus explicitly sexist and hate speech. TE further exploits this structure by amplifying shared signals—such as intensity modifiers and target mentions—only for the tasks that benefit from them. As a result, the TA-equipped models can capitalise on beneficial cross-task cues while suppressing misleading correlations that caused the baseline MTL model to underperform. Similarly, for the Arabic cross-validation experiment, the MTL-TE model outperformed all others in Offensive Language and obtained competitive results for Sexism and Hate-speech detection. In the official training-test split experiment, the MTL-TAI obtained the top result for Offensive-language detection and became the new HSArabic-2023 SOTA model for the task. These results were consistent throughout the aggregated evaluations for cross-validation and the official training-test split, where the MTL-TA models consistently outperformed the traditional MTL model.

Arabic datasets exhibit stronger dialectal variation and class imbalance than their Spanish counterparts, which amplifies negative transfer when models rely solely on shared representations. Injecting task context via TAI helps the encoder focus on morphological patterns that are discriminative for each phenomenon (e.g., misogynistic verb forms versus generic insults), while TE recalibrates the representation to handle skewed label distributions by emphasising features that consistently characterise the minority class. This explains why the largest relative improvements arise on HSArabic-2023, where offensive content is substantially rarer than neutral statements. The analysis of the bar charts presenting the best results for each model across all datasets reveals a consistent pattern: when negative transfer is identified—indicated by the classic MTL model underperforming the STL model—the MTL-TA model effectively mitigates this issue, outperforming both the classic MTL and STL models. Furthermore, the MTL-TA model often achieves superior performance compared to the classic MTL model, even in scenarios where negative transfer is not evident, as demonstrated by the comparative analysis of the classic MTL and STL models' results. In summary, the incorporation of TAI and TE mechanisms in MTL models not only provides a robust solution to the negative transfer problem but also opens up new possibilities for enhancing overall performance. The MTL-TAI and MTL-TE models emerge as the most effective MTL models across the evaluated tasks, demonstrating the exciting potential for performance improvement in MTL scenarios.

Across both languages, a common trend is that TA mitigates situations where dataset-specific annotation guidelines diverge. By explicitly signalling the task objective, the model can maintain separate decision boundaries for nuanced categories while still sharing underlying lexical knowledge. This supports our hypothesis that negative transfer stems from conflating task intent rather than from a lack of shared information. Consequently, TA delivers the most benefit when tasks are semantically related yet operationalised differently—a regime that typifies harmful language detection benchmarks.

TABLE 15 | Aggregated Arabic training-test results for the MTL models by model type.

	Task heads			
	Sexism		Sexism	
Models	Offensive-language	Hate-speech	Offensive-language	Hate-speech
MTL	0.755	0.828	0.691	0.761
MTL-TAI	0.752	0.835	0.705	0.768
MTL-TE	**0.763**	**0.839**	**0.710**	**0.774**

Note: Bold values indicate the highest score across all analysed models within a column.

6.2 | Limitations

Despite the promising results demonstrated by the incorporation of TAI and TE mechanisms in MTL models, several limitations should be acknowledged. Firstly, the negative transfer phenomenon remains a challenge, especially for certain task combinations. While the TA mechanisms effectively mitigate this issue, the extent of their effectiveness is not fully known. We are still unable to determine whether the TA mechanisms entirely eliminate negative transfer, and traditional MTL models continue to suffer from its effects, particularly in tasks such as Sexism and Offensive Language detection. The evaluation is limited to specific datasets and tasks within the Spanish and Arabic languages. This scope may not fully capture the generalisability

of the findings across different languages, datasets, or task domains. Future work should explore a broader range of datasets and tasks to validate the robustness of the proposed mechanisms. Additionally, achieving strong performance with the two MTL-TA models hinges on employing a potent encoder. This dependence could pose difficulties for computational systems with limited resources that lack the capacity for deep learning frameworks like Transformers (Vaswani et al. 2017) as the encoder. The inclusion of supplementary layers and mechanisms might also escalate computational requirements, potentially restricting the scalability of these models for practical, real-world uses. A detailed examination of the balance between performance gains and computational overhead is warranted. Handling an increased number of tasks necessitates more task heads, consequently enlarging the model's parameter count. As a result, fine-tuning MTL-TA models demands greater computational resources. This escalation in resource needs could present a considerable barrier for applications operating under tight computational constraints. A further consideration is whether the fine-tuning process, which uses task-specific information, diminishes the MTL-TA models' capacity for adapting to novel, previously unseen tasks (e.g., in few-shot learning or instruction-following scenarios). The specialisation towards specific tasks during fine-tuning could potentially impede their adaptability and generalisation capabilities for new tasks, an area deserving investigation in future studies.

7 | Conclusion and Future Work

This paper introduced the TA strategy aimed at tackling the negative transfer issue (Wu, Zhang, and Ré 2020; Vandenhende et al. 2022; Li et al. 2023) encountered during MTL training phases. Our approch presented two distinct mechanisms: TAI and TE. The TAI mechanism enhances the MTL model encoder's input by integrating task description details. Concurrently, the TE method adds a TEB, an extra module processing the encoder's latent output alongside a Task Identification Vector (TIV). Through these mechanisms, the MTL model can generate task-tailored representations, which effectively reduce negative transfer effects and boost overall model performance.

Our experimental results demonstrate that the TA mechanisms significantly reduce negative transfer and improve performance over standard MTL models across different tasks. Notably, we achieved competitive results compared to SOTA methods for both the Spanish and Arabic datasets. The proposed models set new SOTA benchmarks on the EXIST-2021 (Rodríguez-Sánchez et al. 2021) and HatEval-2019 (Basile et al. 2019) datasets for Spanish, as well as on the HSArabic-2023 dataset for Arabic. These findings underscore the generalisability and effectiveness of the TA approach in mitigating negative transfer across different languages and tasks.

Beyond the immediate results, the broader impact of our approach lies in its potential to shape future research in NLP. By enabling more accurate and efficient MTL systems, the introduction of TA mechanisms paves the way for improved detection and moderation of harmful content on social media platforms. This has significant implications for the development of automated systems tasked with moderating online spaces, reducing human bias, and fostering safer digital environments.

Furthermore, the ability of TA-equipped models to enhance performance across multiple languages and tasks suggests broader applicability in multilingual and cross-linguistic NLP challenges. This opens the door for future research to explore TA in other languages and domains, where traditional single-task models often struggle due to data scarcity and computational constraints.

For future work, it would be valuable to further investigate the minimum amount of labelled data or information volume required for MTL to outperform STL models. Additionally, exploring the augmentation of MTL models with low-level task supervision, where the decoder leverages the entirety or a portion of the encoder's hidden states, could provide further performance gains. We also plan to extend the application of MTL combined with TA into novel areas, such as identifying Sexism within learning-with-disagreement paradigms (Uma et al. 2021; Plaza et al. 2024; Plaza, de Carrillo-Albornoz, Morante, Amigó, et al. 2023; Plaza, de Carrillo-Albornoz, Morante, Gonzalo, et al. 2023), where multiple annotator labels are considered rather than relying on a single aggregated gold label (Frenda et al. 2025).

Finally, future research will also focus on incorporating unsupervised learning techniques to enhance the proposed models for identifying Hate Speech, Toxic Language, and Sexism. Potential techniques include Latent Dirichlet Allocation (Blei et al. 2003), Self-Organising Maps (Miljković 2017), and K-Means Clustering (Ezugwu et al. 2022), which could offer further improvements in model robustness and accuracy across different linguistic and cultural contexts.

Author Contributions

Angel Felipe Magnossão de Paula: conceptualization, data curation, formal analysis, investigation, methodology, resources, software, validation, writing – original draft. **Imene Bensalem:** data curation, resources, supervision, writing – original draft. **Damiano Spina:** conceptualization, methodology, resources, supervision, funding acquisition, writingen – original draft. **Paolo Rosso:** conceptualization, methodology, resources, supervision, funding acquisition.

Acknowledgements

TThis research is partially supported by the Australian Research Council (ARC) Centre of Excellence for Automated Decision-Making and Society (ADM+S, CE200100005). The research work of Paolo Rosso was in the framework of the Malicious Actors Profiling and Detection in Online Social Networks Through Artificial Intelligence (MARTINI) project(Grant PCI2022-135008-2) funded by MCIN/AEI/ 10.13039/501100011033 and by European Union NextGenerationEU/PRTR.

Funding

This work was supported by MCIN/AEI/10.13039/501100011033 (PCI2022-135008-2); European Union NextGenerationEU/PRTR.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are available on request from the corresponding sources. The data are not publicly available due to privacy or ethical restrictions.

Endnotes

[1] https://github.com/AngelFelipeMP/Arabic-MultiTask-Learning.

References

Abburi, H., P. Parikh, N. Chhaya, and V. Varma. 2020. "Semi-Supervised Multi-Task Learning for Multi-Label Fine-Grained Sexism Classification." In *Proceedings of the 28th International Conference on Computational Linguistics*, 5810–5820. International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.511.

Abdelali, A., S. Hassan, H. Mubarak, K. Darwish, and Y. Samih. 2021. "Pre-Training BERT on Arabic Tweets: Practical Considerations." arXiv:2102.10684.

Abdul-Mageed, M., and A. Elmadany. 2021. "ARBERT & MARBERT: Deep Bidirectional Transformers for Arabic." In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* (Volume 1: Long Papers), 7088–7105.

Abu Farha, I., and W. Magdy. 2020. "Multitask Learning for Arabic Offensive Language and Hate-Speech Detection." In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, With a Shared Task on Offensive Language Detection*, edited by H. Al-Khalifa, W. Magdy, K. Darwish, T. Elsayed, and H. Mubarak, 86–90. European Language Resource Association. https://aclanthology.org/2020.osact-1.14.

Aguilar, G., S. Maharjan, A. P. López-Monroy, and T. Solorio. 2017. "A Multi-Task Approach for Named Entity Recognition in Social Media Data." In *Proceedings of the 3rd Workshop on Noisy User-Generated Text*, edited by L. Derczynski, W. Xu, A. Ritter, and T. Baldwin, 148–153. Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-4419.

Alkomah, F., and X. Ma. 2022. "A Literature Review of Textual Hate Speech Detection Methods and Datasets." *Information* 13: 273. https://doi.org/10.3390/info13060273.

Alshaabi, T., D. R. Dewhurst, J. R. Minot, et al. 2021. "The Growing Amplification of Social Media: Measuring Temporal and Social Contagion Dynamics for Over 150 Languages on Twitter for 2009–2020." *EPJ Data Science* 10: 15. https://doi.org/10.1140/epjds/s13688-021-00271-0.

Antoun, W., F. Baly, and H. Hajj. 2020. "AraBERT: Transformer-Based Model for Arabic Language Understanding." In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, With a Shared Task on Offensive Language Detection*, 9–15. European Language Resource Association. https://aclanthology.org/2020.osact-1.2.

Basile, V., C. Bosco, E. Fersini, et al. 2019. "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*, 54–63. Association for Computational Linguistics. https://doi.org/10.18653/v1/S19-2007.

Bensalem, I., P. Rosso, and H. Zitouni. 2024. "Toxic Language Detection: A Systematic Review of Arabic Datasets." *Expert Systems* 41: e13551. https://doi.org/10.1111/exsy.13551.

Blei, D. M., A. Y. Ng, and M. I. Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3: 993–1022. https://doi.org/10.5555/944919.944937.

Blizard, W. D. 1988. "Multiset Theory." *Notre Dame Journal of Formal Logic* 30: 36–66. https://doi.org/10.1305/ndjfl/1093634995.

Bojanowski, P., E. Grave, A. Joulin, and T. Mikolov. 2017. "Enriching Word Vectors With Subword Information." *Transactions of the Association for Computational Linguistics* 5: 135–146. https://doi.org/10.1162/tacl_a_00051.

Boser, B. E., I. M. Guyon, and V. N. Vapnik. 1992. "A Training Algorithm for Optimal Margin Classifiers." In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory COLT '92*, 144–152. Association for Computing Machinery. https://doi.org/10.1145/130385.130401.

Canete, J., G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, and J. Pérez. 2020. "Spanish Pre-trained Bert Model and Evaluation Data." In *Practical Machine Learning for Developing Countries (PML4DC) at Eleventh International Conference on Learning Representations (ICLR)*, 2020, 1–10. https://pml4dc.github.io/iclr2020/papers/PML4DC2020_10.pdf.

Caruana, R., S. Lawrence, and L. Giles. 2000. "Overfitting in Neural Nets: Backpropagation, Conjugate Gradient, and Early Stopping." In *Proceedings of the 13th International Conference on Neural Information Processing Systems NIPS'00*, 381–387. MIT Press. https://doi.org/10.5555/3008751.3008807.

Charfi, A., M. Besghaier, R. Akasheh, A. Atalla, and W. Zaghouani. 2024. "Hate Speech Detection With ADHAR: A Multi-Dialectal Hate Speech Corpus in Arabic." *Frontiers in Artificial Intelligence* 7: 1391472. https://doi.org/10.3389/frai.2024.1391472.

Chen, S., Y. Zhang, and Q. Yang. 2024. "Multi-Task Learning in Natural Language Processing: An Overview." *ACM Computing Surveys* 56: 1–32. https://doi.org/10.1145/3663363.

Chen, Z., V. Badrinarayanan, C.-Y. Lee, and A. Rabinovich. 2018. "GradNorm: Gradient Normalization for Adaptive Loss Balancing in Deep Multitask Networks." In *Proceedings of the 35th International Conference on Machine Learning* (794–803). PMLR volume 80 of *Proceedings of Machine Learning Research*. http://proceedings.mlr.press/v80/chen18a/chen18a.pdf.

Cipolla, R., Y. Gal, and A. Kendall. 2018. "Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics." In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA*, 7482–7491. https://doi.org/10.1109/CVPR.2018.00781.

De Freitas, J. M., S. Berg, B. C. Geiger, and M. Mucke. 2022. "Compressed Hierarchical Representations for Multi-Task Learning and Task Clustering." In *2022 International Joint Conference on Neural Networks (IJCNN)*, 1–8. IEEE. https://doi.org/10.1109/ijcnn55064.2022.9892342.

Derczynski, L., M. Guerini, D. Nozza, F. M. del Plaza-Arco, J. Sorensen, and M. Zampieri. 2024. "Countering Hateful and Offensive Speech Online—Open Challenges." In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, edited by J. Li and F. Liu, 11–16. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-tutorials.2.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2019. "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*, 4171–4186. https://doi.org/10.18653/v1/N19-1423.

Duan, L., D. Xu, and I. W. Tsang. 2012. "Learning With Augmented Features for Heterogeneous Domain Adaptation." In *Proceedings of the 29th International Coference on International Conference on Machine Learning ICML'12*, 667–674. Omnipress. https://doi.org/10.5555/3042573.3042661.

Ezugwu, A. E., A. M. Ikotun, O. O. Oyelade, et al. 2022. "A Comprehensive Survey of Clustering Algorithms: State-of-the-Art Machine Learning Applications, Taxonomy, Challenges, and Future Research Prospects." *Engineering Applications of Artificial Intelligence* 110: 104743. https://doi.org/10.1016/j.engappai.2022.104743.

Fang, L., G. Liu, and R. Zhang. 2022. "Sense-aware BERT and Multi-task Fine-tuning for Multimodal Sentiment Analysis." In *2022 International Joint Conference on Neural Networks (JCNN)*, 1–8. https://doi.org/10.1109/IJCNN55064.2022.9892116.

Frenda, S., G. Abercrombie, V. Basile, et al. 2025. "Perspectivist Approaches to Natural Language Processing: A Survey." *Language Resources and Evaluation* 59: 1719–1746. https://doi.org/10.1007/s10579-024-09766-4.

Frenda, S., B. Ghanem, M. M. y Gómez, and P. Rosso. 2019. "Online Hate Speech Against Women: Automatic Identification of Misogyny and Sexism on Twitter." *Journal of Intelligent & Fuzzy Systems* 36: 4743–4752. https://doi.org/10.3233/JIFS-179023.

Guo, M., A. Haque, D.-A. Huang, S. Yeung, and L. Fei-Fei. 2018. "Dynamic Task Prioritization for Multitask Learning." In *Computer Vision—ECCV 2018: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part XVI*, 282–299. Springer-Verlag. https://doi.org/10.1007/978-3-030-01270-0_17.

Indurthi, S., M. A. Zaidi, N. Kumar Lakumarapu, et al. 2021. "Task Aware Multi-Task Learning for Speech to Text Tasks." In *ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7723–7727. IEEE. https://doi.org/10.1109/ICASSP39728.2021.9414703.

Jamil, S., M. Jalil Piran, and O.-J. Kwon. 2023. "A Comprehensive Survey of Transformers for Computer Vision." *Drones* 7: 287. https://doi.org/10.3390/drones7050287.

Knight, P., and R. Duan. 2023. "Multi-Task Learning With Summary Statistics." In *Proceedings of the 37th International Conference on Neural Information Processing Systems NIPS '23*, vol. 36, 54020–54031. Curran Associates Inc. https://doi.org/10.5555/3666122.3668472.

Kulis, B., K. Saenko, and T. Darrell. 2011. "What You Saw Is Not What You Get: Domain Adaptation Using Asymmetric Kernel Transforms." In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition CVPR '11*, 1785–1792. IEEE Computer Society. https://doi.org/10.1109/CVPR.2011.5995702.

Lauriola, I., A. Lavelli, and F. Aiolli. 2022. "An Introduction to Deep Learning in Natural Language Processing: Models, Techniques, and Tools." *Neurocomputing* 470: 443–456. https://doi.org/10.1016/j.neucom.2021.05.103.

Li, D., H. Nguyen, and H. R. Zhang. 2023. "Identification of Negative Transfers in Multitask Learning Using Surrogate Models." Transactions on Machine Learning Research. https://openreview.net/forum?id=KgfFAI9f3E.

Liu, Y., Y. Lu, H. Liu, et al. 2023. "Hierarchical Prompt Learning for Multi-Task Learning." In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10888–10898. https://doi.org/10.1109/CVPR52729.2023.01048.

Long, M., Z. Cao, J. Wang, and P. S. Yu. 2017. "Learning Multiple Tasks with Multilinear Relationship Networks." In *Proceedings of the 31st International Conference on Neural Information Processing Systems NIPS'17*, 1593–1602. Curran Associates Inc. https://doi.org/10.5555/3294771.3294923.

Loshchilov, I., and F. Hutter. 2019. "Decoupled Weight Decay Regularization." In *7th International Conference on Learning Representations, ICLR*, New Orleans, LA, USA. https://openreview.net/pdf?id=Bkg6RiCqY7.

Lu, Y., A. Kumar, S. Zhai, Y. Cheng, T. Javidi, and R. Feris. 2017. "Fully-Adaptive Feature Sharing in Multi-Task Networks With Applications in Person Attribute Classification." In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1131–1140. IEEE. https://doi.org/10.1109/CVPR.2017.126.

Magnossão de Paula, A. F., R. F. da Silva, and I. B. Schlicht. 2021. "Sexism Prediction in Spanish and English Tweets Using Monolingual and Multilingual BERT and Ensemble Models." In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021) Co-Located With the XXXVII International Conference of the Spanish Society for Natural Language Processing (SEPLN 2021)*, 356–373. CEUR. https://ceur-ws.org/Vol-2943/exist_paper2.pdf.

Magnossão de Paula, A. F., P. Rosso, and D. Spina. 2023. "Mitigating Negative Transfer With Task Awareness for Sexism, Hate Speech, and Toxic Language Detection." In *2023 International Joint Conference on Neural Networks (IJCNN)*, 1–8. https://doi.org/10.1109/IJCNN54540.2023.10191347.

Magnossão de Paula, A. F., and I. B. Schlicht. 2021. "AI-UPV at IberLEF-2021 DETOXIS Task: Toxicity Detection in Immigration-Related Web News Comments Using Transformers and Statistical Models." In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021) Co-Located With the XXXVII International Conference of the Spanish Society for Natural Language Processing (SEPLN 2021)*, 547–566. CEUR. https://ceur-ws.org/Vol-2943/detoxis_paper2.pdf.

Mahdaouy, A. E., A. E. Mekki, A. Oumar, H. Mousannif, and I. Berrada. 2021. "Deep Multi-Task Models for Misogyny Identification and Categorization on Arabic Social Media." In *Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, Gandhinagar, India*, 852–860. CEUR-WS.org Volume 3159 of CEUR Workshop Proceedings. https://ceur-ws.org/Vol-3159/T5-5.pdf.

Mikolov, T., K. Chen, G. Corrado, and J. Dean. 2013. "Efficient Estimation of Word Representations in Vector Space." In *1st International Conference on Learning Representations*, ICLR 2013, Scottsdale, Arizona, USA, May 2–4, 2013, Workshop Track Proceedings. https://people.fjfi.cvut.cz/vybirja2/Seminar/word2vec_1301.3781.pdf.

Miljković, D. 2017. "Brief Review of Self-Organizing Maps." In *2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 1061–1066. https://doi.org/10.23919/MIPRO.2017.7973581.

Mostafa, A., O. Mohamed, and A. Ashraf. 2022. "GOF at Arabic Hate Speech 2022: Breaking the Loss Function Convention for Data-Imbalanced Arabic Offensive Text Detection." In *Proceedings of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools With Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection*, 167–175. European Language Resources Association. https://aclanthology.org/2022.osact-1.21.

Mubarak, H., H. Al-Khalifa, and A. Al-Thubaity. 2022. "Overview of OSACT5 Shared Task on Arabic Offensive Language and Hate Speech Detection." In *Proceedings of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools With Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection (Pp. 162–166)*. European Language Resources Association. https://aclanthology.org/2022.osact-1.20.

Mubarak, H., K. Darwish, W. Magdy, T. Elsayed, and H. Al-Khalifa. 2020. "Overview of OSACT4 Arabic Offensive Language Detection Shared Task." In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, With a Shared Task on Offensive Language Detection*, edited by H. Al-Khalifa, W. Magdy, K. Darwish, T. Elsayed, and H. Mubarak, 48–52. European Language Resource Association. https://aclanthology.org/2020.osact-1.7.

Mulki, H., and B. Ghanem. 2021. "ArMI at FIRE 2021: Overview of the First Shared Task on Arabic Misogyny Identification." In *Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation*, Gandhinagar, India, December 13–17, 2021, 820–830. CEUR-WS.org volume 3159 of CEUR Workshop Proceedings. https://ceur-ws.org/Vol-3159/T5-1.pdf.

Otter, D. W., J. R. Medina, and J. K. Kalita. 2020. "A Survey of the Usages of Deep Learning for Natural Language Processing." *IEEE Transactions on Neural Networks and Learning Systems* 32: 604–624. https://doi.org/10.1109/TNNLS.2020.2979670.

Pachinger, P., A. Hanbury, J. Neidhardt, and A. Planitzer. 2023. "Toward Disambiguating the Definitions of Abusive, Offensive, Toxic, and Uncivil Comments." In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, 107–113. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.c3nlp-1.11.

Pan, S. J., and Q. Yang. 2009. "A Survey on Transfer Learning." *IEEE Transactions on Knowledge and Data Engineering* 22: 1345–1359. https://doi.org/10.1109/TKDE.2009.191.

Pérez, J. M., and F. M. Luque. 2019. "Atalaya at SemEval 2019 Task 5: Robust Embeddings for Tweet Classification." In *In Proceedings of the 13th International Workshop on Semantic Evaluation*, 64–69. Association for Computational Linguistics. https://doi.org/10.18653/v1/S19-2008.

Plaza, L., J. de Carrillo-Albornoz, R. Morante, et al. 2023. "Overview of EXIST 2023—Learning With Disagreement for Sexism Identification and Characterization." In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, edited by A. Arampatzis, E. Kanoulas, T. Tsikrika, et al., 316–342. Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-42448-9_23.

Plaza, L., J. de Carrillo-Albornoz, R. Morante, et al. 2023. "Overview of EXIST 2023: sEXism Identification in Social neTworks." In *Proceedings of ECIR'23*, 593–599. Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-28241-6_68.

Plaza, L., J. de Carrillo-Albornoz, V. Ruiz, et al. 2024. "Overview of EXIST 2024—Learning With Disagreement for Sexism Identification and Characterization in Tweets and Memes." In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, 93–117. Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-71908-0_5.

Plaza-del-Arco, F. M., M. D. Molina-González, and L. Alfonso. 2021. "SINAI at IberLEF-2021 DETOXIS Task: Exploring Features as Tasks in a Multi-Task Learning Approach to Detecting Toxic Comments." In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021) Co-Located With the XXXVII International Conference of the Spanish Society for Natural Language Processing (SEPLN 2021)*, 580–590. Málaga, Spain. https://ceur-ws.org/Vol-2943/detoxis_paper5.pdf.

Plaza-del-Arco, F. M., M. D. Molina-González, L. A. Ureña-López, and M. T. Martín-Valdivia. 2021a. "A Multi-Task Learning Approach to Hate Speech Detection Leveraging Sentiment Analysis." *IEEE Access* 9: 112478–112489. https://doi.org/10.1109/ACCESS.2021.3103697.

Plaza-del-Arco, F. M., M. D. Molina-González, L. A. Ureña-López, and M. T. Martín-Valdivia. 2021b. "Comparing Pre-Trained Language Models for Spanish Hate Speech Detection." *Expert Systems with Applications* 166: 114120. https://doi.org/10.1016/j.eswa.2020.114120.

Poletto, F., V. Basile, M. Sanguinetti, C. Bosco, and V. Patti. 2021. "Resources and Benchmark Corpora for Hate Speech Detection: A Systematic Review." *Language Resources and Evaluation* 55: 477–523. https://doi.org/10.1007/s10579-020-09502-8.

Prettenhofer, P., and B. Stein. 2010. "Cross-Language Text Classification Using Structural Correspondence Learning." In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 1118–1127. Association for Computational Linguistics. https://aclanthology.org/P10-1114.

Rodríguez-Sánchez, F., J. de Carrillo-Albornoz, L. Plaza, et al. 2021. "Overview of EXIST 2021: sEXism Identification in Social neTworks." *Procesamiento del Lenguaje Natural* 67: 195–207. https://doi.org/10.26342/2021-67-17.

Ruder, S. 2017. "An Overview of Multi-Task Learning in Deep Neural Networks." CoRR, abs/1706.05098. http://arxiv.org/abs/1706.05098.

Ruder, S., M. E. Peters, S. Swayamdipta, and T. Wolf. 2019. "Transfer Learning in Natural Language Processing." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, 15–18. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-5004.

Sener, O., and V. Koltun. 2018. "Multi-Task Learning as Multi-Objective Optimization." In *Proceedings of the 32nd International Conference on Neural Information Processing Systems NIPS'18*, 525–536. Curran Associates Inc. https://doi.org/10.5555/3326943.3326992.

Sinha, A. T., A. Rabinovich, Z. Chen, and V. Badrinarayanan. 2021. "Gradient Adversarial Training of Neural Networks." US Patent App. 17/051, 982.

Taulé, M., A. Ariza, M. Nofre, E. Amigó, and P. Rosso. 2021. "Overview of DETOXIS at IberLEF 2021: DEtection of TOxicity in Comments in Spanish." *Procesamiento del Lenguaje Natural* 67: 209–221. https://doi.org/10.26342/2021-67-18.

Uma, A., T. Fornaciari, A. Dumitrache, et al. 2021. "SemEval-2021 Task 12: Learning With Disagreements." In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, edited by A. Palmer, N. Schneider, N. Schluter, G. Emerson, A. Herbelot, and X. Zhu, 338–347. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.semeval-1.41.

Vandenhende, S., S. Georgoulis, L. V. Gool, and B. D. Brabandere. 2020. "Branched Multi-Task Networks: Deciding What Layers to Share." In *Proceedings of the 31st British Machine Vision Conference BMVC '20*. BMVA Press. https://www.bmvc2020-conference.com/assets/papers/0213.pdf.

Vandenhende, S., S. Georgoulis, W. Van Gansbeke, M. Proesmans, D. Dai, and L. Van Gool. 2022. "Multi-Task Learning for Dense Prediction Tasks: A Survey." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44: 3614–3633. https://doi.org/10.1109/TPAMI.2021.3054719.

Vaswani, A., N. Shazeer, N. Parmar, et al. 2017. "Attention Is All You Need." In *Proceedings of the 31st International Conference on Neural Information Processing Systems NIPS'17*, 6000–6010. Curran Associates Inc. https://doi.org/10.5555/3295222.3295349.

Voulodimos, A., N. Doulamis, A. Doulamis, E. Protopapadakis, and D. Andina. 2018. "Deep Learning for Computer Vision: A Brief Review." *Computational Intelligence and Neuroscience* 2018: 1–13. https://doi.org/10.1155/2018/7068349.

Wang, C., and S. Mahadevan. 2011. "Heterogeneous Domain Adaptation Using Manifold Alignment." In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*, 1541–1546. AAAI Press. https://doi.org/10.5555/2283516.2283652.

Wang, Y., M. Xu, Y. Yan, T. Zhao, Y. Chen, and J. Yang. 2022. "Exploring Topic Supervision With BERT for Text Matching." In *2022 International Joint Conference on Neural Networks (IJCNN)*, 1–7. https://doi.org/10.1109/IJCNN55064.2022.9892023.

Weiss, K., T. M. Khoshgoftaar, and D. Wang. 2016. "A Survey of Transfer Learning." *Journal of Big Data* 3: 1–40. https://doi.org/10.1186/s40537-016-0043-6.

Worsham, J., and J. Kalita. 2020. "Multi-Task Learning for Natural Language Processing in the 2020s: Where Are We Going?" *Pattern Recognition Letters* 136: 120–126. https://doi.org/10.1016/j.patrec.2020.05.031.

Wu, S., H. Fei, and D. Ji. 2020. "Aggressive Language Detection With Joint Text Normalization via Adversarial Multi-Task Learning." In *Natural Language Processing and Chinese Computing: 9th CCF International Conference, NLPCC 2020, Zhengzhou, China*, 683–696. Springer-Verlag. https://doi.org/10.1007/978-3-030-60450-9_54.

Wu, S., H. R. Zhang, and C. Ré. 2020. "Understanding and Improving Information Transfer in Multi-Task Learning." In *8th International Conference on Learning Representations, ICLR 2020*, Addis Ababa, Ethiopia. OpenReview.net. https://openreview.net/forum?id=SylzhkBtDB.

Xu, D., W. Ouyang, X. Wang, and N. Sebe. 2018. "Pad-net: Multi-Asks Guided Prediction and Distillation Network for Simultaneous Depth Estimation and Scene Parsing." In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 675–684. https://doi.org/10.1109/CVPR.2018.00077.

Zhang, Y., and Q. Yang. 2022. "A Survey on Multi-Task Learning." *IEEE Transactions on Knowledge and Data Engineering* 34: 5586–5609. https://doi.org/10.1109/TKDE.2021.3070203.

Zhang, Z., Z. Cui, C. Xu, Z. Jie, X. Li, and J. Yang. 2018. "Joint Task-Recursive Learning for Semantic Segmentation and Depth Estimation." In *Computer Vision—ECCV 2018: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part X*, 238–255. Springer-Verlag. https://doi.org/10.1007/978-3-030-01249-6_15.

Zhang, Z., Z. Cui, C. Xu, Y. Yan, N. Sebe, and J. Yang. 2019. "Pattern-Affinitive Propagation Across Depth, Surface Normal and Semantic Segmentation." In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4106–4115. IEEE. https://doi.org/10.1109/CVPR.2019.00423.

Zhao, X., H. Li, X. Shen, X. Liang, and Y. Wu. 2018. "A Modulation Module for Multi-Task Learning With Applications in Image Retrieval." In *Proceedings of the European Conference on Computer Vision (ECCV)*, 415–432. Springer International Publishing. https://doi.org/10.1007/978-3-030-01246-5_25.

Zhu, Z., K. Lin, A. K. Jain, and J. Zhou. 2023. "Transfer Learning in Deep Reinforcement Learning: A Survey." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45: 13344–13362. https://doi.org/10.1109/TPAMI.2023.3292075.