Towards a Unified Framework for Adaptable Problematic Content Detection via Continual Learning

Ali Omrani[*]
University of Southern Califronia
aomrani@usc.edu

Alireza S. Ziabari[*]
University of Southern California
salkhord@usc.edu

Preni Golazizian
University of Southern California
golazizi@usc.edu

Jeffrey Sorensen
Jigsaw
sorenj@google.com

Morteza Dehghani
University of Southern California
mdehghan@usc.edu

Abstract

Detecting problematic content, such as hate speech, is a multifaceted and ever-changing task, influenced by social dynamics, user populations, diversity of sources, and evolving language. There has been significant efforts, both in academia and in industry, to develop annotated resources that capture various aspects of problematic content. Due to researchers' diverse objectives, these annotations are often inconsistent and hence, reports of progress on the detection of problematic content are fragmented. This pattern is expected to persist unless we pool these resources, taking into account the dynamic nature of this issue. In this paper, we propose integrating the available resources, leveraging their dynamic nature to break this pattern, and introduce a continual learning framework and benchmark for problematic content detection. Our benchmark, comprising 84 related tasks, creates a novel measure of progress: prioritizing the adaptability of classifiers to evolving tasks over excelling in specific tasks. To ensure continuous relevance, our benchmark is designed for seamless integration of new tasks. Our results demonstrate that continual learning methods outperform static approaches by up to 17% and 4% AUC in capturing the evolving content and adapting to novel forms of problematic content.
***Warning**: datasets contain offensive language.*

1 Introduction

Our social contexts continuously evolve and adapt to new situations, a characteristic that has empowered us to navigate through various challenges such as wars or pandemics. Peoples' expressions of hate, toxicity, and incivility, among other types of biases and prejudices, undergo adaptations in response to such changing circumstances. For instance, when there is a shift in the social or economic context, novel forms of hate speech emerge (Tahmasbi et al., 2021). In such scenarios, fear and uncertainty contribute to the proliferation of stereotypical beliefs and the attribution of blame to particular groups (Cinelli et al., 2020). The rise of anti-asian racism during the Covid-19 pandemic (Cowan, 2021) or surges of anti-muslim and anti-semitic hate during the Israel-Hamas conflict (Frenkel and Myers, 2023) are two recent examples of such cases. Even in stable social situations, differences in countries, contexts, and perspectives shape the boundaries of what is considered problematic (Klonick, 2017).

The field of problematic content detection has produced an abundance of resources aiming to capture various aspects of this ever-changing phenomenon (Poletto et al., 2021; Vidgen and Derczynski, 2020). While the accumulation of such resources may appear to bring us closer to effectively addressing this problem, the static viewpoint adopted by each resource has resulted in heterogeneity among them, posing a significant challenge for integration of their knowledge into models. This heterogeneity has also caused fragmentation in progress reports on the automatic detection of problematic content (Yin and Zubiaga, 2021). Therefore, it is crucial to establish a benchmark that integrates these annotated resources while capturing the dynamic nature of this problem. Such a benchmark would provide a more practical setting to test our models under stress and offer a new way to measure progress.

In this paper, we introduce a continual learning benchmark and framework for problematic content detection comprising 84 related tasks encompassing 15 annotation schemas from 8 sources. By doing so, we present a novel perspective to address the problem of problematic content detection. Instead of focusing solely on specific aspects, such as the toxicity of a snapshot of a platform, we advocate for a dynamic formulation that builds on the ever-changing nature of problematic content.

Further, we propose a framework for identifying

[*]These authors contributed equally to this work.

Proceedings of the The 8th Workshop on Online Abuse and Harms (WOAH), pages 68–109

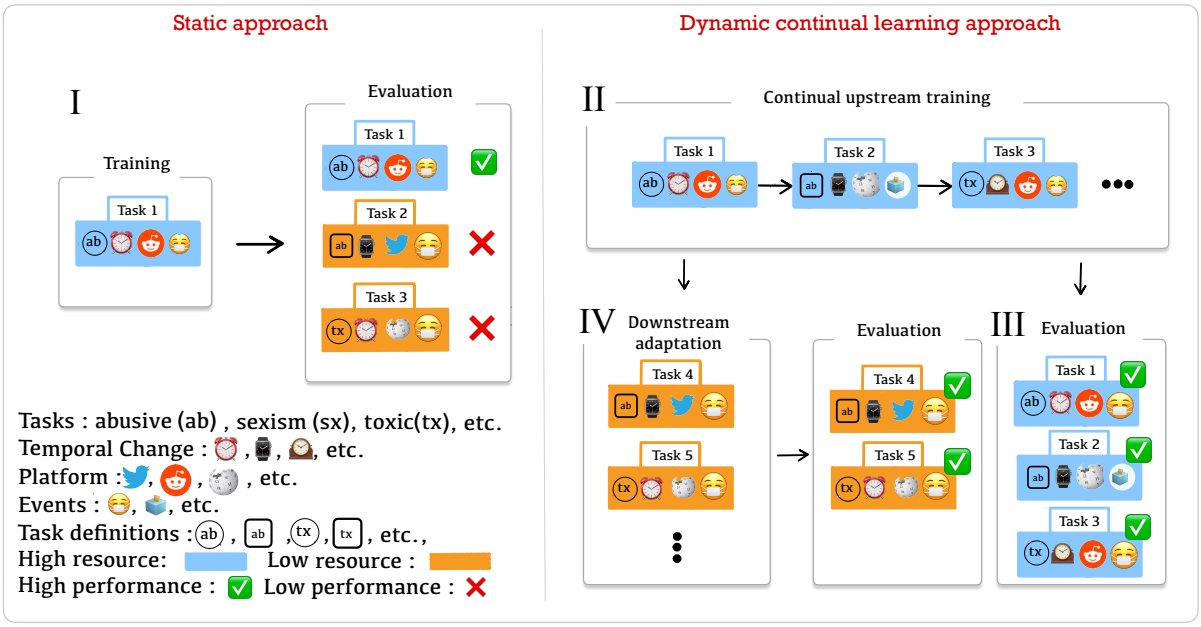

Figure 1: Current static approaches (I) train and evaluate models on a fixed set of datasets. Our benchmark embraces the dynamic aspects of problematic content detection in two stages. The upstream training (II) and evaluation (III) where data is assumed to be coming in a stream, and downstream fewshot evaluation (IV) that measure models' generalization to novel forms of problematic content.

problematic content in a dynamic setting which satisfies the following two objectives: First, an optimal model should have the capability to acquire and retain knowledge about various types of problematic content. This capability is particularly crucial for effectively utilizing the diverse datasets that exist for detecting problematic content. We model this capability through a continual learning formulation, drawing inspiration from previous research (Robins, 1995; de Masson D'Autume et al., 2019; Sun et al., 2019). Our models are designed to learn and understand the intricacies of problematic content by performing a diverse set of related tasks. Second, an optimal model should also have the ability to quickly learn and recognize new instances of problematic content, regardless of whether they appear on new platforms, in different languages, or target new groups. To assess and reward models that can adapt rapidly to emerging problematic content, we employ a few-shot evaluation benchmark on a separate set of related tasks, as suggested by recent work (Jin et al., 2021).

Through these objectives, we establish criteria for an ideal model that can effectively handle the dynamic nature of problematic content. We define metrics and evaluations that capture these criteria, and we create a benchmark that accurately reflects the complexities of the problem. In constructing this benchmark, we integrate existing resources in the field, leveraging their strengths to develop a comprehensive framework for studying the evolution of problematic content online.

To validate the effectiveness of our proposed approach in a practical setting, we set up our experiments to simulate the evolution of problematic content research (§5). Our results show that dynamic continual learning approaches outperform static methods in all the identified criteria for an ideal model, namely, accumulating knowledge and generalizing to novel forms of problematic content (§6). In sum, by addressing the dynamic nature of problematic content and embracing its complexities, our framework, benchmark, and experiments offer valuable insights, resources, and practical solutions for combating problematic content [1].

2 Background

2.1 Problematic Content Detection

Social media platforms offer individuals means to freely express themselves. However, certain features of social media, such as partial anonymity, which may promote freedom of expression, can also result in dissemination of problematic content.

[1]Our benchmark and experiments are available at https://github.com/USC-CSSL/Adaptable-Problematic-Content-Detection

Researchers and social media companies recognize this issue and have developed various strategies to tackle it, including automated systems to identify problematic content. Consequently, multiple definitions of problematic content have been proposed (Poletto et al., 2021), encompassing specific areas like misogyny detection (e.g., Fersini et al., 2018), to hate speech (e.g., Kennedy et al., 2022) and broader categories such as offensive language detection (e.g., Davidson et al., 2017). Ideally, such systems should possess the capability to identify undesirable content irrespective of factors such as timing, specific linguistic form, or the social media platform used. However, recent studies have revealed limited generalizability of such systems, particularly in the context of hate speech detection (Yin and Zubiaga, 2021; Ramponi and Tonelli, 2022). Yin and Zubiaga (2021) recognized that the scarcity of hate speech in sources poses a challenge to constructing datasets and models. They also acknowledged the difficulty in modeling implicit notions of problematic content. Combining diverse datasets can alleviate both issues by reducing the scarcity of problematic content and enhancing a model's ability to identify implicit notions through exposure to a broader range of data.

2.2 Multitask Learning for Problematic Content

In recent years, multitask learning (Caruana, 1997) has gained considerable attention as a promising approach for problematic content detection (Kapil and Ekbal, 2021; Plaza-Del-Arco et al., 2021; Farha and Magdy, 2020; Kapil and Ekbal, 2020; Talat et al., 2018). Multitask learning leverages the inherent relationships and shared characteristics among related tasks (e.g., hate speech, racism, sexism detection etc. in the context problematic content) to improve performance over a model that learns the tasks individually. By jointly training on multiple related tasks, the models can benefit from knowledge transfer and information sharing across different domains. Furthermore, empirical evidence shows the advantage of multitask learning in enhancing generalization and robustness. This advantage could potentially be due to the model's ability to learn common patterns and effectively differentiate between various forms of harmful language across different tasks (Mao et al., 2020; Zhou et al., 2019; Kapil and Ekbal, 2020).

Although multitask learning has demonstrated potential in the field of problematic content detection, it is not exempt from limitations. A significant drawback is the expense involved in retraining the model whenever a new task is introduced to the existing set. As the number of tasks grows, so does the complexity and computational resources needed for retraining. This becomes particularly challenging in the context of a dynamic landscape of problematic content, where new types of hate speech or toxic behavior emerge constantly. Multitask learning encounters various other challenges apart from computational complexity. These challenges include task interference, a phenomenon wherein the acquisition of multiple tasks concurrently can exert a detrimental impact on each other's learning processes, and catastrophic forgetting, which entails the loss of previously acquired knowledge when learning new tasks (Robins, 1995; Kirkpatrick et al., 2017; Wu et al., 2023).

2.3 Continual Learning and Few Shot Generalization

Continual learning is an approach that has emerged to address challenges like task interference, computational complexity, and catastrophic forgetting faced by multitask learning; instead of simultaneously learning all the tasks, continual learning models learn new tasks over time while maintaining knowledge of previous tasks (Robins, 1995). This incremental approach allows for efficient adaptation to new tasks while preserving the knowledge acquired from the previous tasks (Parisi et al., 2019). By leveraging techniques such as parameter isolation, rehearsal, or regularization, continual learning mitigates catastrophic forgetting and ensures that the model retains its proficiency in previously learned tasks (Kirkpatrick et al., 2017; de Masson D'Autume et al., 2019; Wang et al., 2020; Schwarz et al., 2018). Moreover, the capability to incrementally update the model alleviates the computational burden associated with retraining the entire multitask model every time new tasks are added. As a result, continual learning presents a promising approach to tackle the scalability and adaptability issues inherent in multitask learning. This framework becomes particularly attractive for tasks like hate speech detection, toxicity detection, and similar endeavors within a rapidly changing environment of problematic content. The only work in this space is Qian et al. (2021) which applies continual learning to detect hate speech on Twitter. However, their focus is limited to a single definition of hate speech and they analyze a single snapshot

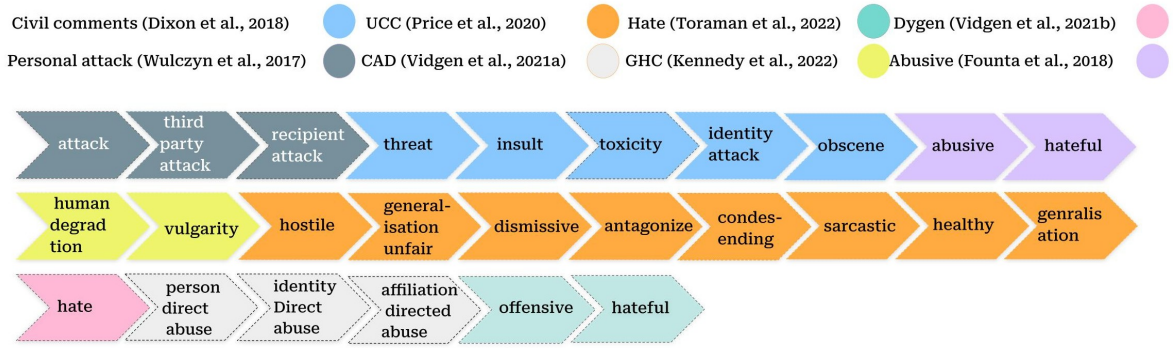

Figure 2: Sequence of upstream tasks in the experiment with chronological task order. Note that datasets are ordered according to the earliest publication date of the data and tasks (i.e., labels) within each dataset are ordered randomly.

of Twitter data. Consequently, their approach does not fully account for the dynamic nature of problematic content across the internet.

3 Continual Learning Benchmark for Problematic Content Detection

3.1 Problem Formulation

Our objective is to develop models that are not only agile in detecting new manifestations of problematic content but are also capable of accumulating knowledge from diverse instances across different time periods and platforms. Such models should possess the ability to rapidly learn and identify new manifestations of problematic content on novel platforms, even when only limited data is available. As time progresses, we anticipate a natural increase in the availability of resources for problematic content detection. Therefore, to encourage building models that leverage this increase in resources, we consider the existing resources as a continuous stream of incoming data. In this context, we make the assumption that there exists a problematic content detection model denoted as f, which undergoes continual learning on a stream of problematic content detection binary classification tasks ($T^u = [T^u_1, \ldots, T^u_{N_u}]$) over time. We refer to this set of tasks as *upstream* tasks. In addition to accumulating knowledge from the stream of tasks, this continual learning model should be able to rapidly generalize its knowledge to numerous related unseen tasks (Jin et al., 2021). We formulate this ability as few-shot learning over a separate set of binary classification tasks $T^d = [T^d_1, \ldots, T^d_{N_d}]$, referred to as *downstream* tasks.

3.2 Training and Evaluation

During the continual learning stage, the model encounters a sequentially ordered list of N_u upstream tasks: $[T^u_1, \ldots, T^u_{N_u}]$, where each task has its own distinct training and test sets. To evaluate the few-shot learning capability of the sequentially trained model f, we proceed to adapt it to a collection of N_d few-shot tasks individually represented as T^d_i. In this scenario, each unseen task is associated with only a small number of training examples.

For evaluation purposes, a task is considered "new" if the model hasn't been exposed to labels from that task. This applies to the i_{th} upstream task (T^u_i) in the upstream training process before the model's upstream training reaches T^u_i, as well as to all downstream tasks (Figure 1). The paucity of problematic content online results in most datasets used in this work being quite unbalanced. In the evaluation of models trained on such unbalanced datasets, Area Under the Curve (AUC) often takes precedence over the F_1 score (Bradley, 1997). AUC serves as a measure of a model's ability to differentiate between positive and negative classes, calculated by assessing the area under the Receiver Operating Characteristic (ROC) curve. Hence, we chose AUC as our primary evaluation metric for both the upstream training and downstream adaptation processes. We acknowledge that the selection of an evaluation metric is not without its controversies. The rationale behind this choice primarily stems from the extensive adoption of the AUC in the problematic content detection literature. In the context of this work, it is important to note that our conclusions would have remained consistent even if we had opted for the F_1 score as our primary metric (§A.7.) To enable fair comparisons, we used a fixed set of held-out test data for all models. Be-

low we outline the specific measures we employ to characterize the desired attributes of each model.

Few-Shot Performance To assess the model's few-shot generalization ability, we evaluate the continually trained model f on unseen tasks by individually fine-tuning it for each task T_i^v using a few annotated examples. The few-shot AUC for task T_i^d is denoted as AUC_i^{FS}, and we report the average few-shot AUC across all downstream tasks.

Final Performance To assess the accumulation of knowledge in upstream tasks, we evaluate the AUC of f at the end of the continual learning over upstream tasks. This evaluation allows us to determine the extent to which model f forgets the knowledge pertaining to a specific task once it acquires the ability to solve additional tasks. We report the average AUC over all upstream tasks.

Instant Performance To assess the extent of positive transfer among upstream tasks, we evaluate the AUC of f on task T_i^u right after the model is trained on T_i^u. We report the average of instant performance across all upstream tasks.

3.3 Datasets

We have selected datasets for our benchmark based on the following criteria: 1) must be related to problematic content detection, 2) must be in English, and 3) must include a classification task (or a task transformable into classification). We aimed to use datasets that span different sources and time periods, and rely on different definitions of problematic content. Even though we currently focus on one language, the dynamic nature of our formulation easily allows for expansion of this benchmark to other languages (see §8 for more details). Our benchmark currently covers data from 8 different sources, namely, Twitter, Reddit, Wikipedia, Gab, Stromfront, chat dialogues, and synthetically generated text. These datasets cover a wide range of definitions of problematic content, from focused definitions such as sexism and misogyny to broader definitions such as toxicity. All datasets in our work are publicly available for research purposes. We do not redistribute these datasets but offer instructions in our repository for downloading and recreating the benchmark from publicly available sources. In addition, we provide license information for all datasets, along with descriptive statistics in §A.2. For all datasets, we use the original train/test/dev splits when available, otherwise split the data 80/10/10 randomly. We briefly discuss each dataset below; [U] and [D] denote upstream and downstream datasets respectively.

Call Me Sexists, But (CMSB; Liakhovets et al., 2022) [D] Consists of 6,325 tweets from two sources: 1) Twitter data that was previously annotated for sexism and racism (Waseem and Hovy, 2016), and 2) Twitter data collected between 2008 and 2019 using the phrase "call me sexist, but." Each tweet is labeled for sexist content and sexist phrasing, with both being single-choice options.

US-election (Grimminger and Klinger, 2021) [D] Consists of 3,000 tweets, covering hate speech and offensive language, which were collected during the six weeks prior to the 2020 presidential election, until one week after the election. Each tweet was annotated for being hateful or not, without considering whether the target is a group or an individual.

Misogyny Detection (misogyny; Guest et al., 2021) [D] Contains 6,567 Reddit Posts from 34 subreddits identified as misogynistic from February to May 2020 annotated with a three level hierarchical taxonomy. We only use the top level annotations which are binary labels for misogynistic content.

Contextual Abuse Dataset (CAD; Vidgen et al., 2021a) [U] Consists of 25k Reddit posts collected from 16 Subreddits more likely to contain a diverse range of abusive language, and focused on taking the context of the conversations into account. A hierarchical annotation schema is proposed which takes the context of the conversation into account; Level 1: abusive, non-abusive, and Level 2: for abusive (i) identity-directed, (ii) affiliation-directed and (iii) person-directed. In our benchmark, we use the three labels from the second level to stress test models' ability in learning variations of abuse.

Ex-Machina: Personal Attacks at Scale (Personal attack; Wulczyn et al., 2017) [U] Includes 100k annotated comments from a public dump of Wikipedia from 2004-2015. Annotators were asked to label comments that contain personal attack or harassment in addition to some finer labels about the category of attack or harassment. We included the detecting personal attacks, quoted personal attacks (QA), and personal attack targeted at third party (TPA) as separate tasks in our benchmark.

Unhealthy Comment Corpus (UCC; Price et al., 2020) [U] Consists of 44,355 comments collected from the Globe and Mail news site. Every comment is annotated according to a two-level hierarchy; Level 1: healthy or unhealthy. Level 2: binary labels indicating the presence or absence of six specific unhealthy subattributes: (i) hostility, (ii) antagonism, (iii) insults, (iv) provocation, (v)

trolling, (vi) dismissiveness, (vii) condescension, (viii) sarcasm, and (ix) generalization.

The Gab Hate Corpus (GHC; Kennedy et al., 2022)[U] Contains 27,665 posts from Gab.com, spanning January to October, 2018, annotated based on a typology for hate speech derived from definitions across legal precedent. Posts were annotated for Call for Violence (CV), Human degradation (HD), Vulgarity and/or Offensive language (VO), and explicit or implicit language.

Stormfront (De Gibert et al., 2018) [D] Includes a 10,568 sentences collected from 22 sub-forums of Stormfront.org spanning from 2002 to 2017. Each sentence has been classified as containing hate or not depending on whether they meet the following three premises: "a) deliberate attack, b) directed towards a specific group of people, and c) motivated by aspects of the group's identity."

Dialogue Safety (Miller et al., 2017; Xu et al., 2021) [D] The Dialogue Safety dataset includes five datasets in the domain of dialogue safety. Three datasets, namely ParlAI single standard, ParlAI single adversarial, and ParlAI multi, are sourced from ParlAI (Miller et al., 2017). The other two datasets, BAD2 and BAD4, are from Bot-Adversarial Dialogue (Xu et al., 2021). The ParlAI datasets consist of 30,000 samples, while the BAD datasets consist of 5,784 samples. Conversations in the BAD dataset can span up to 14 turns, and following (Xu et al., 2021), we consider the last two and four utterances of the conversation (BAD2 and BAD4) in our benchmark. All dialogue safety datasets provide toxic or safe labels.

Dygen (Vidgen et al., 2021b) [hate U, rest D] Consists of 41,255 samples dynamically generated using the human-and-model-in-the-loop setting to train more robust hate detection models. The authors collected four rounds data using *Dynabench* (Kiela et al., 2021), and annotated each sample hierarchically; Level 1: binary hate/non-hate label, Level 2: subclasses of hate (i.e., derogation, animosity, threatening language, support for hateful entities and dehumanization) and 29 target identities (e.g., immigrant, muslim, woman, etc.). We use Level 1 for upstream training and Level 2 for downstream adaptation.

Hatecheck (Röttger et al., 2021) [D] Contains of 3,728 synthetically generated sentences motivated by 29 hate speech detection model functionalities; 18 of these functionalities test for hateful content and cover distinct expressions of hate, and the other 11 functionalities test for non-hateful content and cover contrastive non-hate.

Multitarget-CONAN (CONAN; Fanton et al., 2021) [D] Consists of 5003 samples of hate speech and counter-narrative pairs targeting different target groups (LGBTQ+, Migrants, Muslims, etc.) created using human-in-the-loop methodology, in which the generative language model generates new samples and, after confirmation by expert annotators, would get added to the dataset. In our benchmark we included detection of hate speech toward each target group as a separate task.

Civil-comments (Dixon et al., 2018) [U] Includes two million comments from the Civil Comments platform annotated by human raters for various toxic conversational attributes. Each comment has a toxicity label and several additional toxicity subtype attributes which are severe toxicity, obscene, threat, insult, identity attack, sexual explicit.

Twitter Abusive (Abusive; Founta et al., 2018) [U] Contains 80k tweets from March to April 2017 annotated for multiple fine-grained aspects of abuse, namely, offensiveness, abusiveness, hateful speech, aggression, cyberbullying, and spam.

Large-Scale Hate Speech Detection with Cross-Domain Transfer (hate; Toraman et al., 2022) [U] Includes 100k tweets from 2020 and 2021, each annotated by five annotators for hate speech. Tweets are labeled as hate if "they target, incite violence against, threaten, or call for physical damage for an individual or a group of people because of some identifying trait or characteristic."

4 Models and Methods

4.1 Models

We represent all tasks in a consistent binary classification format and conduct our experiments using a pretrained language model, specifically BART-Basew (Lewis et al., 2020). In addition to fine-tuning all the model weights of BART-Base, we also explore two other variations 1) **Adapter:** We experiment with Adapters (Houlsby et al., 2019). In addition to the classification head, adapter training only trains parameters of Adapters, which are two-layer multilayer perceptrons inserted after each layer of BART. We used a hidden size of 256 for all Adapter layers. 2) **BiHNet:** The hypernetwork (h) accepts a task representation z as input and generates model parameters for a separate prediction model, denoted as f, in order to address the specific task at hand (Jin et al., 2021).

	Final		Instant		Fewshot	
Model	AUC	Δ AUC	AUC	Δ AUC	AUC	Δ AUC
Adapter-Single	-	-	0.879	-	0.806	-
BiHNet-Single	-	-	0.870	-	0.786	-
Adapter-Vanilla	0.518	-	0.882	-	0.765	-
BiHNet-Vanilla	0.617	-	0.878	-	0.772	-
BiHNet-Reg	0.792	+0.174	0.882	+0.003	0.819	+0.047
BiHNet-EWC	0.676	+0.059	0.881	+0.003	0.766	-0.006
Adapter-Multitask	0.873	+0.355	-	-	0.816	+0.052
BiHNet-Multitask	0.834	+0.216	-	-	0.796	+0.024

Table 1: AUC scores for chronological experiment. Δ values are calculated in comparison to the corresponding Vanilla model.

4.2 Upstream Training

Single Task Learning We finetune a pretrained model on each of the tasks separately. Note that this model completely ignores the sequential order imposed on our upstream tasks and serves as a baseline for evaluating the performance of the base model each tasks without any knowledge transfer.

Sequential Finetuning (Vanilla) We also finetune a pretrained model on the sequence of upstream tasks $[T^u_1, \ldots, T^u_{Nu}]$ without any continual learning algorithms. Previous research suggests that this model will suffer from catastrophic forgetting (Robins, 1995). Comparing the final performance of this model with a continual learning algorithm will give us a measure of the ability of these algorithms in knowledge accumulation.

Multitask Learning (MTL) To assess the upper bound of knowledge accumulation on the set of upstream tasks we finetune a pretrained model with multitask learning on all upstream tasks implemented via hard parameter sharing. For **Adapter-Multitask** models we shared only the adapter parameters and for **BiHNet-Multitask** models we used a shared BiHNet for all tasks.

Continual Learning Finally, we finetune a model continually on a sequence of upstream tasks $[T^u_1, \ldots, T^u_{Nu}]$. This model should ideally be able to 1) use knowledge from previous tasks to learn a new upstream task, and 2) retain knowledge of the seen upstream tasks. We experiment with two regularization-based continual learning algorithms:

Bi-level Hypernetworks for Adapters with Regularization (**BiHNet-Reg**: Jin et al., 2021). This model is specifically designed to enhance the generation of adapter weights by optimizing bi-level long and short-)task representations. Its primary objective is to address two important challenges: mitigating catastrophic forgetting and enhancing the overall generalizability of the model. Towards the first challenge, regularization is imposed on the generated adapters. To improve generalization this model learns two representations for each task task; one for high-resource settings and one for few-shot cases. We calculated the long task representation by averaging the embedding of all text samples in the training split of a dataset. short task representations were computed by averaging embeddings of 64 texts sampled from the training set.

Elastic Weight Consolidation (**EWC**: Kirkpatrick et al., 2017): leverages the principles of Bayesian inference, suggesting a method that selectively slows down learning on the weights important for previous tasks. The model retains old knowledge by assigning a larger penalty to changes in crucial parameters, effectively making them "elastic".

4.3 Downstream Adaptation

An ideal model for problematic content detection should be able to learn its new manifestations quickly. Therefore, we evaluate our models' ability on learning unseen datasets of problematic content using only a few examples. We report the performances using $k = 16$ shots. Sensitivity analysis on the number of shots is provided in §A.5.

5 Experiments

Most of the datasets in our benchmark include annotations for various aspects of problematic content (e.g., UCC includes labels for antagonism, insults, etc.). To ensure flexibility, we treated each label as a separate task. This choice is rooted in the likely possibility that we will need to introduce additional labels to the existing set in the future. To accommodate potential future updates to the label

taxonomy, it is preferable to have models that can quickly adapt and learn new labels.

In order to minimize the exchange of information between the upstream and downstream tasks, across all our datasets with the exception of Dygen, we categorized all tasks within the dataset as either upstream or downstream. Our selection of larger datasets for the upstream tasks was driven by both the data requirements of upstream training and the fact that larger datasets typically encompass a broader range of problematic content. This decision enables the model to accumulate knowledge on general notions of problematic content, which aligns with our objectives. Subsequently, we assigned tasks as downstream that 1) had limited labeled data, and 2) had minimal overlap (e.g., same domain or labels) with the upstream tasks.

To assess the efficacy of our proposed framework in practical scenarios, we ran our main experiments by ordering the upstream tasks *chronologically*. Specifically, we used the earliest publication date of each dataset as the temporal reference point to order the upstream datasets. Note that each dataset consists of multiple labels (i.e., tasks). Since we don't have any information about the temporal order of tasks within datasets, we chose this order at random. This experiment allowed us to capture the evolution of the research landscape on problematic content detection, thereby providing a more nuanced understanding of the progress of model performance over time. Figure 2 shows the order of upstream tasks in this experiment. We experiment with alternative orders of upstream tasks in §A.4.

6 Results

Baselines: To determine the learning capabilities of each model, we finetune a classifier from each architecture on each task. The average fewshot, final, and instant performance of Adapter-Vanilla, and BiHNet-Vanilla is presented in the first two rows of table 1 respectively. We see the largest gap in performance for these models on the final performance metrics. This can be attributed to BiHNet's meta learning capabilities.

Multitask Upperbound: When there are no adversarial tasks, multitask learning is often used as an empirical upper bound for continual learning. The last two rows of table 1 show the few shot and final evaluation of multitask models. Note that since these models see all tasks at the same time, instant performance is not defined for them.

Does the collection of problematic content tasks help with learning new upstream tasks? In other words, do the models benefit from upstream training when learning a new task with substantial amount of annotated data available? To answer this question, compare the instant performance of a CL model on T_i^u with a pretrained model finetuned on just T_i^u. Our results (Δ Instant AUC) show evidence of slight positive transfer, however, the magnitude of this transfer is negligible.

Does continual learning improve knowledge retention? The final AUC values, as shown in Table 1, indicate the models' ability to retain knowledge from a sequence of tasks at the end of training. Our results suggest that all continual learning variations outperform naive training. Most notably, BiHNet-Reg outperforms BiHNet-Vanilla by almost 18% in AUC, indicating its potential to mitigate catastrophic forgetting, while falling only 4% short of the multitask counterpart.

Does upstream learning help with generalization to new manifestations of problematic content? Comparing the single-task baselines with continual and multitask learning, our results (Table 1) demonstrate a noteworthy improvement in models' generalization ability as a result of upstream training. Specifically, BiHNet-Reg shows remarkable generalization ability in fewshot settings, outperforming the BiHNet-Vanilla by nearly 5% in AUC.

7 Discussion and Conclusion

In conclusion, we propose a continual learning benchmark and framework for detecting problematic content, that realizes its dynamic and adaptable nature. We define essential characteristics of an ideal model and create a continual learning benchmark and evaluation metrics to capture the variability in problematic content. Our benchmark has two key components: First, an upstream sequence of problematic tasks over which we measure a model's ability in accumulating knowledge, and second, a separate set of downstream few-shot tasks on which we gauge a model's agility in learning new manifestations of problematic content. Our experiments clearly demonstrate the effectiveness of this formulation, particularly in its ability to adapt to new types of problematic content. To keep the benchmark up-to-date, we have designed it with continuous updates in mind; tasks can be effortlessly added, removed, or repositioned. We encourage the community to actively contribute to

and expand this benchmark, as it serves as a collaborative platform for advancements in the field.

8 Limitation and Negative Societal Impact

We emphasize that this is only one experimental scenario for dividing the tasks into upstream and downstream. Our benchmark's modular design allows for easy experimentation with other scenarios allowing researchers to further study various continual learning setups and evaluate a variety of continual learning algorithms. The social science examination of the evolution of problematic content carries its own importance and follows a dedicated line of inquiry. Due to space constraints, we have not provided an exhaustive discussion of this subject. We recommend referring to (Klonick, 2017; Atlantic-Council, 2023) for a comprehensive overview of this area. We acknowledge that the experiments in our paper are limited to the continual learning methods employed. We encourage future researchers to explore other continual learning approaches. The benchmark under discussion is currently designed only for English language content, neglecting the challenges posed by problematic content in other languages and cultures. Our design, however, allows for an easy expansion of the benchmark to include other languages. We have outlined the procedure to expand the benchmark on the accompanying repository and encourage the community to contribute to the benchmark. Though it presents a new measure of progress and baseline results, further investigations and extensive experimentation are needed to fully evaluate the potential of continual learning in detecting evolving problematic content. The study's approach, predominantly using majority label datasets, potentially leads to bias and overgeneralization in detecting problematic content, given the inherent subjectivity of such content influenced by cultural norms, individual sensitivities, and societal changes over time. The effectiveness of this benchmark could significantly vary due to the diversity of sources and annotation schemas, potentially leading to cultural bias and an overreliance on AI for content detection, thereby neglecting the importance of nuanced human moderation. Future work can explore the potential considering this subjectivity under our continual learning framework. Moreover, the benchmark opens possibilities for misuse, including training models to generate problematic content or designing adversarial attacks, where malicious actors can exploit the understanding of detection systems to craft content that evades detection.

Datasets used in this benchmark may have a high prevalence of problematic content targeting certain social groups. Hence, models trained on these datasets could produce unfair outcomes, such as higher false positive rates for the aforementioned groups (Dixon et al., 2018; Wiegand et al., 2019). Recently, various methods have been proposed to mitigate these biases, such as those by Mostafazadeh Davani et al. (2021); Kennedy et al. (2020); Omrani et al. (2023). Future research could examine the extent of biases' influence on the model within our framework and the effectiveness of the mentioned techniques in mitigating them. Moreover, some datasets may hold personally identifiable information or data from which individual details can be inferred. Since we are not redistributing any of the datasts, to address this concern, we suggest applying Google's DLP, a tool designed to scan and classify sensitive data, to the datasets. Another concern in research on problematic content detection is the potential misuse for censorship. However, we emphasize that, in contrast to private methods concealed behind corporate doors, an open-access or academic approach to detecting problematic content fosters transparency. This allows the public to understand and critique the detection criteria. Such transparency ensures accountability, given that academic methods frequently undergo peer review and public scrutiny, thereby addressing biases and mistakes.

References

Atlantic-Council. 2023. Scaling trust on the web. Technical report, Atlantic Council.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48.

Andrew P Bradley. 1997. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159.

Rich Caruana. 1997. Multitask learning. *Machine learning*, 28(1):41–75.

Matteo Cinelli, Walter Quattrociocchi, Alessandro Galeazzi, Carlo Michele Valensise, Emanuele Brugnoli, Ana Lucia Schmidt, Paola Zola, Fabiana Zollo, and Antonio Scala. 2020. The COVID-19 social media infodemic. *Scientific Reports*, 10(1).

Jill Cowan. 2021. Looking at the rise of anti-asian racism in the pandemi.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515.

Ona De Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate speech dataset from a white supremacy forum. *arXiv preprint arXiv:1809.04444*.

Cyprien de Masson D'Autume, Sebastian Ruder, Lingpeng Kong, and Dani Yogatama. 2019. Episodic memory in lifelong language learning. *Advances in Neural Information Processing Systems*, 32.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73.

Jeffrey L Elman. 1993. Learning and development in neural networks: The importance of starting small. *Cognition*, 48(1):71–99.

Margherita Fanton, Helena Bonaldi, Serra Sinem Tekiroğlu, and Marco Guerini. 2021. Human-in-the-loop for data collection: a multi-target counter narrative dataset to fight online hate speech. 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)*, pages 3226–3240.

Ibrahim Abu Farha and Walid 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*, pages 86–90.

Elisabetta Fersini, Paolo Rosso, and Maria Anzovino. 2018. Overview of the task on automatic misogyny identification at ibereval 2018. *Ibereval@ sepln*, 2150:214–228.

Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Proceedings of the international AAAI conference on web and social media*, volume 12.

Sheera Frenkel and Steven Lee Myers. 2023. Anti-semitic and anti-muslim hate speech surges across the internet.

Lara Grimminger and Roman Klinger. 2021. Hate towards the political opponent: A Twitter corpus study of the 2020 US elections on the basis of offensive speech and stance detection. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 171–180, Online. Association for Computational Linguistics.

Ella Guest, Bertie Vidgen, Alexandros Mittos, Nishanth Sastry, Gareth Tyson, and Helen Margetts. 2021. An expert annotated dataset for the detection of online misogyny. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1336–1350.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR.

Xisen Jin, Bill Yuchen Lin, Mohammad Rostami, and Xiang Ren. 2021. Learn continually, generalize rapidly: Lifelong knowledge accumulation for few-shot learning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 714–729, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Prashant Kapil and Asif Ekbal. 2020. A deep neural network based multi-task learning approach to hate speech detection. *Knowledge-Based Systems*, 210:106458.

Prashant Kapil and Asif Ekbal. 2021. Leveraging multi-domain, heterogeneous data using deep multitask learning for hate speech detection. *ArXiv*, abs/2103.12412.

Zixuan Ke and Bing Liu. 2022. Continual learning of natural language processing tasks: A survey. *arXiv preprint arXiv:2211.12701*.

Brendan Kennedy, Mohammad Atari, Aida Mostafazadeh Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kris Coombs, Shreya Havaldar, Gwenyth Portillo-Wightman, Elaine Gonzalez, et al. 2022. Introducing the gab hate corpus: defining and applying hate-based rhetoric to social media posts at scale. *Language Resources and Evaluation*, pages 1–30.

Brendan Kennedy, Xisen Jin, Aida Mostafazadeh Davani, Morteza Dehghani, and Xiang Ren. 2020. Contextualizing hate speech classifiers with post-hoc explanation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5435–5442.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, et al. 2021. Dynabench: Rethinking benchmarking in nlp. *arXiv preprint arXiv:2104.14337*.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu,

Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

Kate Klonick. 2017. The new governors: The people, rules, and processes governing online speech. *Harv. L. Rev.*, 131:1598.

Kai A Krueger and Peter Dayan. 2009. Flexible shaping: How learning in small steps helps. *Cognition*, 110(3):380–394.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Daria Liakhovets, Mina Schütz, Jaqueline Böck, Medina Andresel, Armin Kirchknopf, Andreas Babic, Djordje Slijepčević, Jasmin Lampert, Alexander Schindler, and Matthias Zeppelzauer. 2022. Transfer learning for automatic sexism detection with multilingual transformer models.

Chengzhi Mao, Amogh Gupta, Vikram Nitin, Baishakhi Ray, Shuran Song, Junfeng Yang, and Carl Vondrick. 2020. Multitask learning strengthens adversarial robustness. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 158–174. Springer.

Alexander Miller, Will Feng, Dhruv Batra, Antoine Bordes, Adam Fisch, Jiasen Lu, Devi Parikh, and Jason Weston. 2017. Parlai: A dialog research software platform. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 79–84.

Aida Mostafazadeh Davani, Ali Omrani, Brendan Kennedy, Mohammad Atari, Xiang Ren, and Morteza Dehghani. 2021. Improving counterfactual generation for fair hate speech detection. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 92–101, Online. Association for Computational Linguistics.

Ali Omrani, Alireza Salkhordeh Ziabari, Charles Yu, Preni Golazizian, Brendan Kennedy, Mohammad Atari, Heng Ji, and Morteza Dehghani. 2023. Social-group-agnostic bias mitigation via the stereotype content model. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4123–4139, Toronto, Canada. Association for Computational Linguistics.

German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. 2019. Continual lifelong learning with neural networks: A review. *Neural networks*, 113:54–71.

Flor Miriam Plaza-Del-Arco, M Dolores Molina-González, L Alfonso Ureña-López, and María Teresa Martín-Valdivia. 2021. A multi-task learning approach to hate speech detection leveraging sentiment analysis. *IEEE Access*, 9:112478–112489.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55:477–523.

Ilan Price, Jordan Gifford-Moore, Jory Flemming, Saul Musker, Maayan Roichman, Guillaume Sylvain, Nithum Thain, Lucas Dixon, and Jeffrey Sorensen. 2020. Six attributes of unhealthy conversations. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 114–124, Online. Association for Computational Linguistics.

Jing Qian, Hong Wang, Mai ElSherief, and Xifeng Yan. 2021. Lifelong learning of hate speech classification on social media. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2304–2314, Online. Association for Computational Linguistics.

Alan Ramponi and Sara Tonelli. 2022. Features or spurious artifacts? data-centric baselines for fair and robust hate speech detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3027–3040. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

Anthony Robins. 1995. Catastrophic forgetting, rehearsal and pseudorehearsal. *Connection Science*, 7(2):123–146.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. HateCheck: Functional tests for hate speech detection models. 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)*, pages 41–58, Online. Association for Computational Linguistics.

Jonathan Schwarz, Wojciech Czarnecki, Jelena Luketina, Agnieszka Grabska-Barwinska, Yee Whye Teh, Razvan Pascanu, and Raia Hadsell. 2018. Progress & compress: A scalable framework for continual learning. In *International conference on machine learning*, pages 4528–4537. PMLR.

Fan-Keng Sun, Cheng-Hao Ho, and Hung-Yi Lee. 2019. Lamol: Language modeling for lifelong language learning. *arXiv preprint arXiv:1909.03329*.

Fatemeh Tahmasbi, Leonard Schild, Chen Ling, Jeremy Blackburn, Gianluca Stringhini, Yang Zhang, and Savvas Zannettou. 2021. "go eat a bat, chang!": On the emergence of sinophobic behavior on web communities in the face of covid-19. In *Proceedings of the Web Conference 2021*, WWW '21, page 1122–1133, New York, NY, USA. Association for Computing Machinery.

Zeerak Talat, James Thorne, and Joachim Bingel. 2018. Correction to: Bridging the gaps: Multi task learning for domain transfer of hate speech entection. *Online Harassment*, pages C1–C1.

Cagri Toraman, Furkan Şahinuç, and Eyup Yilmaz. 2022. Large-scale hate speech detection with cross-domain transfer. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2215–2225, Marseille, France. European Language Resources Association.

Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *Plos one*, 15(12):e0243300.

Bertie Vidgen, Dong Nguyen, Helen Margetts, Patricia Rossini, and Rebekah Tromble. 2021a. Introducing cad: the contextual abuse dataset. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2289–2303.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021b. Learning from the worst: Dynamically generated datasets to improve online hate detection. 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)*, pages 1667–1682, Online. Association for Computational Linguistics.

Zirui Wang, Sanket Vaibhav Mehta, Barnabás Póczos, and Jaime Carbonell. 2020. Efficient meta lifelong-learning with limited memory. *arXiv preprint arXiv:2010.02500*.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of abusive language: the problem of biased datasets. In *Proceedings of the 2019 conference of the North American Chapter of the Association for Computational Linguistics: human language technologies, volume 1 (long and short papers)*, pages 602–608.

Zihao Wu, Huy Tran, Hamed Pirsiavash, and Soheil Kolouri. 2023. Is multi-task learning an upper bound for continual learning? In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th international conference on world wide web*, pages 1391–1399.

Jing Xu, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. 2021. Bot-adversarial dialogue for safe conversational agents. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2950–2968.

Wenjie Yin and Arkaitz Zubiaga. 2021. Towards generalisable hate speech detection: a review on obstacles and solutions. *PeerJ Computer Science*, 7:e598.

Shuyan Zhou, Xiangkai Zeng, Yingqi Zhou, Antonios Anastasopoulos, and Graham Neubig. 2019. Improving robustness of neural machine translation with multi-task learning. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 565–571, Florence, Italy. Association for Computational Linguistics.

A Supplementary Material

A.1 Hardware and Runtimes

Experiments were conducted on Nvidia Quadro 6000 GPUs with Cuda version 11.4. Each upstream training for 26 tasks takes around 12 hours, and few-shot training and evaluation for all 58 downstream tasks for a single model takes around 6 hours to complete.

A.2 Data Sources, Statistics, and License Information

All of the datasets used in this benchmark are publicly available for research purposes. Table 5 provides license information for all datasets. We do not redistribute these datasets. In our Github repository[2] we offer a clear guide on how to create a local copy of all the datasets used in our benchmark, from the original sources. Our benchmark consists of English classification datasets that contain tasks related to problematic content detection. Each label from each dataset is treated as a separate task and we only used tasks with more than 100 positive examples in their training sets. Table 2 and 3 show dataset statistics along with the number of positive samples per task for downstream and upstream tasks, respectively. Table 4 shows number of datasets from each source.

[2] https://github.com/USC-CSSL/Adaptable-Problematic-Content-Detection

Dataset	Labels
Abusive	abusive (2763); hateful (503); total (8597)
CAD	affiliation directed abuse (242) ; identity directed abuse (514); person directed abuse (237); total (5307)
Dygen	hate (2268); total (4120)
GHC	human degradation (491); vulgarity (369); total (5510)
Gate	hateful (170); offensive (1247); total (10207)
Civil comments	identity attack (687); insult (5776); obscene (543); threat (221); toxicity (7777); total (97320)
Personal attack	attack (3056) ; recipient attack (1999) ; third party attack (204); total (23178)
UCC	antagonize (203); condescending (269) ; dismissive (150) ; generalisation (96) ; generalisation unfair (91) ; healthy (320) ; hostile (108) ; sarcastic (201) ; total (4425)

Table 2: Number of label occurrences in upstream tasks test sets.

Dataset	Labels
Dygen	Black men (7); African (8); Muslim women (10); Asylum seekers (13); Asians (15); Indigenous people (18); Gender minorities (21); Chinese people (25); Foreigners (26); Black women (27); Travellers (27); Non-whites (28); Mixed race (30); Gay women (31); East Asians (32); South Asians (32). Gay men (34); support (35); Arabs (45); threatening (48); Refguees (51); dehumanization (70); People with disabilities (79); Gay people (81); Immigrants (81); Trans people (90); Jewish people (111); Muslims (126); Black (211); animosity (315); derogation (1036); total (3009)
CONAN	disabled (22); jews (59); muslims (134); migrant (96); woman (67); LGBT (62); people of color (35); total (501)
Hatecheck	trans (42); black (44); immigrants (45); muslims (47); gay (48); disabled (50); women (60); hate (117); total (373)
single adversarial multi	toxic (300); total (3000)
BAD2 BAD4	toxic (44); total (190)
Stormfront	hate (239); total (478)
US-election	hateful (31); total (300)
GHC	calls for violence (24); total (5510)
CAD	counter-speech (66); total (5307)
Misogyny	misogynistic (73); total (657)
CMSB	sexist (181); total (2363)

Table 3: Number of label occurrences in downstream tasks test sets.

Source: Twitter (6); Reddit (2); Wikipedia (2); Gab (1) ; Stormfront (1); Chat dialogue (4); Synthetically generated (2); Civil Comments (1).

Table 4: Number of datasets by source.

Table 5: License information for all datasets used in the benchmark. According to this information, all datasets can be used for research purposes

Name	License	Source
UCC and Ex Machina	CC-BY-SA	https://en.wikipedia.org/wiki/Wikipedia:Copyrights#Contributors'_rights_and_obligations
Civil Comments Corpus	CC0	https://www.kaggle.com/ competitions/jigsaw unintended bias in toxicity classification/data
Misogyny Detection	MIT	https://github.com/ellamguest/online-misogyny-eacl2021
CAD	CC-By Attribution 4.0 International	https://zenodo.org/record/4881008
DYGEN	CC By 4.0	Footnote of the first page of the paper: https://dl.acm.org/doi/pdf/10.1145/3580305.3599318
HateCheck	CC By 4.0	https://github.com/paul-rottger/hatecheck-data/blob/main/LICENSE
CONAN	"resources can be used for research purposes"	https://github.com/marcoguerini/CONAN
Stormfront	CC-BY-SA	https://github.com/Vicomtech/hate-speech-dataset
GHC	CC-By Attribution 4.0 International	The GHC is available on the Open Science Framework (OSF, https://osf.io/edua3/), and the license is discussed in detail in section 4 of the paper
CMSB	CC BY-NC-SA 4.0	https://data.gesis.org/sharing/#!Detail/10.7802/2251
Large-Scale Hate Speech Detection with Cross-Domain Transfer	CC-BY-SA 4.0	https://github.com/avaapm/hatespeech/blob/master/LICENSE
US Election	data is publicly available	https://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/stance-hof/
Dialogue Safety	MIT	https://github.com/facebookresearch/ParlAI/blob/main/LICENSE
Twitter Abusive	CC-By Attribution 4.0 International	https://zenodo.org/record/2657374

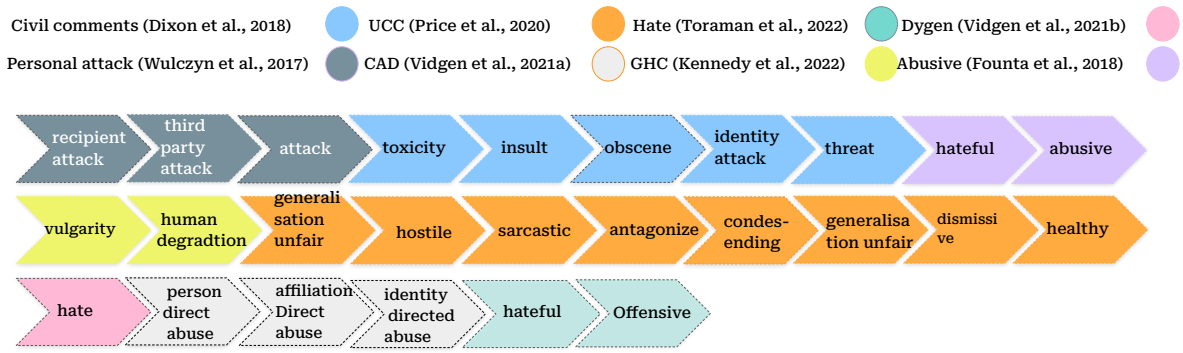

Figure 3: Shuffled sequence of tasks for the chronological experiment.

A.3 Model Implementation Details

For all experiments, we used a batch size of 32 and trained the models for at most 100 epochs. To prevent the model from overfitting, we used early stopping with a patience of three and chose the best model based on the F_1 score.Due to the paucity of problematic content online most of the datasets in this benchmark are heavily sparse. This sparsity poses challenges to the optimization process. To address this, we used a weighted random sampler ensuring each batch consists of at least 30% positive samples.

Adapter: To implement Adapter models, we added an adapter (Houlsby et al., 2019) between each layer of BART transformers. The adapter consists of an autoencoder with input and output layers of size equal to embedding dimensions and a hidden layer of size of 256 in the middle.

BiHNet: The BiHNet uses is an extension of the hypernetworks. BiHNet computes two different losses using two forms of task representation, long task representation and short task representation, to generate wights for the classification model. In our experiments, we calculated the long task representation by averaging the embedding of all text samples in the training split of a dataset. The short task embeddings, which are designed to help the model in few-shot settings, were computed by averaging embeddings of 64 texts sampled from the training set. For both long and short task representations, we used sentence-transformers (Reimers and Gurevych, 2019) [3] with mean pooling. The final model weights are calculated as the sum of weights generated using long and short task representations. Following Jin et al. (2021), we used a two-layer MLP model with a hidden size of 32 as our weight generator hypernetwork for each classification model. When BiHNet was used in a model variatoin that utilizes adapters, we used it to only generate the weights of all adapters in addition to each classification head.

Multitask Learning: In the multi-task setting, we used hard parameter sharing. For Adapter-Multitask models we shared only the adapter parameters and for BiHNet-Multitask models we used a shared BiHNet for all tasks. We use the BiHNet to generate task-specific parameters using the long and short task-specific representations.

Continual Learning Parameters: For BiHNet-Reg and BiHNet-EWC, both of which are regularization-based approaches (Ke and Liu, 2022), we used regularization coefficient of 0.01.

Downstream Adaptation: For downstream adaptation, we conducted few-shot training for 800 epochs with a batch size of 8 for 8-shot experiments. For 16-shot and 32-shot experiments, we used a batch size of 16. Since the total number of training samples is less than 64 in our downstream few-shot adaptations, we only use the long task representation for BiHNet models. For Adapter-Multitask, we initialize a new classification head for each downstream task. However, for the Adapter-Vanilla model, we keep the existing classification head.

A.4 The Impact of Upstream Task Order

Both humans and animals demonstrate enhanced learning abilities when examples are presented in a deliberate sequence (Elman, 1993; Krueger and Dayan, 2009). Curriculum learning, a strategy involving the organized presentation of examples or tasks to expedite learning, has been proven to significantly influence the performance of neural models (Bengio et al., 2009). In the context of our proposed framework, a crucial question arises: to

[3] https://huggingface.co/sentence-transformers/paraphrase-xlm-r-multilingual-v1

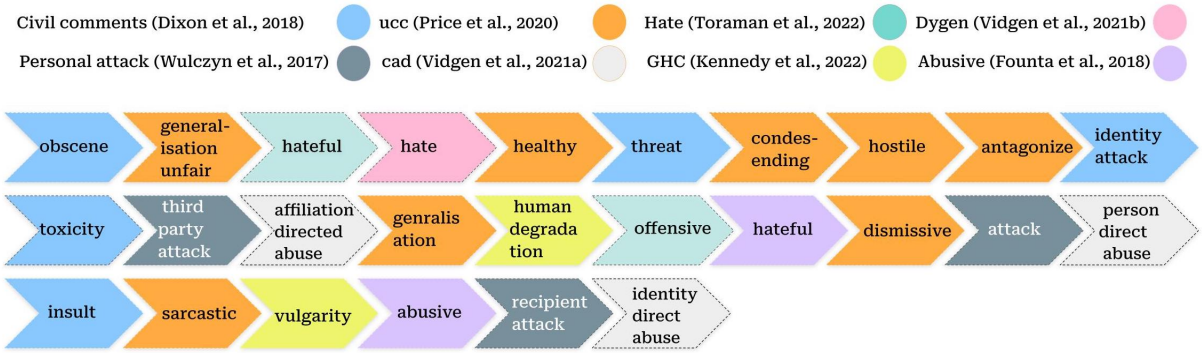

Figure 4: Random sequence of upstream tasks.

what extent does the sequence of upstream tasks impact the performance of different strategies on both upstream and downstream tasks? Furthermore, can we find the optimal ordering for upstream tasks? While the exhaustive exploration of these questions is beyond the scope of the current work, we investigate two alternative orders of upstream tasks. We emphasize that the modular design of our benchmarks allows for the effortless reordering of upstream tasks and facilitates seamless experimentation with curriculum learning. Specifically, we first modify our experiment in section 5 by keeping the upstream dataset order intact but modifying the order of tasks within each dataset. Additionally, we present results with a completely random order of tasks. Overall, these experiments show that BiHNet-Reg, the top-performing model in our main experiment, is also the least sensitive to task order, in comparison to other approaches. These results suggest BiHNet-Reg is a robust architecture for practical settings were the sequence of upstream tasks frequently evolves.

A.4.1 Chronological Upstream Datasets with Shuffled Tasks

In our chronological experiment, we initially assigned tasks within each dataset in a random order, as we lacked information regarding their precedence. To gauge the potential influence of the selected task sequence on our results, we train all model variations again but use an alternative random task order reshuffling while maintaining the dataset order intact. The sequence of upstream tasks in this experiment is illustrated in figure 3.

Our results reflect a similar pattern as the initial experiment (Table 6) Specifically, the few-shot AUC of BiHNet-Reg improves by nearly 2% compared to BiHNet-Vanilla, falling only 1.2% short of BiHnet-Adapter-Multitask. In terms of the final AUC, once again, BiHnet-Reg outperforms all sequential fine-tuning variations, and the instant AUC of all models falls within a close range. Overall, this experiment suggests that our proposed approach is robust to task perturbations within datasets. In other words, while the order of tasks within a dataset affects the resulting model's performance, the order of performance among different algorithms remains consistent.

A.4.2 Random Upstream Task Order

To show the efficacy of our proposed continual learning approach in adapting to any scenario, we randomly ordered the upstream tasks. Figure 4 shows upstream task sequence used in our experiments. Note that, we kept the dataset splits (i.e. train/dev/test) consistent with chronological experiment. This approach ensures that our comparison remains fair and valid, allowing for a meaningful assessment of the model's performance under the altered evaluation conditions. Overall, we observe similar performance patterns among the different algorithms, but the differences in performance are now less pronounced (Table 6). Below we discuss the results in detail;

Baselines: Interestingly, in this experiment, the Adapter-Vanilla baseline performs exceptionally well on downstream tasks despite achieving lower final performance. This could be attributed to the order of tasks, specifically the tasks at the end of the upstream. While this result might be favorable, the Adapter-Vanilla is not well-suited for practical settings where the of upstream tasks constantly evolve. This is evident from the high variations in the final and few-shot performance of the model across experiments.

Multitask Upperbound: The final and few-shot evaluation results for multitask models are displayed in the last two rows of table 6. It is impor-

	Method	Upstream		Downstream			
		Final	Instant	Few-shot	Δ Final	Δ Instant	Δ Few-shot
Chronological	Adapter-Vanilla	0.7648	0.8844	0.7568	—	—	—
	BiHNet-Vanilla	0.7594	0.8815	0.7865	-0.0054	-0.0031	+0.0297
	BiHNet-Reg	0.7963	0.8830	0.8043	+0.0315	-0.0014	+0.0475
	BiHNet-EWC	0.7513	0.8783	0.7702	-0.0135	-0.0061	+0.0134
Random Order	Adapter-Vanilla	0.6784	0.8859	0.8321	—	—	—
	BiHNet-Vanilla	0.7115	0.8838	0.8146	+0.0331	-0.0021	-0.0175
	BiHNet-Reg	0.7859	0.8846	0.8087	+0.1075	-0.0013	-0.0234
	BiHNet-EWC	0.6571	0.8863	0.8190	-0.0213	+0.0004	-0.0131
	Adapter-Multitask	0.8752	—	0.8531	—-	—	—
	BiHNet-Multitask	0.8321	—-	0.8215	—-	—	—

Table 6: Results in AUC for experiments with alternative upstream task order. Rows marked with "Chronological" show the results of experiments with chronologically ordered datasets but shuffled task orders within a dataset. Rows marked with "Random Order" show the results on complete random order of upstream tasks. The Δ values are computed in comparison to Adapter-Vanilla in each experiment. Notably, BiHNet+Reg demonstrates very stable performance regardless of the upstream task order.

tant to note that these models, having been exposed to all tasks simultaneously, do not have an instant performance metric defined for them.

Does the collection of problematic content tasks help with learning new upstream tasks? To address this inquiry, we can assess the immediate performance of a continual learning (CL) model when applied to $[T_1^u, T_2^u, ..., T_i^u]$ and compare it to a pretrained model fine-tuned exclusively on T_i^u. Our results (Δ Instance) show evidence of slight positive transfer, however, the magnitude of this transfer is negligible.

Does continual learning improve knowledge retention? The final AUC values, as shown in the first column of Table 6, indicate the models' ability to retain knowledge from a sequence of tasks at the end of training. Our results suggest that continual learning (BiHNet-Reg) outperforms naive training (BiHNet-Vanilla) by at least 0.07 in AUC, indicating its potential to mitigate catastrophic forgetting. However, BiHNet-Reg falls 0.04 short of the multitask counterpart. Further investigation is needed to understand this difference.

Does upstream learning help generalize new manifestations of problematic content? Comparing the single-task baselines with continual and multitask learning, our results demonstrate a noteworthy improvement in models' generalization ability due to upstream training.

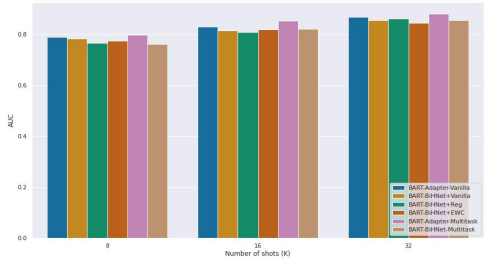

Figure 5: Few-shot performance (AUC) based on number of shots (K)

A.5 The Impact of Number of Shots in Downstream Adaptation

We performed a sensitivity analysis on the number of shots to examine how it affects our models. Specifically, we conducted few-shot training using 8, 16, and 32 shots. You can find the corresponding results in Figure 5. Our results show a consistent pattern; all models improve as the number of shots increases and the order between models stays the same. Interestingly, there is only one exception. BiHNet-Reg outperforms BiHNet-Vanilla with more shots. We leave further investigation of this effect is left for future work.

	Adapter-Multitask	Adapter-Vanilla	BiHnet-MultiTask	BiHNet-EWC	BiHNet-Reg
Adapter-Vanilla	0.015	-	-	-	-
BiHNet-Multitask	0.334	0.144	-	-	-
BiHNet-Reg	0.018*	0.955	0.159	-	-
BiHNet- EWC	0.916	0.012*	0.284	0.014*	-
BiHNet-Vanilla	0.037*	0.738	0.259	0.781	0.028*

Table 7: P values to pairwise T-test between the fewshot performances for experiments with the chronological order of upstream tasks.

A.6 Qualitative Analysis

We provide qualitative examples of texts correctly classified by the BIHNet-Reg and misclassified by adapter-vanilla below. Examples from CMSB dataset with sexism present.

- *This is the exact reason why Women shouldn't be involved. Not sexist. But situations like this will always be blown out of proportion.*
- *I'm not sexist, but women are inferior. proving that you can still be an idiot regardless of your "high IQ"*

Examples from CMSB dataset labeled as not sexist.

- *I'm not sensitive... But if in this modern era, a good adult is judged as one that pays the bills? A good adult is also one that can cook!*
- *I do not like dumb refs for football....*
- *Advice for adults: Think like an adult "act" like a pro*
- *I almost hate every song by any Southern country artist*

As demonstrated in the first two examples, BiHNet-Reg is able to correctly classify instances with a direct mention of "not sexist" but the vanilla model fails to do so. In the later examples, the vanilla model misclassifies texts that mention any gender stereotypes despite the fact that the mentions are not used in the context of gender.

A.7 Detailed Results

Below we provide detailed results, including AUC and F_1 scores, for all upstream and downstream tasks in our experiments. Specifically, tables 8 and 9 show detailed results for upstream training on experiments with chronological and random upstream task order. Table 10 and 11 provide detailed results on all downstream tasks for chronological and random upstream task order respectively.

Table 7 shows the p values for pairwise T-tests conducted on the fewshot AUC of various models. Our results indicate a significant difference between Adapter-Vanilla and BiHNet-Reg in downstream adaptation (i.e., few-shot). Furthermore, there is no significant difference between the BiHNet-Reg and Multitask models which are considered as the upper bounds. However, BiHNet-Reg significantly outperforms classic continual learning approaches such as EWC. These findings underscore the importance of developing continual learning approaches that have an emphasis on generalization as solutions to practical scenarios for dealing with the ever-evolving nature of problematic content.

order	dataset	task	model	final-f1	instant-f1	final-auc	instant-auc
0	personal-attack	a	BART-Adapter-Vanilla	0.305006	0.750957	0.540933	0.962732
0	personal-attack	a	BART-BiHNet+Vanilla	0.265491	0.749760	0.727950	0.957005
0	personal-attack	a	BART-BiHNet+Reg	0.743326	0.751853	0.956610	0.959739
0	personal-attack	a	BART-BiHNet+EWC	0.295845	0.737895	0.896827	0.954632
-	personal-attack	a	BART-Adapter-Multitask	0.703736	-	0.957941	-
-	personal-attack	a	BART-BiHNet-Multitask	0.747288	-	0.954593	-
1	personal-attack	tpa	BART-Adapter-Vanilla	0.062567	0.321267	0.461003	0.948346
1	personal-attack	tpa	BART-BiHNet+Vanilla	0.061224	0.296041	0.639220	0.938166
1	personal-attack	tpa	BART-BiHNet+Reg	0.093847	0.224464	0.884294	0.929268
1	personal-attack	tpa	BART-BiHNet+EWC	0.051033	0.275862	0.826217	0.924251
-	personal-attack	tpa	BART-Adapter-Multitask	0.311203	-	0.940889	-
-	personal-attack	tpa	BART-BiHNet-Multitask	0.100756	-	0.894707	-
2	personal-attack	ra	BART-Adapter-Vanilla	0.360275	0.722284	0.601506	0.968865
2	personal-attack	ra	BART-BiHNet+Vanilla	0.340474	0.729980	0.786089	0.969692
2	personal-attack	ra	BART-BiHNet+Reg	0.684231	0.712880	0.965032	0.968443
2	personal-attack	ra	BART-BiHNet+EWC	0.385978	0.733111	0.924772	0.968828
-	personal-attack	ra	BART-Adapter-Multitask	0.678799	-	0.970798	-
-	personal-attack	ra	BART-BiHNet-Multitask	0.682053	-	0.958645	-
3	jigsaw	threat	BART-Adapter-Vanilla	0.105263	0.099762	0.863857	0.987086
3	jigsaw	threat	BART-BiHNet+Vanilla	0.084746	0.119318	0.839035	0.983698
3	jigsaw	threat	BART-BiHNet+Reg	0.013133	0.130612	0.747348	0.983460
3	jigsaw	threat	BART-BiHNet+EWC	0.037736	0.086580	0.741358	0.986048
-	jigsaw	threat	BART-BiHNet-Multitask	0.031847	-	0.944563	-
-	jigsaw	threat	BART-Adapter-Multitask	0.067901	-	0.981188	-
4	jigsaw	insult	BART-Adapter-Vanilla	0.130737	0.560664	0.485944	0.948078
4	jigsaw	insult	BART-BiHNet+Vanilla	0.079646	0.548204	0.595428	0.943907
4	jigsaw	insult	BART-BiHNet+Reg	0.422827	0.556169	0.887573	0.944491
4	jigsaw	insult	BART-BiHNet+EWC	0.025848	0.586301	0.646605	0.944762
-	jigsaw	insult	BART-BiHNet-Multitask	0.483731	-	0.925866	-
-	jigsaw	insult	BART-Adapter-Multitask	0.496063	-	0.946926	-
5	jigsaw	toxicity	BART-Adapter-Vanilla	0.145535	0.569122	0.497406	0.937929
5	jigsaw	toxicity	BART-BiHNet+Vanilla	0.087558	0.575450	0.615918	0.930685
5	jigsaw	toxicity	BART-BiHNet+Reg	0.433930	0.576525	0.875425	0.933619
5	jigsaw	toxicity	BART-BiHNet+EWC	0.024783	0.545455	0.652824	0.934314
-	jigsaw	toxicity	BART-BiHNet-Multitask	0.552734	-	0.924050	-
-	jigsaw	toxicity	BART-Adapter-Multitask	0.495274	-	0.935170	-
6	jigsaw	identity-attack	BART-Adapter-Vanilla	0.053476	0.191682	0.542978	0.982650
6	jigsaw	identity-attack	BART-BiHNet+Vanilla	0.040000	0.173077	0.623106	0.981113
6	jigsaw	identity-attack	BART-BiHNet+Reg	0.041958	0.142012	0.822579	0.973798
6	jigsaw	identity-attack	BART-BiHNet+EWC	0.045977	0.160883	0.610788	0.982633
-	jigsaw	identity-attack	BART-BiHNet-Multitask	0.072587	-	0.918391	-
-	jigsaw	identity-attack	BART-Adapter-Multitask	0.166365	-	0.971636	-
7	jigsaw	obscene	BART-Adapter-Vanilla	0.045977	0.199513	0.422526	0.972589
7	jigsaw	obscene	BART-BiHNet+Vanilla	0	0.288973	0.676161	0.978831
7	jigsaw	obscene	BART-BiHNet+Reg	0.051030	0.156701	0.900776	0.968669
7	jigsaw	obscene	BART-BiHNet+EWC	0	0.171779	0.651265	0.976982
-	jigsaw	obscene	BART-BiHNet-Multitask	0.066298	-	0.949782	-
-	jigsaw	obscene	BART-Adapter-Multitask	0.113821	-	0.961654	-
8	abusive	abusive	BART-Adapter-Vanilla	0.044897	0.906134	0.165191	0.976826
8	abusive	abusive	BART-BiHNet+Vanilla	0.041800	0.904637	0.512914	0.974778

Continued on next page

Continued from previous page

order	dataset	task	model	final-f1	instant-f1	final-auc	instant-auc
8	abusive	abusive	BART-BiHNet+Reg	0.782107	0.906317	0.911893	0.974146
8	abusive	abusive	BART-BiHNet+EWC	0.032074	0.901350	0.686447	0.975405
-	abusive	abusive	BART-BiHNet-Multitask	0.871585	-	0.930225	-
-	abusive	abusive	BART-Adapter-Multitask	0.900779	-	0.973485	-
9	abusive	hateful	BART-Adapter-Vanilla	0.074675	0.392539	0.476841	0.862842
9	abusive	hateful	BART-BiHNet+Vanilla	0.067797	0.433871	0.591083	0.861912
9	abusive	hateful	BART-BiHNet+Reg	0.206936	0.391649	0.772192	0.857521
9	abusive	hateful	BART-BiHNet+EWC	0.080279	0.419483	0.724583	0.860007
-	abusive	hateful	BART-BiHNet-Multitask	0.188304	-	0.779141	-
-	abusive	hateful	BART-Adapter-Multitask	0.430430	-	0.832692	-
10	ghc	hd	BART-Adapter-Vanilla	0.183333	0.422484	0.522991	0.870717
10	ghc	hd	BART-BiHNet+Vanilla	0.138568	0.437736	0.607691	0.859721
10	ghc	hd	BART-BiHNet+Reg	0.370881	0.389571	0.839976	0.863519
10	ghc	hd	BART-BiHNet+EWC	0.062827	0.413381	0.701330	0.865431
-	ghc	hd	BART-Adapter-Multitask	0.422880	-	0.862535	-
-	ghc	hd	BART-BiHNet-Multitask	0.380296	-	0.836859	-
11	ghc	vo	BART-Adapter-Vanilla	0.223844	0.491176	0.542028	0.904490
11	ghc	vo	BART-BiHNet+Vanilla	0.168937	0.501466	0.675985	0.905508
11	ghc	vo	BART-BiHNet+Reg	0.325468	0.497396	0.850554	0.907117
11	ghc	vo	BART-BiHNet+EWC	0.089457	0.504425	0.737284	0.898693
-	ghc	vo	BART-Adapter-Multitask	0.461366	-	0.892245	-
-	ghc	vo	BART-BiHNet-Multitask	0.394544	-	0.863356	-
12	ucc	hostile	BART-Adapter-Vanilla	0.166667	0.209677	0.565535	0.847778
12	ucc	hostile	BART-BiHNet+Vanilla	0.058394	0.218274	0.582643	0.811822
12	ucc	hostile	BART-BiHNet+Reg	0.103139	0.205379	0.722313	0.852443
12	ucc	hostile	BART-BiHNet+EWC	0.018018	0.201754	0.614855	0.832668
-	ucc	hostile	BART-Adapter-Multitask	0.189474	-	0.819008	-
-	ucc	hostile	BART-BiHNet-Multitask	0.138947	-	0.773304	-
13	ucc	generalisation-unfair	BART-Adapter-Vanilla	0.082759	0.156250	0.449118	0.826243
13	ucc	generalisation-unfair	BART-BiHNet+Vanilla	0.079365	0.198925	0.542561	0.853333
13	ucc	generalisation-unfair	BART-BiHNet+Reg	0.140312	0.182796	0.839903	0.867649
13	ucc	generalisation-unfair	BART-BiHNet+EWC	0.040000	0.167173	0.641156	0.836470
-	ucc	generalisation-unfair	BART-Adapter-Multitask	0.184332	-	0.848067	-
-	ucc	generalisation-unfair	BART-BiHNet-Multitask	0.104545	-	0.768346	-
14	ucc	dismissive	BART-Adapter-Vanilla	0.100000	0.193750	0.601274	0.789372
14	ucc	dismissive	BART-BiHNet+Vanilla	0.032967	0.208333	0.564912	0.790613
14	ucc	dismissive	BART-BiHNet+Reg	0.103784	0.230624	0.642689	0.808192
14	ucc	dismissive	BART-BiHNet+EWC	0.012903	0.224543	0.594760	0.804139
-	ucc	dismissive	BART-Adapter-Multitask	0.240642	-	0.797518	-
-	ucc	dismissive	BART-BiHNet-Multitask	0.162362	-	0.741484	-
15	ucc	antagonize	BART-Adapter-Vanilla	0.095238	0.226455	0.553457	0.825648
15	ucc	antagonize	BART-BiHNet+Vanilla	0.018868	0.253859	0.571401	0.825812
15	ucc	antagonize	BART-BiHNet+Reg	0.154799	0.243506	0.711969	0.830656
15	ucc	antagonize	BART-BiHNet+EWC	0	0.244898	0.607714	0.831594
-	ucc	antagonize	BART-Adapter-Multitask	0.239080	-	0.789518	-
-	ucc	antagonize	BART-BiHNet-Multitask	0.182469	-	0.744412	-
16	ucc	condescending	BART-Adapter-Vanilla	0.067797	0.240994	0.537908	0.774688
16	ucc	condescending	BART-BiHNet+Vanilla	0.021739	0.250000	0.495190	0.776334
16	ucc	condescending	BART-BiHNet+Reg	0.137736	0.251880	0.631144	0.786145
16	ucc	condescending	BART-BiHNet+EWC	0.008000	0.246575	0.538720	0.759145

Continued on next page

Continued from previous page

order	dataset	task	model	final-f1	instant-f1	final-auc	instant-auc
-	ucc	condescending	BART-Adapter-Multitask	0.248175	-	0.759093	-
-	ucc	condescending	BART-BiHNet-Multitask	0.174603	-	0.700839	-
17	ucc	sarcastic	BART-Adapter-Vanilla	0.039683	0.146974	0.524464	0.697387
17	ucc	sarcastic	BART-BiHNet+Vanilla	0.017167	0.153846	0.521057	0.693673
17	ucc	sarcastic	BART-BiHNet+Reg	0.102000	0.173307	0.579365	0.707746
17	ucc	sarcastic	BART-BiHNet+EWC	0.009662	0.164539	0.489642	0.712868
-	ucc	sarcastic	BART-Adapter-Multitask	0.074349	-	0.664485	-
-	ucc	sarcastic	BART-BiHNet-Multitask	0.113014	-	0.629825	-
18	ucc	healthy	BART-Adapter-Vanilla	0.071247	0.247126	0.537509	0.727002
18	ucc	healthy	BART-BiHNet+Vanilla	0.026738	0.238141	0.567775	0.715351
18	ucc	healthy	BART-BiHNet+Reg	0.211990	0.250223	0.665776	0.716110
18	ucc	healthy	BART-BiHNet+EWC	0.005747	0.256209	0.575490	0.730077
-	ucc	healthy	BART-BiHNet-Multitask	0.180055	-	0.691591	-
-	ucc	healthy	BART-Adapter-Multitask	0.194357	-	0.701764	-
19	ucc	generalisation	BART-Adapter-Vanilla	0.078431	0.230530	0.453378	0.836325
19	ucc	generalisation	BART-BiHNet+Vanilla	0.074627	0.215730	0.544156	0.819732
19	ucc	generalisation	BART-BiHNet+Reg	0.152809	0.239700	0.835866	0.843875
19	ucc	generalisation	BART-BiHNet+EWC	0.037037	0.236111	0.642791	0.845400
-	ucc	generalisation	BART-BiHNet-Multitask	0.118451	-	0.763144	-
-	ucc	generalisation	BART-Adapter-Multitask	0.227642	-	0.832400	-
20	dygen	hate	BART-Adapter-Vanilla	0.162119	0.777707	0.556406	0.829644
20	dygen	hate	BART-BiHNet+Vanilla	0.107438	0.771440	0.536290	0.806773
20	dygen	hate	BART-BiHNet+Reg	0.618577	0.737288	0.667232	0.761661
20	dygen	hate	BART-BiHNet+EWC	0.058160	0.774381	0.520744	0.819920
-	dygen	hate	BART-Adapter-Multitask	0.732227	-	0.810266	-
-	dygen	hate	BART-BiHNet-Multitask	0.712602	-	0.759315	-
21	cad	persondirectedabuse	BART-Adapter-Vanilla	0.170492	0.411765	0.482379	0.867650
21	cad	persondirectedabuse	BART-BiHNet+Vanilla	0.120141	0.422330	0.574214	0.870717
21	cad	persondirectedabuse	BART-BiHNet+Reg	0.084211	0.408094	0.612836	0.880580
21	cad	persondirectedabuse	BART-BiHNet+EWC	0.114068	0.412698	0.659694	0.883367
-	cad	persondirectedabuse	BART-Adapter-Multitask	0.435216	-	0.893343	-
-	cad	persondirectedabuse	BART-BiHNet-Multitask	0.274268	-	0.811561	-
22	cad	identitydirectedabuse	BART-Adapter-Vanilla	0.127341	0.400000	0.531804	0.808435
22	cad	identitydirectedabuse	BART-BiHNet+Vanilla	0.097656	0.401665	0.575728	0.794394
22	cad	identitydirectedabuse	BART-BiHNet+Reg	0.146830	0.379535	0.566856	0.795727
22	cad	identitydirectedabuse	BART-BiHNet+EWC	0.085366	0.424365	0.599927	0.801885
-	cad	identitydirectedabuse	BART-Adapter-Multitask	0.362812	-	0.770885	-
-	cad	identitydirectedabuse	BART-BiHNet-Multitask	0.263666	-	0.729921	-
23	cad	affiliationdirectedabuse	BART-Adapter-Vanilla	0.069364	0.433613	0.380418	0.879725
23	cad	affiliationdirectedabuse	BART-BiHNet+Vanilla	0.098765	0.456835	0.524418	0.874797
23	cad	affiliationdirectedabuse	BART-BiHNet+Reg	0.423462	0.440514	0.846201	0.883356
23	cad	affiliationdirectedabuse	BART-BiHNet+EWC	0.073090	0.445652	0.562836	0.860070
-	cad	affiliationdirectedabuse	BART-Adapter-Multitask	0.402010	-	0.853287	-
-	cad	affiliationdirectedabuse	BART-BiHNet-Multitask	0.353623	-	0.807568	-
24	hate	offensive	BART-Adapter-Vanilla	0.094327	0.804835	0.391743	0.976767
24	hate	offensive	BART-BiHNet+Vanilla	0.064655	0.802737	0.643551	0.975667
24	hate	offensive	BART-BiHNet+Reg	0.143131	0.815429	0.898465	0.979279
24	hate	offensive	BART-BiHNet+EWC	0.041481	0.806264	0.818924	0.977839
-	hate	offensive	BART-Adapter-Multitask	0.791569	-	0.979571	-
-	hate	offensive	BART-BiHNet-Multitask	0.785226	-	0.966649	-

Continued on next page

Continued from previous page

order	dataset	task	model	final-f1	instant-f1	final-auc	instant-auc
25	hate	hateful	BART-Adapter-Vanilla	0.327078	0.347305	0.771206	0.927272
25	hate	hateful	BART-BiHNet+Vanilla	0.368421	0.377907	0.919701	0.946408
25	hate	hateful	BART-BiHNet+Reg	0.372951	0.395238	0.945276	0.946778
25	hate	hateful	BART-BiHNet+EWC	0.292308	0.344828	0.915899	0.938439
-	hate	hateful	BART-BiHNet-Multitask	0.143653	-	0.913388	-
-	hate	hateful	BART-Adapter-Multitask	0.382353	-	0.944837	-

Table 8: Final and instant AUC and F1 scores for upstream tasks for the chronological experiment

order	dataset	task	model	final-f1	instant-f1	final-auc	instant-auc
1	jigsaw	obscene	Adapter-Vanilla	0.020779	0.199005	0.634348	0.977025
1	jigsaw	obscene	BiHNet+Vanilla	0.026471	0.194175	0.726092	0.979034
1	jigsaw	obscene	BiHNet+Reg	0.117117	0.208877	0.946478	0.978208
1	jigsaw	obscene	BiHNet+EWC	0.035088	0.298387	0.649704	0.976722
-	jigsaw	obscene	Adapter-Multitask	0.202667	-	0.970656	-
-	jigsaw	obscene	BiHNet-Multitask	0.092511	-	0.944722	-
2	ucc	generalisation-unfair	Adapter-Vanilla	0.123967	0.256198	0.658976	0.860923
2	ucc	generalisation-unfair	BiHNet+Vanilla	0.107817	0.222222	0.706271	0.853472
2	ucc	generalisation-unfair	BiHNet+Reg	0.083832	0.206061	0.682753	0.860750
2	ucc	generalisation-unfair	BiHNet+EWC	0.105263	0.222841	0.653317	0.871959
-	ucc	generalisation-unfair	Adapter-Multitask	0.185714	-	0.838597	-
-	ucc	generalisation-unfair	BiHNet-Multitask	0.113861	-	0.707083	-
3	hate	hateful	Adapter-Vanilla	0.100000	0.396985	0.688817	0.940574
3	hate	hateful	BiHNet+Vanilla	0.080491	0.396450	0.693829	0.939336
3	hate	hateful	BiHNet+Reg	0.119177	0.334096	0.774023	0.940949
3	hate	hateful	BiHNet+EWC	0.071477	0.389423	0.544535	0.944195
-	hate	hateful	Adapter-Multitask	0.407692	-	0.960242	-
-	hate	hateful	BiHNet-Multitask	0.152436	-	0.914408	-
4	dygen	hate	Adapter-Vanilla	0.586525	0.772302	0.734833	0.828820
4	dygen	hate	BiHNet+Vanilla	0.637133	0.782263	0.706050	0.837907
4	dygen	hate	BiHNet+Reg	0.606033	0.748860	0.613006	0.762699
4	dygen	hate	BiHNet+EWC	0.547778	0.790928	0.706884	0.850217
-	dygen	hate	Adapter-Multitask	0.750575	-	0.819942	-
-	dygen	hate	BiHNet-Multitask	0.713164	-	0.760064	-
5	ucc	healthy	Adapter-Vanilla	0.089796	0.252822	0.607956	0.723211
5	ucc	healthy	BiHNet+Vanilla	0.130506	0.245672	0.607529	0.717350
5	ucc	healthy	BiHNet+Reg	0.200000	0.280778	0.680537	0.720583
5	ucc	healthy	BiHNet+EWC	0.124567	0.239151	0.602608	0.709075
-	ucc	healthy	Adapter-Multitask	0.224204	-	0.690258	-
-	ucc	healthy	BiHNet-Multitask	0.207002	-	0.690280	-
6	jigsaw	threat	Adapter-Vanilla	0.011019	0.123077	0.590772	0.987871
6	jigsaw	threat	BiHNet+Vanilla	0.006211	0.109375	0.693627	0.985106
6	jigsaw	threat	BiHNet+Reg	0.012539	0.095455	0.823852	0.989349
6	jigsaw	threat	BiHNet+EWC	0.008119	0.107969	0.606869	0.989180
-	jigsaw	threat	Adapter-Multitask	0.094808	-	0.980725	-
-	jigsaw	threat	BiHNet-Multitask	0.047511	-	0.947328	-
7	ucc	condescending	Adapter-Vanilla	0.056122	0.246080	0.569447	0.785604
7	ucc	condescending	BiHNet+Vanilla	0.084130	0.243767	0.570273	0.783299
7	ucc	condescending	BiHNet+Reg	0.162839	0.232461	0.646058	0.776889
7	ucc	condescending	BiHNet+EWC	0.098160	0.238443	0.587424	0.787313

Continued on next page

Continued from previous page

order	dataset	task	model	final-f1	instant-f1	final-auc	instant-auc
-	ucc	condescending	Adapter-Multitask	0.207407	-	0.746329	-
-	ucc	condescending	BiHNet-Multitask	0.169611	-	0.703610	-
8	ucc	hostile	Adapter-Vanilla	0.079051	0.210169	0.601122	0.837135
8	ucc	hostile	BiHNet+Vanilla	0.070652	0.193853	0.594370	0.813944
8	ucc	hostile	BiHNet+Reg	0.190476	0.213992	0.789258	0.855591
8	ucc	hostile	BiHNet+EWC	0.105572	0.206522	0.602163	0.831534
-	ucc	hostile	Adapter-Multitask	0.213198	-	0.828848	-
-	ucc	hostile	BiHNet-Multitask	0.150235	-	0.803156	-
9	ucc	antagonize	Adapter-Vanilla	0.085366	0.260870	0.627160	0.824417
9	ucc	antagonize	BiHNet+Vanilla	0.101545	0.239726	0.620268	0.823707
9	ucc	antagonize	BiHNet+Reg	0.200000	0.244275	0.760923	0.830485
9	ucc	antagonize	BiHNet+EWC	0.095465	0.259819	0.577579	0.803287
-	ucc	antagonize	Adapter-Multitask	0.201780	-	0.790624	-
-	ucc	antagonize	BiHNet-Multitask	0.187373	-	0.786051	-
10	jigsaw	identity-attack	Adapter-Vanilla	0.100503	0.213043	0.841030	0.979880
10	jigsaw	identity-attack	BiHNet+Vanilla	0.082739	0.241470	0.877627	0.982284
10	jigsaw	identity-attack	BiHNet+Reg	0.033691	0.223350	0.805231	0.982487
10	jigsaw	identity-attack	BiHNet+EWC	0.040332	0.232295	0.761092	0.981215
-	jigsaw	identity-attack	Adapter-Multitask	0.145833	-	0.973271	-
-	jigsaw	identity-attack	BiHNet-Multitask	0.085837	-	0.905618	-
11	jigsaw	toxicity	Adapter-Vanilla	0.177102	0.576288	0.686841	0.938429
11	jigsaw	toxicity	BiHNet+Vanilla	0.222537	0.580645	0.696388	0.935391
11	jigsaw	toxicity	BiHNet+Reg	0.552076	0.543160	0.918422	0.930403
11	jigsaw	toxicity	BiHNet+EWC	0.173575	0.577108	0.622396	0.937142
-	jigsaw	toxicity	Adapter-Multitask	0.573469	-	0.935680	-
-	jigsaw	toxicity	BiHNet-Multitask	0.552855	-	0.922125	-
12	personal-attack	tpa	Adapter-Vanilla	0.071197	0.365297	0.713532	0.949232
12	personal-attack	tpa	BiHNet+Vanilla	0.065125	0.357942	0.806359	0.912470
12	personal-attack	tpa	BiHNet+Reg	0.072626	0.366197	0.841588	0.934629
12	personal-attack	tpa	BiHNet+EWC	0.074959	0.364000	0.756201	0.930620
-	personal-attack	tpa	Adapter-Multitask	0.364035	-	0.947569	-
-	personal-attack	tpa	BiHNet-Multitask	0.105491	-	0.902844	-
13	cad	affiliationdirectedabuse	Adapter-Vanilla	0.148270	0.494845	0.618436	0.887943
13	cad	affiliationdirectedabuse	BiHNet+Vanilla	0.151282	0.470825	0.664817	0.888610
13	cad	affiliationdirectedabuse	BiHNet+Reg	0.129193	0.419682	0.643099	0.879390
13	cad	affiliationdirectedabuse	BiHNet+EWC	0.104972	0.502530	0.550398	0.908008
-	cad	affiliationdirectedabuse	Adapter-Multitask	0.449064	-	0.878271	-
-	cad	affiliationdirectedabuse	BiHNet-Multitask	0.317204	-	0.804172	-
14	ucc	generalisation	Adapter-Vanilla	0.120000	0.235897	0.660351	0.848341
14	ucc	generalisation	BiHNet+Vanilla	0.122016	0.237288	0.705748	0.859008
14	ucc	generalisation	BiHNet+Reg	0.096677	0.226164	0.685203	0.875159
14	ucc	generalisation	BiHNet+EWC	0.107955	0.232258	0.653448	0.874206
-	ucc	generalisation	Adapter-Multitask	0.219178	-	0.834728	-
-	ucc	generalisation	BiHNet-Multitask	0.125604	-	0.710813	-
15	ghc	hd	Adapter-Vanilla	0.351351	0.425131	0.763803	0.870509
15	ghc	hd	BiHNet+Vanilla	0.351544	0.443587	0.793900	0.879318
15	ghc	hd	BiHNet+Reg	0.308617	0.412698	0.780278	0.872039
15	ghc	hd	BiHNet+EWC	0.291815	0.428850	0.697856	0.878094
-	ghc	hd	Adapter-Multitask	0.391257	-	0.854813	-
-	ghc	hd	BiHNet-Multitask	0.363448	-	0.827565	-

Continued on next page

Continued from previous page

order	dataset	task	model	final-f1	instant-f1	final-auc	instant-auc
16	hate	offensive	Adapter-Vanilla	0.352511	0.802792	0.685974	0.978245
16	hate	offensive	BiHNet+Vanilla	0.371750	0.805515	0.720766	0.977552
16	hate	offensive	BiHNet+Reg	0.781868	0.785835	0.955545	0.979944
16	hate	offensive	BiHNet+EWC	0.373037	0.809084	0.594099	0.976769
-	hate	offensive	Adapter-Multitask	0.799446	-	0.976373	-
-	hate	offensive	BiHNet-Multitask	0.766355	-	0.962001	-
17	abusive	hateful	Adapter-Vanilla	0.270035	0.458667	0.763553	0.858683
17	abusive	hateful	BiHNet+Vanilla	0.278997	0.410728	0.770081	0.854976
17	abusive	hateful	BiHNet+Reg	0.165092	0.424520	0.666253	0.864749
17	abusive	hateful	BiHNet+EWC	0.275524	0.421230	0.728667	0.849809
-	abusive	hateful	Adapter-Multitask	0.420432	-	0.843342	-
-	abusive	hateful	BiHNet-Multitask	0.189639	-	0.774595	-
18	ucc	dismissive	Adapter-Vanilla	0.047138	0.235589	0.588588	0.825034
18	ucc	dismissive	BiHNet+Vanilla	0.060748	0.220994	0.591715	0.822811
18	ucc	dismissive	BiHNet+Reg	0.146835	0.207299	0.681038	0.819899
18	ucc	dismissive	BiHNet+EWC	0.065327	0.229508	0.576748	0.808745
-	ucc	dismissive	Adapter-Multitask	0.145923	-	0.801140	-
-	ucc	dismissive	BiHNet-Multitask	0.162839	-	0.769410	-
19	personal-attack	a	Adapter-Vanilla	0.430756	0.774558	0.797523	0.962485
19	personal-attack	a	BiHNet+Vanilla	0.519235	0.760917	0.857912	0.963369
19	personal-attack	a	BiHNet+Reg	0.733024	0.748555	0.947966	0.961693
19	personal-attack	a	BiHNet+EWC	0.455738	0.761735	0.829767	0.962449
-	personal-attack	a	Adapter-Multitask	0.755801	-	0.961488	-
-	personal-attack	a	BiHNet-Multitask	0.708326	-	0.950576	-
20	cad	persondirectedabuse	Adapter-Vanilla	0.116608	0.381703	0.589687	0.878956
20	cad	persondirectedabuse	BiHNet+Vanilla	0.165088	0.381356	0.637047	0.864690
20	cad	persondirectedabuse	BiHNet+Reg	0.141732	0.391681	0.609079	0.880668
20	cad	persondirectedabuse	BiHNet+EWC	0.139053	0.396552	0.569930	0.869262
-	cad	persondirectedabuse	Adapter-Multitask	0.381963	-	0.868548	-
-	cad	persondirectedabuse	BiHNet-Multitask	0.264045	-	0.801124	-
21	jigsaw	insult	Adapter-Vanilla	0.159140	0.548837	0.673561	0.951626
21	jigsaw	insult	BiHNet+Vanilla	0.168421	0.618182	0.663345	0.950417
21	jigsaw	insult	BiHNet+Reg	0.561667	0.525070	0.934685	0.949777
21	jigsaw	insult	BiHNet+EWC	0.134516	0.555082	0.589250	0.947814
-	jigsaw	insult	Adapter-Multitask	0.591755	-	0.948925	-
-	jigsaw	insult	BiHNet-Multitask	0.483471	-	0.916784	-
22	ucc	sarcastic	Adapter-Vanilla	0.051576	0.179817	0.537452	0.715202
22	ucc	sarcastic	BiHNet+Vanilla	0.058700	0.156682	0.535132	0.707973
22	ucc	sarcastic	BiHNet+Reg	0.090909	0.158956	0.632267	0.710375
22	ucc	sarcastic	BiHNet+EWC	0.090703	0.158163	0.615797	0.714295
-	ucc	sarcastic	Adapter-Multitask	0.115385	-	0.675992	-
-	ucc	sarcastic	BiHNet-Multitask	0.057582	-	0.590061	-
23	ghc	vo	Adapter-Vanilla	0.339791	0.474674	0.784665	0.893579
23	ghc	vo	BiHNet+Vanilla	0.333333	0.494453	0.810356	0.897837
23	ghc	vo	BiHNet+Reg	0.435155	0.471446	0.891330	0.899036
23	ghc	vo	BiHNet+EWC	0.324538	0.488114	0.735190	0.890318
-	ghc	vo	Adapter-Multitask	0.492221	-	0.902838	-
-	ghc	vo	BiHNet-Multitask	0.430180	-	0.887518	-
24	abusive	abusive	Adapter-Vanilla	0.237068	0.909381	0.637408	0.975141
24	abusive	abusive	BiHNet+Vanilla	0.296675	0.906077	0.784075	0.974635

Continued on next page

Continued from previous page

order	dataset	task	model	final-f1	instant-f1	final-auc	instant-auc
24	abusive	abusive	BiHNet+Reg	0.891249	0.897924	0.966972	0.972513
24	abusive	abusive	BiHNet+EWC	0.296176	0.905965	0.681150	0.975408
-	abusive	abusive	Adapter-Multitask	0.902729	-	0.974823	-
-	abusive	abusive	BiHNet-Multitask	0.868651	-	0.940765	-
25	personal-attack	ra	Adapter-Vanilla	0.439443	0.746765	0.822923	0.972592
25	personal-attack	ra	BiHNet+Vanilla	0.521540	0.750300	0.881657	0.974125
25	personal-attack	ra	BiHNet+Reg	0.728748	0.743187	0.966885	0.972671
25	personal-attack	ra	BiHNet+EWC	0.440975	0.741830	0.851548	0.974420
-	personal-attack	ra	Adapter-Multitask	0.728530	-	0.969852	-
-	personal-attack	ra	BiHNet-Multitask	0.668837	-	0.955089	-
26	cad	identitydirectedabuse	Adapter-Vanilla	0.349686	0.352399	0.759334	0.780956
26	cad	identitydirectedabuse	BiHNet+Vanilla	0.396285	0.405063	0.784712	0.800906
26	cad	identitydirectedabuse	BiHNet+Reg	0.390533	0.396292	0.791699	0.799686
26	cad	identitydirectedabuse	BiHNet+EWC	0.369469	0.390764	0.740702	0.802461
-	cad	identitydirectedabuse	Adapter-Multitask	0.369803	-	0.781649	-
-	cad	identitydirectedabuse	BiHNet-Multitask	0.292017	-	0.757460	-

Continued on next page

Table 9: Instant and final AUC and F1 scores for upstream tasks for the random order experiment

dataset	task	model	few-shot-auc	few-shot-f1
BAD2	-	BART-Adapter-Vanilla	0.626491	0.475584
BAD2	-	BART-BiHNet+Vanilla	0.591835	0.442589
BAD2	-	BART-BiHNet+Reg	0.627312	0.469799
BAD2	-	BART-BiHNet+EWC	0.624396	0.483940
BAD2	-	BART-Adapter-Multitask	0.643871	0.492441
BAD2	-	BART-BiHNet-Multitask	0.661902	0.482916
BAD4	-	BART-Adapter-Vanilla	0.590429	0.335484
BAD4	-	BART-BiHNet+Vanilla	0.560764	0.404692
BAD4	-	BART-BiHNet+Reg	0.591853	0.445521
BAD4	-	BART-BiHNet+EWC	0.623405	0.448454
BAD4	-	BART-Adapter-Multitask	0.628114	0.482385
BAD4	-	BART-BiHNet-Multitask	0.637908	0.474747
cad	counterspeech	BART-Adapter-Vanilla	0.947467	0.004090
cad	counterspeech	BART-BiHNet+Vanilla	0.940275	0.004717
cad	counterspeech	BART-BiHNet+Reg	0.994684	0.003210
cad	counterspeech	BART-BiHNet+EWC	0.890557	0.004376
cad	counterspeech	BART-Adapter-Multitask	0.973734	0.003040
cad	counterspeech	BART-BiHNet-Multitask	0.933083	0.004785
cmsb	sexist	BART-Adapter-Vanilla	0.800860	0.401189
cmsb	sexist	BART-BiHNet+Vanilla	0.791143	0.428305
cmsb	sexist	BART-BiHNet+Reg	0.847109	0.464678
cmsb	sexist	BART-BiHNet+EWC	0.788794	0.433862
cmsb	sexist	BART-Adapter-Multitask	0.838390	0.458685
cmsb	sexist	BART-BiHNet-Multitask	0.858623	0.487342
conan	disabled	BART-Adapter-Vanilla	0.904717	0.413793
conan	disabled	BART-BiHNet+Vanilla	0.971757	0.424242
conan	disabled	BART-BiHNet+Reg	0.970236	0.500000
conan	disabled	BART-BiHNet+EWC	0.964673	0.451613

Continued on next page

Continued from previous page

dataset	task	model	few-shot-auc	few-shot-f1
conan	disabled	BART-Adapter-Multitask	0.988589	0.555556
conan	disabled	BART-BiHNet-Multitask	0.932389	0.344262
conan	jews	BART-Adapter-Vanilla	0.929167	0.606452
conan	jews	BART-BiHNet+Vanilla	0.916136	0.563830
conan	jews	BART-BiHNet+Reg	0.986761	0.814286
conan	jews	BART-BiHNet+EWC	0.955000	0.658683
conan	jews	BART-Adapter-Multitask	0.971250	0.769231
conan	jews	BART-BiHNet-Multitask	0.911648	0.625000
conan	lgbt	BART-Adapter-Vanilla	0.826356	0.436975
conan	lgbt	BART-BiHNet+Vanilla	0.841163	0.455446
conan	lgbt	BART-BiHNet+Reg	0.890511	0.426230
conan	lgbt	BART-BiHNet+EWC	0.726521	0.318519
conan	lgbt	BART-Adapter-Multitask	0.876452	0.448430
conan	lgbt	BART-BiHNet-Multitask	0.864446	0.454148
conan	migrant	BART-Adapter-Vanilla	0.937178	0.787879
conan	migrant	BART-BiHNet+Vanilla	0.933143	0.764706
conan	migrant	BART-BiHNet+Reg	0.948523	0.783019
conan	migrant	BART-BiHNet+EWC	0.889955	0.616601
conan	migrant	BART-Adapter-Multitask	0.961840	0.833333
conan	migrant	BART-BiHNet-Multitask	0.925652	0.697248
conan	muslims	BART-Adapter-Vanilla	0.973152	0.869863
conan	muslims	BART-BiHNet+Vanilla	0.961423	0.807818
conan	muslims	BART-BiHNet+Reg	0.966340	0.835017
conan	muslims	BART-BiHNet+EWC	0.946108	0.762500
conan	muslims	BART-Adapter-Multitask	0.987032	0.880795
conan	muslims	BART-BiHNet-Multitask	0.953043	0.845361
conan	poc	BART-Adapter-Vanilla	0.705530	0.242105
conan	poc	BART-BiHNet+Vanilla	0.930292	0.492063
conan	poc	BART-BiHNet+Reg	0.848664	0.309524
conan	poc	BART-BiHNet+EWC	0.856897	0.400000
conan	poc	BART-Adapter-Multitask	0.907496	0.394737
conan	poc	BART-BiHNet-Multitask	0.757419	0.259740
conan	woman	BART-Adapter-Vanilla	0.945992	0.659091
conan	woman	BART-BiHNet+Vanilla	0.927384	0.629213
conan	woman	BART-BiHNet+Reg	0.921676	0.744828
conan	woman	BART-BiHNet+EWC	0.938102	0.608696
conan	woman	BART-Adapter-Multitask	0.982824	0.745562
conan	woman	BART-BiHNet-Multitask	0.898216	0.612022
dygen	african	BART-Adapter-Vanilla	0.697561	0.031546
dygen	african	BART-BiHNet+Vanilla	0.889696	0.043103
dygen	african	BART-BiHNet+Reg	0.822976	0.032895
dygen	african	BART-BiHNet+EWC	0.789526	0.028846
dygen	african	BART-Adapter-Multitask	0.791274	0.031496
dygen	african	BART-BiHNet-Multitask	0.894539	0.030848
dygen	animosity	BART-Adapter-Vanilla	0.545165	0.164412
dygen	animosity	BART-BiHNet+Vanilla	0.553239	0.164929
dygen	animosity	BART-BiHNet+Reg	0.556119	0.166000
dygen	animosity	BART-BiHNet+EWC	0.541385	0.156479
dygen	animosity	BART-Adapter-Multitask	0.528676	0.157377
dygen	animosity	BART-BiHNet-Multitask	0.577321	0.181818

Continued on next page

Continued from previous page

dataset	task	model	few-shot-auc	few-shot-f1
dygen	arab	BART-Adapter-Vanilla	0.706551	0.048900
dygen	arab	BART-BiHNet+Vanilla	0.684826	0.043584
dygen	arab	BART-BiHNet+Reg	0.771614	0.061776
dygen	arab	BART-BiHNet+EWC	0.673449	0.043222
dygen	arab	BART-Adapter-Multitask	0.720759	0.061135
dygen	arab	BART-BiHNet-Multitask	0.769525	0.055470
dygen	asi	BART-Adapter-Vanilla	0.722597	0.021341
dygen	asi	BART-BiHNet+Vanilla	0.602426	0.016985
dygen	asi	BART-BiHNet+Reg	0.680983	0.016416
dygen	asi	BART-BiHNet+EWC	0.639644	0.018154
dygen	asi	BART-Adapter-Multitask	0.637484	0.013106
dygen	asi	BART-BiHNet-Multitask	0.672150	0.018490
dygen	asi.chin	BART-Adapter-Vanilla	0.684886	0.040449
dygen	asi.chin	BART-BiHNet+Vanilla	0.822891	0.050505
dygen	asi.chin	BART-BiHNet+Reg	0.900363	0.057221
dygen	asi.chin	BART-BiHNet+EWC	0.740221	0.048408
dygen	asi.chin	BART-Adapter-Multitask	0.750432	0.040080
dygen	asi.chin	BART-BiHNet-Multitask	0.813962	0.046875
dygen	asi.east	BART-Adapter-Vanilla	0.599577	0.017668
dygen	asi.east	BART-BiHNet+Vanilla	0.719864	0.032698
dygen	asi.east	BART-BiHNet+Reg	0.792294	0.062257
dygen	asi.east	BART-BiHNet+EWC	0.738057	0.031034
dygen	asi.east	BART-Adapter-Multitask	0.566423	0.021692
dygen	asi.east	BART-BiHNet-Multitask	0.673008	0.022508
dygen	asi.south	BART-Adapter-Vanilla	0.694890	0.060086
dygen	asi.south	BART-BiHNet+Vanilla	0.670054	0.050000
dygen	asi.south	BART-BiHNet+Reg	0.820420	0.086275
dygen	asi.south	BART-BiHNet+EWC	0.669341	0.057803
dygen	asi.south	BART-Adapter-Multitask	0.804298	0.065906
dygen	asi.south	BART-BiHNet-Multitask	0.702177	0.055749
dygen	asylum	BART-Adapter-Vanilla	0.741776	0.010909
dygen	asylum	BART-BiHNet+Vanilla	0.818531	0.013187
dygen	asylum	BART-BiHNet+Reg	0.913690	0.026549
dygen	asylum	BART-BiHNet+EWC	0.704966	0.013015
dygen	asylum	BART-Adapter-Multitask	0.841792	0.011976
dygen	asylum	BART-BiHNet-Multitask	0.959743	0.027211
dygen	bla	BART-Adapter-Vanilla	0.663344	0.218642
dygen	bla	BART-BiHNet+Vanilla	0.676250	0.214612
dygen	bla	BART-BiHNet+Reg	0.783386	0.273713
dygen	bla	BART-BiHNet+EWC	0.662496	0.197213
dygen	bla	BART-Adapter-Multitask	0.743135	0.222460
dygen	bla	BART-BiHNet-Multitask	0.769149	0.235669
dygen	bla.man	BART-Adapter-Vanilla	0.843789	0.021505
dygen	bla.man	BART-BiHNet+Vanilla	0.853931	0.032680
dygen	bla.man	BART-BiHNet+Reg	0.913739	0.022346
dygen	bla.man	BART-BiHNet+EWC	0.826485	0.018116
dygen	bla.man	BART-Adapter-Multitask	0.914314	0.020374
dygen	bla.man	BART-BiHNet-Multitask	0.817650	0.019305
dygen	bla.wom	BART-Adapter-Vanilla	0.886206	0.046218
dygen	bla.wom	BART-BiHNet+Vanilla	0.713370	0.025974

Continued on next page

Continued from previous page

dataset	task	model	few-shot-auc	few-shot-f1
dygen	bla.wom	BART-BiHNet+Reg	0.865667	0.033537
dygen	bla.wom	BART-BiHNet+EWC	0.740031	0.033028
dygen	bla.wom	BART-Adapter-Multitask	0.869987	0.034321
dygen	bla.wom	BART-BiHNet-Multitask	0.796928	0.024691
dygen	dehumanization	BART-Adapter-Vanilla	0.763208	0.142857
dygen	dehumanization	BART-BiHNet+Vanilla	0.746079	0.151111
dygen	dehumanization	BART-BiHNet+Reg	0.790485	0.160643
dygen	dehumanization	BART-BiHNet+EWC	0.739724	0.129693
dygen	dehumanization	BART-Adapter-Multitask	0.723382	0.117130
dygen	dehumanization	BART-BiHNet-Multitask	0.727210	0.130159
dygen	derogation	BART-Adapter-Vanilla	0.589725	0.455206
dygen	derogation	BART-BiHNet+Vanilla	0.576981	0.459941
dygen	derogation	BART-BiHNet+Reg	0.651349	0.545455
dygen	derogation	BART-BiHNet+EWC	0.591059	0.495477
dygen	derogation	BART-Adapter-Multitask	0.596901	0.507422
dygen	derogation	BART-BiHNet-Multitask	0.692075	0.578187
dygen	dis	BART-Adapter-Vanilla	0.664966	0.094241
dygen	dis	BART-BiHNet+Vanilla	0.653491	0.087855
dygen	dis	BART-BiHNet+Reg	0.794327	0.111288
dygen	dis	BART-BiHNet+EWC	0.626324	0.085202
dygen	dis	BART-Adapter-Multitask	0.684887	0.091082
dygen	dis	BART-BiHNet-Multitask	0.726102	0.124748
dygen	for	BART-Adapter-Vanilla	0.833637	0.047970
dygen	for	BART-BiHNet+Vanilla	0.725930	0.039927
dygen	for	BART-BiHNet+Reg	0.929193	0.107023
dygen	for	BART-BiHNet+EWC	0.769685	0.036474
dygen	for	BART-Adapter-Multitask	0.832336	0.055202
dygen	for	BART-BiHNet-Multitask	0.903980	0.076372
dygen	gay	BART-Adapter-Vanilla	0.813890	0.081784
dygen	gay	BART-BiHNet+Vanilla	0.721734	0.075269
dygen	gay	BART-BiHNet+Reg	0.805713	0.076312
dygen	gay	BART-BiHNet+EWC	0.734685	0.079681
dygen	gay	BART-Adapter-Multitask	0.875041	0.097087
dygen	gay	BART-BiHNet-Multitask	0.826741	0.081169
dygen	gay.man	BART-Adapter-Vanilla	0.719518	0.056338
dygen	gay.man	BART-BiHNet+Vanilla	0.671613	0.050633
dygen	gay.man	BART-BiHNet+Reg	0.677750	0.039052
dygen	gay.man	BART-BiHNet+EWC	0.669622	0.044304
dygen	gay.man	BART-Adapter-Multitask	0.751199	0.047478
dygen	gay.man	BART-BiHNet-Multitask	0.669411	0.039216
dygen	gay.wom	BART-Adapter-Vanilla	0.653895	0.048780
dygen	gay.wom	BART-BiHNet+Vanilla	0.578229	0.037037
dygen	gay.wom	BART-BiHNet+Reg	0.682982	0.060302
dygen	gay.wom	BART-BiHNet+EWC	0.640716	0.039634
dygen	gay.wom	BART-Adapter-Multitask	0.696146	0.045296
dygen	gay.wom	BART-BiHNet-Multitask	0.763081	0.058027
dygen	gendermin	BART-Adapter-Vanilla	0.688054	0.024578
dygen	gendermin	BART-BiHNet+Vanilla	0.711625	0.021362
dygen	gendermin	BART-BiHNet+Reg	0.842811	0.029173
dygen	gendermin	BART-BiHNet+EWC	0.639510	0.021116

Continued on next page

Continued from previous page

dataset	task	model	few-shot-auc	few-shot-f1
dygen	gendermin	BART-Adapter-Multitask	0.880199	0.035587
dygen	gendermin	BART-BiHNet-Multitask	0.790749	0.029173
dygen	immig	BART-Adapter-Vanilla	0.743909	0.083019
dygen	immig	BART-BiHNet+Vanilla	0.781631	0.144828
dygen	immig	BART-BiHNet+Reg	0.821696	0.170492
dygen	immig	BART-BiHNet+EWC	0.708115	0.078704
dygen	immig	BART-Adapter-Multitask	0.840829	0.120000
dygen	immig	BART-BiHNet-Multitask	0.771645	0.093700
dygen	indig	BART-Adapter-Vanilla	0.817480	0.033195
dygen	indig	BART-BiHNet+Vanilla	0.718626	0.024263
dygen	indig	BART-BiHNet+Reg	0.800475	0.040201
dygen	indig	BART-BiHNet+EWC	0.847406	0.038278
dygen	indig	BART-Adapter-Multitask	0.917906	0.043689
dygen	indig	BART-BiHNet-Multitask	0.766115	0.022191
dygen	jew	BART-Adapter-Vanilla	0.786166	0.118902
dygen	jew	BART-BiHNet+Vanilla	0.781324	0.146597
dygen	jew	BART-BiHNet+Reg	0.846148	0.200000
dygen	jew	BART-BiHNet+EWC	0.839360	0.169133
dygen	jew	BART-Adapter-Multitask	0.784537	0.129713
dygen	jew	BART-BiHNet-Multitask	0.774725	0.106667
dygen	mixed.race	BART-Adapter-Vanilla	0.531906	0.019569
dygen	mixed.race	BART-BiHNet+Vanilla	0.646306	0.022857
dygen	mixed.race	BART-BiHNet+Reg	0.555626	0.017429
dygen	mixed.race	BART-BiHNet+EWC	0.611304	0.029412
dygen	mixed.race	BART-Adapter-Multitask	0.558827	0.016863
dygen	mixed.race	BART-BiHNet-Multitask	0.638592	0.023468
dygen	mus	BART-Adapter-Vanilla	0.755388	0.135472
dygen	mus	BART-BiHNet+Vanilla	0.797697	0.148014
dygen	mus	BART-BiHNet+Reg	0.765743	0.122754
dygen	mus	BART-BiHNet+EWC	0.772548	0.143113
dygen	mus	BART-Adapter-Multitask	0.816584	0.150289
dygen	mus	BART-BiHNet-Multitask	0.698485	0.104031
dygen	mus.wom	BART-Adapter-Vanilla	0.645392	0.016438
dygen	mus.wom	BART-BiHNet+Vanilla	0.717868	0.010417
dygen	mus.wom	BART-BiHNet+Reg	0.833229	0.014545
dygen	mus.wom	BART-BiHNet+EWC	0.736740	0.018059
dygen	mus.wom	BART-Adapter-Multitask	0.766520	0.016807
dygen	mus.wom	BART-BiHNet-Multitask	0.758558	0.012945
dygen	non.white	BART-Adapter-Vanilla	0.824000	0.061093
dygen	non.white	BART-BiHNet+Vanilla	0.696062	0.056604
dygen	non.white	BART-BiHNet+Reg	0.824159	0.070866
dygen	non.white	BART-BiHNet+EWC	0.801129	0.068100
dygen	non.white	BART-Adapter-Multitask	0.838850	0.058925
dygen	non.white	BART-BiHNet-Multitask	0.839195	0.076577
dygen	ref	BART-Adapter-Vanilla	0.834419	0.098039
dygen	ref	BART-BiHNet+Vanilla	0.868346	0.123348
dygen	ref	BART-BiHNet+Reg	0.788232	0.068966
dygen	ref	BART-BiHNet+EWC	0.814017	0.076923
dygen	ref	BART-Adapter-Multitask	0.908773	0.126482
dygen	ref	BART-BiHNet-Multitask	0.856012	0.095745

Continued on next page

Continued from previous page

dataset	task	model	few-shot-auc	few-shot-f1
dygen	support	BART-Adapter-Vanilla	0.606195	0.013962
dygen	support	BART-BiHNet+Vanilla	0.794912	0.060606
dygen	support	BART-BiHNet+Reg	0.451712	0.007207
dygen	support	BART-BiHNet+EWC	0.682239	0.016563
dygen	support	BART-Adapter-Multitask	0.696765	0.017021
dygen	support	BART-BiHNet-Multitask	0.740696	0.021645
dygen	threatening	BART-Adapter-Vanilla	0.852452	0.139013
dygen	threatening	BART-BiHNet+Vanilla	0.793205	0.112735
dygen	threatening	BART-BiHNet+Reg	0.798413	0.113725
dygen	threatening	BART-BiHNet+EWC	0.810625	0.136709
dygen	threatening	BART-Adapter-Multitask	0.882179	0.145631
dygen	threatening	BART-BiHNet-Multitask	0.866154	0.121008
dygen	trans	BART-Adapter-Vanilla	0.558231	0.096525
dygen	trans	BART-BiHNet+Vanilla	0.619845	0.106538
dygen	trans	BART-BiHNet+Reg	0.817006	0.146132
dygen	trans	BART-BiHNet+EWC	0.615229	0.093352
dygen	trans	BART-Adapter-Multitask	0.735171	0.135189
dygen	trans	BART-BiHNet-Multitask	0.714170	0.124077
dygen	trav	BART-Adapter-Vanilla	0.646662	0.020243
dygen	trav	BART-BiHNet+Vanilla	0.564392	0.021053
dygen	trav	BART-BiHNet+Reg	0.762115	0.029350
dygen	trav	BART-BiHNet+EWC	0.611448	0.023576
dygen	trav	BART-Adapter-Multitask	0.664540	0.028169
dygen	trav	BART-BiHNet-Multitask	0.606042	0.022814
dygen	wom	BART-Adapter-Vanilla	0.666830	0.191529
dygen	wom	BART-BiHNet+Vanilla	0.772368	0.252459
dygen	wom	BART-BiHNet+Reg	0.849288	0.369515
dygen	wom	BART-BiHNet+EWC	0.702072	0.194139
dygen	wom	BART-Adapter-Multitask	0.769987	0.248322
dygen	wom	BART-BiHNet-Multitask	0.757370	0.227474
ghc	cv	BART-Adapter-Vanilla	0.812127	0.062893
ghc	cv	BART-BiHNet+Vanilla	0.781179	0.062500
ghc	cv	BART-BiHNet+Reg	0.838447	0.060403
ghc	cv	BART-BiHNet+EWC	0.824924	0.062176
ghc	cv	BART-Adapter-Multitask	0.825069	0.072000
ghc	cv	BART-BiHNet-Multitask	0.818089	0.045977
hatecheck	black	BART-Adapter-Vanilla	0.789423	0.425000
hatecheck	black	BART-BiHNet+Vanilla	0.843558	0.496552
hatecheck	black	BART-BiHNet+Reg	0.931186	0.641791
hatecheck	black	BART-BiHNet+EWC	0.876891	0.448087
hatecheck	black	BART-Adapter-Multitask	0.926859	0.552632
hatecheck	black	BART-BiHNet-Multitask	0.856827	0.426230
hatecheck	disabled	BART-Adapter-Vanilla	0.886520	0.507463
hatecheck	disabled	BART-BiHNet+Vanilla	0.880580	0.624204
hatecheck	disabled	BART-BiHNet+Reg	0.954725	0.870968
hatecheck	disabled	BART-BiHNet+EWC	0.906063	0.584795
hatecheck	disabled	BART-Adapter-Multitask	0.965245	0.622222
hatecheck	disabled	BART-BiHNet-Multitask	0.894543	0.538462
hatecheck	gay	BART-Adapter-Vanilla	0.906400	0.512195
hatecheck	gay	BART-BiHNet+Vanilla	0.932067	0.615385

Continued on next page

Continued from previous page

dataset	task	model	few-shot-auc	few-shot-f1
hatecheck	gay	BART-BiHNet+Reg	0.902274	0.517647
hatecheck	gay	BART-BiHNet+EWC	0.890527	0.580645
hatecheck	gay	BART-Adapter-Multitask	0.959058	0.646617
hatecheck	gay	BART-BiHNet-Multitask	0.797588	0.413793
hatecheck	hate	BART-Adapter-Vanilla	0.779787	0.742597
hatecheck	hate	BART-BiHNet+Vanilla	0.711358	0.669704
hatecheck	hate	BART-BiHNet+Reg	0.745539	0.738854
hatecheck	hate	BART-BiHNet+EWC	0.768348	0.750000
hatecheck	hate	BART-Adapter-Multitask	0.822555	0.786957
hatecheck	hate	BART-BiHNet-Multitask	0.798437	0.806517
hatecheck	immigrants	BART-Adapter-Vanilla	0.862502	0.502857
hatecheck	immigrants	BART-BiHNet+Vanilla	0.919529	0.592593
hatecheck	immigrants	BART-BiHNet+Reg	0.915845	0.704000
hatecheck	immigrants	BART-BiHNet+EWC	0.842041	0.443114
hatecheck	immigrants	BART-Adapter-Multitask	0.930885	0.502732
hatecheck	immigrants	BART-BiHNet-Multitask	0.936488	0.615385
hatecheck	muslims	BART-Adapter-Vanilla	0.909837	0.617647
hatecheck	muslims	BART-BiHNet+Vanilla	0.929787	0.633094
hatecheck	muslims	BART-BiHNet+Reg	0.940720	0.588235
hatecheck	muslims	BART-BiHNet+EWC	0.923197	0.616438
hatecheck	muslims	BART-Adapter-Multitask	0.937066	0.544218
hatecheck	muslims	BART-BiHNet-Multitask	0.887850	0.545455
hatecheck	trans	BART-Adapter-Vanilla	0.751396	0.291339
hatecheck	trans	BART-BiHNet+Vanilla	0.891404	0.561644
hatecheck	trans	BART-BiHNet+Reg	0.940533	0.678899
hatecheck	trans	BART-BiHNet+EWC	0.825156	0.395939
hatecheck	trans	BART-Adapter-Multitask	0.876546	0.361991
hatecheck	trans	BART-BiHNet-Multitask	0.851881	0.454545
hatecheck	women	BART-Adapter-Vanilla	0.861084	0.485981
hatecheck	women	BART-BiHNet+Vanilla	0.941924	0.681319
hatecheck	women	BART-BiHNet+Reg	0.954110	0.747253
hatecheck	women	BART-BiHNet+EWC	0.948801	0.609524
hatecheck	women	BART-Adapter-Multitask	0.952622	0.646465
hatecheck	women	BART-BiHNet-Multitask	0.860923	0.374269
misogyny	-	BART-Adapter-Vanilla	0.803650	0.362264
misogyny	-	BART-BiHNet+Vanilla	0.814446	0.380567
misogyny	-	BART-BiHNet+Reg	0.853848	0.332248
misogyny	-	BART-BiHNet+EWC	0.817719	0.335766
misogyny	-	BART-Adapter-Multitask	0.858276	0.385185
misogyny	-	BART-BiHNet-Multitask	0.832160	0.341137
multi	-	BART-Adapter-Vanilla	0.643382	0.237037
multi	-	BART-BiHNet+Vanilla	0.631730	0.215385
multi	-	BART-BiHNet+Reg	0.592240	0.182062
multi	-	BART-BiHNet+EWC	0.575144	0.184080
multi	-	BART-Adapter-Multitask	0.632464	0.220779
multi	-	BART-BiHNet-Multitask	0.625541	0.218023
single	-	BART-Adapter-Vanilla	0.923063	0.618852
single	-	BART-BiHNet+Vanilla	0.909798	0.554622
single	-	BART-BiHNet+Reg	0.887218	0.483180
single	-	BART-BiHNet+EWC	0.904630	0.562162

Continued on next page

Continued from previous page

dataset	task	model	few-shot-auc	few-shot-f1
single	-	BART-Adapter-Multitask	0.958845	0.687747
single	-	BART-BiHNet-Multitask	0.869882	0.502370
single-adversarial	-	BART-Adapter-Vanilla	0.836229	0.521739
single-adversarial	-	BART-BiHNet+Vanilla	0.768038	0.366355
single-adversarial	-	BART-BiHNet+Reg	0.831907	0.490991
single-adversarial	-	BART-BiHNet+EWC	0.846279	0.459770
single-adversarial	-	BART-Adapter-Multitask	0.900268	0.592941
single-adversarial	-	BART-BiHNet-Multitask	0.797279	0.402367
stormfront	-	BART-Adapter-Vanilla	0.861921	0.794595
stormfront	-	BART-BiHNet+Vanilla	0.862494	0.740113
stormfront	-	BART-BiHNet+Reg	0.872769	0.779944
stormfront	-	BART-BiHNet+EWC	0.834097	0.774869
stormfront	-	BART-Adapter-Multitask	0.861880	0.776596
stormfront	-	BART-BiHNet-Multitask	0.865701	0.754617
us-election	hof	BART-Adapter-Vanilla	0.751050	0.293103
us-election	hof	BART-BiHNet+Vanilla	0.633272	0.225166
us-election	hof	BART-BiHNet+Reg	0.808955	0.385321
us-election	hof	BART-BiHNet+EWC	0.739496	0.278788
us-election	hof	BART-Adapter-Multitask	0.786699	0.333333
us-election	hof	BART-BiHNet-Multitask	0.792411	0.297030

Table 10: AUC and F1 scores for few-shot downstream tasks for the chronological experiment

dataset	task	model	few-shot-auc	few-shot-f1
BAD2	-	BART-Single	0.635964	0.490090
BAD2	-	BART-Adapter-Single	0.654797	0.483221
BAD2	-	BART-BiHNet-Single	0.620018	0.467909
BAD2	-	BART-Adapter-Vanilla	0.678801	0.475962
BAD2	-	BART-BiHNet+Vanilla	0.582984	0.435165
BAD2	-	BART-BiHNet+Reg	0.660194	0.491484
BAD2	-	BART-BiHNet+EWC	0.633916	0.470588
BAD2	-	BART-Adapter-Multitask	0.702097	0.514039
BAD2	-	BART-BiHNet-Multitask	0.714881	0.537445
BAD4	-	BART-Single	0.689085	0.469841
BAD4	-	BART-Adapter-Single	0.670554	0.455056
BAD4	-	BART-BiHNet-Single	0.661543	0.470270
BAD4	-	BART-Adapter-Vanilla	0.679876	0.468085
BAD4	-	BART-BiHNet+Vanilla	0.603978	0.454918
BAD4	-	BART-BiHNet+Reg	0.604742	0.447552
BAD4	-	BART-BiHNet+EWC	0.613064	0.438889
BAD4	-	BART-Adapter-Multitask	0.655514	0.455056
BAD4	-	BART-BiHNet-Multitask	0.639380	0.480447
CAD	counterspeech	BART-Single	0.622264	0.002805
CAD	counterspeech	BART-Adapter-Single	0.924328	0.004264
CAD	counterspeech	BART-BiHNet-Single	0.636023	0.002685
CAD	counterspeech	BART-Adapter-Vanilla	0.956223	0.005682
CAD	counterspeech	BART-BiHNet+Vanilla	0.988743	0.004640
CAD	counterspeech	BART-BiHNet+Reg	0.833646	0.003597
CAD	counterspeech	BART-BiHNet+EWC	0.950907	0.005013

Continued on next page

Continued from previous page

dataset	task	model	few-shot-auc	few-shot-f1
CAD	counterspeech	BART-Adapter-Multitask	0.988743	0.006369
CAD	counterspeech	BART-BiHNet-Multitask	0.931207	0.004535
CMSB	sexist	BART-Single	0.832720	0.494071
CMSB	sexist	BART-Adapter-Single	0.830289	0.483221
CMSB	sexist	BART-BiHNet-Single	0.819125	0.464088
CMSB	sexist	BART-Adapter-Vanilla	0.857568	0.510242
CMSB	sexist	BART-BiHNet+Vanilla	0.849790	0.509294
CMSB	sexist	BART-BiHNet+Reg	0.855256	0.515625
CMSB	sexist	BART-BiHNet+EWC	0.883429	0.549165
CMSB	sexist	BART-Adapter-Multitask	0.878635	0.531835
CMSB	sexist	BART-BiHNet-Multitask	0.843043	0.483926
CONAN	disabled	BART-Single	0.995150	0.851064
CONAN	disabled	BART-Adapter-Single	0.997623	0.933333
CONAN	disabled	BART-BiHNet-Single	0.995626	0.637681
CONAN	disabled	BART-Adapter-Vanilla	0.951217	0.478873
CONAN	disabled	BART-BiHNet+Vanilla	0.918315	0.357895
CONAN	disabled	BART-BiHNet+Reg	0.989730	0.458333
CONAN	disabled	BART-BiHNet+EWC	0.940044	0.535211
CONAN	disabled	BART-Adapter-Multitask	0.993343	0.666667
CONAN	disabled	BART-BiHNet-Multitask	0.897062	0.295082
CONAN	jews	BART-Single	0.994053	0.931034
CONAN	jews	BART-Adapter-Single	0.992500	0.890625
CONAN	jews	BART-BiHNet-Single	0.977670	0.775194
CONAN	jews	BART-Adapter-Vanilla	0.973902	0.734694
CONAN	jews	BART-BiHNet+Vanilla	0.931477	0.522936
CONAN	jews	BART-BiHNet+Reg	0.953617	0.627907
CONAN	jews	BART-BiHNet+EWC	0.960663	0.684932
CONAN	jews	BART-Adapter-Multitask	0.978371	0.839695
CONAN	jews	BART-BiHNet-Multitask	0.957102	0.548077
CONAN	LGBT	BART-Single	0.912992	0.543353
CONAN	LGBT	BART-Adapter-Single	0.935733	0.577540
CONAN	LGBT	BART-BiHNet-Single	0.895403	0.539326
CONAN	LGBT	BART-Adapter-Vanilla	0.925165	0.538071
CONAN	LGBT	BART-BiHNet+Vanilla	0.937694	0.533937
CONAN	LGBT	BART-BiHNet+Reg	0.889820	0.453125
CONAN	LGBT	BART-BiHNet+EWC	0.925165	0.537313
CONAN	LGBT	BART-Adapter-Multitask	0.937451	0.529412
CONAN	LGBT	BART-BiHNet-Multitask	0.854065	0.494253
CONAN	migrant	BART-Single	0.977594	0.897297
CONAN	migrant	BART-Adapter-Single	0.987959	0.913978
CONAN	migrant	BART-BiHNet-Single	0.983447	0.900000
CONAN	migrant	BART-Adapter-Vanilla	0.948639	0.789744
CONAN	migrant	BART-BiHNet+Vanilla	0.914204	0.663755
CONAN	migrant	BART-BiHNet+Reg	0.901016	0.653386
CONAN	migrant	BART-BiHNet+EWC	0.906675	0.669456
CONAN	migrant	BART-Adapter-Multitask	0.972875	0.841584
CONAN	migrant	BART-BiHNet-Multitask	0.922146	0.664000
CONAN	muslims	BART-Single	0.991436	0.877076
CONAN	muslims	BART-Adapter-Single	0.990668	0.907216
CONAN	muslims	BART-BiHNet-Single	0.992338	0.923077

Continued on next page

Continued from previous page

dataset	task	model	few-shot-auc	few-shot-f1
CONAN	muslims	BART-Adapter-Vanilla	0.991764	0.929577
CONAN	muslims	BART-BiHNet+Vanilla	0.987902	0.858065
CONAN	muslims	BART-BiHNet+Reg	0.957673	0.809211
CONAN	muslims	BART-BiHNet+EWC	0.972783	0.854237
CONAN	muslims	BART-Adapter-Multitask	0.993946	0.860841
CONAN	muslims	BART-BiHNet-Multitask	0.977792	0.787879
CONAN	people of color	BART-Single	0.885714	0.514851
CONAN	people of color	BART-Adapter-Single	0.959324	0.782609
CONAN	people of color	BART-BiHNet-Single	0.981198	0.777778
CONAN	people of color	BART-Adapter-Vanilla	0.898925	0.692308
CONAN	people of color	BART-BiHNet+Vanilla	0.929555	0.560748
CONAN	people of color	BART-BiHNet+Reg	0.889831	0.280374
CONAN	people of color	BART-BiHNet+EWC	0.903195	0.376623
CONAN	people of color	BART-Adapter-Multitask	0.935730	0.640000
CONAN	people of color	BART-BiHNet-Multitask	0.916190	0.528302
CONAN	woman	BART-Single	0.996055	0.870748
CONAN	woman	BART-Adapter-Single	0.998638	0.891892
CONAN	woman	BART-BiHNet-Single	0.995671	0.864865
CONAN	woman	BART-Adapter-Vanilla	0.986123	0.849315
CONAN	woman	BART-BiHNet+Vanilla	0.928379	0.645161
CONAN	woman	BART-BiHNet+Reg	0.980048	0.738095
CONAN	woman	BART-BiHNet+EWC	0.961720	0.648352
CONAN	woman	BART-Adapter-Multitask	0.994484	0.881119
CONAN	woman	BART-BiHNet-Multitask	0.971879	0.754717
Dygen	African	BART-Single	0.709622	0.022642
Dygen	African	BART-Adapter-Single	0.753744	0.023981
Dygen	African	BART-BiHNet-Single	0.807282	0.016970
Dygen	African	BART-Adapter-Vanilla	0.820106	0.036810
Dygen	African	BART-BiHNet+Vanilla	0.760201	0.021008
Dygen	African	BART-BiHNet+Reg	0.821272	0.027027
Dygen	African	BART-BiHNet+EWC	0.782441	0.036630
Dygen	African	BART-Adapter-Multitask	0.857950	0.040541
Dygen	African	BART-BiHNet-Multitask	0.860730	0.023256
Dygen	animosity	BART-Single	0.583085	0.180437
Dygen	animosity	BART-Adapter-Single	0.561059	0.176707
Dygen	animosity	BART-BiHNet-Single	0.506374	0.137174
Dygen	animosity	BART-Adapter-Vanilla	0.564928	0.176871
Dygen	animosity	BART-BiHNet+Vanilla	0.575415	0.191136
Dygen	animosity	BART-BiHNet+Reg	0.534618	0.168067
Dygen	animosity	BART-BiHNet+EWC	0.577934	0.175299
Dygen	animosity	BART-Adapter-Multitask	0.552231	0.168276
Dygen	animosity	BART-BiHNet-Multitask	0.607637	0.193622
Dygen	Arabs	BART-Single	0.635554	0.031128
Dygen	Arabs	BART-Adapter-Single	0.675253	0.039457
Dygen	Arabs	BART-BiHNet-Single	0.748829	0.062640
Dygen	Arabs	BART-Adapter-Vanilla	0.808592	0.076503
Dygen	Arabs	BART-BiHNet+Vanilla	0.735965	0.057851
Dygen	Arabs	BART-BiHNet+Reg	0.636772	0.048780
Dygen	Arabs	BART-BiHNet+EWC	0.801646	0.051051
Dygen	Arabs	BART-Adapter-Multitask	0.834051	0.078329

Continued on next page

Continued from previous page

dataset	task	model	few-shot-auc	few-shot-f1
Dygen	Arabs	BART-BiHNet-Multitask	0.719747	0.040161
Dygen	Asians	BART-Single	0.653580	0.034602
Dygen	Asians	BART-Adapter-Single	0.683574	0.029940
Dygen	Asians	BART-BiHNet-Single	0.688481	0.023437
Dygen	Asians	BART-Adapter-Vanilla	0.846577	0.024024
Dygen	Asians	BART-BiHNet+Vanilla	0.690327	0.016667
Dygen	Asians	BART-BiHNet+Reg	0.742070	0.018817
Dygen	Asians	BART-BiHNet+EWC	0.689384	0.016588
Dygen	Asians	BART-Adapter-Multitask	0.785647	0.016760
Dygen	Asians	BART-BiHNet-Multitask	0.641292	0.014134
Dygen	Chinese people	BART-Single	0.783270	0.044543
Dygen	Chinese people	BART-Adapter-Single	0.815762	0.050481
Dygen	Chinese people	BART-BiHNet-Single	0.812867	0.039356
Dygen	Chinese people	BART-Adapter-Vanilla	0.826175	0.044759
Dygen	Chinese people	BART-BiHNet+Vanilla	0.843012	0.060606
Dygen	Chinese people	BART-BiHNet+Reg	0.829689	0.057225
Dygen	Chinese people	BART-BiHNet+EWC	0.816698	0.052369
Dygen	Chinese people	BART-Adapter-Multitask	0.835825	0.042989
Dygen	Chinese people	BART-BiHNet-Multitask	0.809353	0.041339
Dygen	East Asians	BART-Single	0.692402	0.026871
Dygen	East Asians	BART-Adapter-Single	0.746267	0.024161
Dygen	East Asians	BART-BiHNet-Single	0.777790	0.061674
Dygen	East Asians	BART-Adapter-Vanilla	0.709308	0.034884
Dygen	East Asians	BART-BiHNet+Vanilla	0.760627	0.041667
Dygen	East Asians	BART-BiHNet+Reg	0.677499	0.039437
Dygen	East Asians	BART-BiHNet+EWC	0.703587	0.036000
Dygen	East Asians	BART-Adapter-Multitask	0.824933	0.038647
Dygen	East Asians	BART-BiHNet-Multitask	0.802792	0.036585
Dygen	South Asians	BART-Single	0.684706	0.050251
Dygen	South Asians	BART-Adapter-Single	0.665598	0.051583
Dygen	South Asians	BART-BiHNet-Single	0.662986	0.079365
Dygen	South Asians	BART-Adapter-Vanilla	0.780351	0.073702
Dygen	South Asians	BART-BiHNet+Vanilla	0.733631	0.074675
Dygen	South Asians	BART-BiHNet+Reg	0.747811	0.060790
Dygen	South Asians	BART-BiHNet+EWC	0.714140	0.061281
Dygen	South Asians	BART-Adapter-Multitask	0.738230	0.065574
Dygen	South Asians	BART-BiHNet-Multitask	0.723940	0.062874
Dygen	Asylum seekers	BART-Single	0.959743	0.053571
Dygen	Asylum seekers	BART-Adapter-Single	0.897400	0.021583
Dygen	Asylum seekers	BART-BiHNet-Single	0.786654	0.016854
Dygen	Asylum seekers	BART-Adapter-Vanilla	0.767387	0.013072
Dygen	Asylum seekers	BART-BiHNet+Vanilla	0.930999	0.016227
Dygen	Asylum seekers	BART-BiHNet+Reg	0.875705	0.013187
Dygen	Asylum seekers	BART-BiHNet+EWC	0.919956	0.019608
Dygen	Asylum seekers	BART-Adapter-Multitask	0.843828	0.022901
Dygen	Asylum seekers	BART-BiHNet-Multitask	0.956532	0.028777
Dygen	Black people	BART-Single	0.748573	0.219591
Dygen	Black people	BART-Adapter-Single	0.737509	0.248555
Dygen	Black people	BART-BiHNet-Single	0.727815	0.234192
Dygen	Black people	BART-Adapter-Vanilla	0.790263	0.255428

Continued on next page

Continued from previous page

dataset	task	model	few-shot-auc	few-shot-f1
Dygen	Black people	BART-BiHNet+Vanilla	0.739259	0.243959
Dygen	Black people	BART-BiHNet+Reg	0.735536	0.238202
Dygen	Black people	BART-BiHNet+EWC	0.711824	0.242321
Dygen	Black people	BART-Adapter-Multitask	0.753437	0.237248
Dygen	Black people	BART-BiHNet-Multitask	0.776706	0.230143
Dygen	Black men	BART-Single	0.970776	0.023669
Dygen	Black men	BART-Adapter-Single	0.818695	0.027397
Dygen	Black men	BART-BiHNet-Single	0.820316	0.023419
Dygen	Black men	BART-Adapter-Vanilla	0.912066	0.024390
Dygen	Black men	BART-BiHNet+Vanilla	0.908616	0.019640
Dygen	Black men	BART-BiHNet+Reg	0.908616	0.020374
Dygen	Black men	BART-BiHNet+EWC	0.986930	0.022989
Dygen	Black men	BART-Adapter-Multitask	0.950335	0.025157
Dygen	Black men	BART-BiHNet-Multitask	0.957340	0.024896
Dygen	Black women	BART-Single	0.796041	0.048193
Dygen	Black women	BART-Adapter-Single	0.844900	0.044444
Dygen	Black women	BART-BiHNet-Single	0.836289	0.039911
Dygen	Black women	BART-Adapter-Vanilla	0.814120	0.036735
Dygen	Black women	BART-BiHNet+Vanilla	0.828480	0.031936
Dygen	Black women	BART-BiHNet+Reg	0.815092	0.029605
Dygen	Black women	BART-BiHNet+EWC	0.796470	0.032454
Dygen	Black women	BART-Adapter-Multitask	0.825734	0.034156
Dygen	Black women	BART-BiHNet-Multitask	0.806968	0.037344
Dygen	dehumanization	BART-Single	0.703067	0.175439
Dygen	dehumanization	BART-Adapter-Single	0.653162	0.130233
Dygen	dehumanization	BART-BiHNet-Single	0.729720	0.130719
Dygen	dehumanization	BART-Adapter-Vanilla	0.803086	0.158654
Dygen	dehumanization	BART-BiHNet+Vanilla	0.726701	0.129524
Dygen	dehumanization	BART-BiHNet+Reg	0.730518	0.107981
Dygen	dehumanization	BART-BiHNet+EWC	0.775381	0.165450
Dygen	dehumanization	BART-Adapter-Multitask	0.839332	0.142649
Dygen	dehumanization	BART-BiHNet-Multitask	0.778659	0.107505
Dygen	derogation	BART-Single	0.514538	0.438830
Dygen	derogation	BART-Adapter-Single	0.511880	0.483633
Dygen	derogation	BART-BiHNet-Single	0.523676	0.464508
Dygen	derogation	BART-Adapter-Vanilla	0.705633	0.566964
Dygen	derogation	BART-BiHNet+Vanilla	0.702747	0.573463
Dygen	derogation	BART-BiHNet+Reg	0.632040	0.539097
Dygen	derogation	BART-BiHNet+EWC	0.706101	0.565619
Dygen	derogation	BART-Adapter-Multitask	0.702820	0.587181
Dygen	derogation	BART-BiHNet-Multitask	0.698568	0.566215
Dygen	People with disabilities	BART-Single	0.656806	0.092555
Dygen	People with disabilities	BART-Adapter-Single	0.683058	0.088962
Dygen	People with disabilities	BART-BiHNet-Single	0.672755	0.085106
Dygen	People with disabilities	BART-Adapter-Vanilla	0.764702	0.123404
Dygen	People with disabilities	BART-BiHNet+Vanilla	0.817699	0.201835
Dygen	People with disabilities	BART-BiHNet+Reg	0.760772	0.130536
Dygen	People with disabilities	BART-BiHNet+EWC	0.817542	0.156334
Dygen	People with disabilities	BART-Adapter-Multitask	0.765716	0.145631
Dygen	People with disabilities	BART-BiHNet-Multitask	0.719064	0.104167

Continued on next page

Continued from previous page

dataset	task	model	few-shot-auc	few-shot-f1
Dygen	Foreigners	BART-Single	0.865222	0.064368
Dygen	Foreigners	BART-Adapter-Single	0.884991	0.057034
Dygen	Foreigners	BART-BiHNet-Single	0.820148	0.054250
Dygen	Foreigners	BART-Adapter-Vanilla	0.916614	0.078313
Dygen	Foreigners	BART-BiHNet+Vanilla	0.910111	0.135266
Dygen	Foreigners	BART-BiHNet+Reg	0.785367	0.041420
Dygen	Foreigners	BART-BiHNet+EWC	0.908439	0.079027
Dygen	Foreigners	BART-Adapter-Multitask	0.907064	0.064240
Dygen	Foreigners	BART-BiHNet-Multitask	0.893594	0.065934
Dygen	gay	BART-Single	0.875634	0.130031
Dygen	gay	BART-Adapter-Single	0.833293	0.108911
Dygen	gay	BART-BiHNet-Single	0.795869	0.080495
Dygen	gay	BART-Adapter-Vanilla	0.856252	0.110843
Dygen	gay	BART-BiHNet+Vanilla	0.919566	0.111801
Dygen	gay	BART-BiHNet+Reg	0.876808	0.101053
Dygen	gay	BART-BiHNet+EWC	0.889835	0.104208
Dygen	gay	BART-Adapter-Multitask	0.892645	0.110132
Dygen	gay	BART-BiHNet-Multitask	0.818323	0.065341
Dygen	Gay men	BART-Single	0.654332	0.042169
Dygen	Gay men	BART-Adapter-Single	0.645526	0.038633
Dygen	Gay men	BART-BiHNet-Single	0.614690	0.031835
Dygen	Gay men	BART-Adapter-Vanilla	0.756145	0.052142
Dygen	Gay men	BART-BiHNet+Vanilla	0.759221	0.043302
Dygen	Gay men	BART-BiHNet+Reg	0.737160	0.048696
Dygen	Gay men	BART-BiHNet+EWC	0.748153	0.049689
Dygen	Gay men	BART-Adapter-Multitask	0.796858	0.055738
Dygen	Gay men	BART-BiHNet-Multitask	0.700956	0.042003
Dygen	Gay women	BART-Single	0.575847	0.035961
Dygen	Gay women	BART-Adapter-Single	0.558069	0.028694
Dygen	Gay women	BART-BiHNet-Single	0.553636	0.032258
Dygen	Gay women	BART-Adapter-Vanilla	0.768479	0.061176
Dygen	Gay women	BART-BiHNet+Vanilla	0.740930	0.059754
Dygen	Gay women	BART-BiHNet+Reg	0.744051	0.082474
Dygen	Gay women	BART-BiHNet+EWC	0.635514	0.056206
Dygen	Gay women	BART-Adapter-Multitask	0.799903	0.061758
Dygen	Gay women	BART-BiHNet-Multitask	0.731807	0.038647
Dygen	Gender minorities	BART-Single	0.852108	0.030905
Dygen	Gender minorities	BART-Adapter-Single	0.795011	0.035794
Dygen	Gender minorities	BART-BiHNet-Single	0.778906	0.027231
Dygen	Gender minorities	BART-Adapter-Vanilla	0.868471	0.035461
Dygen	Gender minorities	BART-BiHNet+Vanilla	0.730162	0.022670
Dygen	Gender minorities	BART-BiHNet+Reg	0.780365	0.021053
Dygen	Gender minorities	BART-BiHNet+EWC	0.871331	0.031696
Dygen	Gender minorities	BART-Adapter-Multitask	0.868585	0.031949
Dygen	Gender minorities	BART-BiHNet-Multitask	0.761714	0.025641
Dygen	Immigrants	BART-Single	0.906365	0.182456
Dygen	Immigrants	BART-Adapter-Single	0.845723	0.180602
Dygen	Immigrants	BART-BiHNet-Single	0.780274	0.090909
Dygen	Immigrants	BART-Adapter-Vanilla	0.811105	0.095552
Dygen	Immigrants	BART-BiHNet+Vanilla	0.809194	0.103448

Continued on next page

Continued from previous page

dataset	task	model	few-shot-auc	few-shot-f1
Dygen	Immigrants	BART-BiHNet+Reg	0.808537	0.129032
Dygen	Immigrants	BART-BiHNet+EWC	0.785020	0.089783
Dygen	Immigrants	BART-Adapter-Multitask	0.816552	0.092399
Dygen	Immigrants	BART-BiHNet-Multitask	0.815099	0.088685
Dygen	indig	BART-Single	0.743278	0.040000
Dygen	indig	BART-Adapter-Single	0.864376	0.050955
Dygen	indig	BART-BiHNet-Single	0.879705	0.029126
Dygen	indig	BART-Adapter-Vanilla	0.825127	0.026616
Dygen	indig	BART-BiHNet+Vanilla	0.849291	0.029316
Dygen	indig	BART-BiHNet+Reg	0.845764	0.029474
Dygen	indig	BART-BiHNet+EWC	0.774495	0.026316
Dygen	indig	BART-Adapter-Multitask	0.800475	0.027273
Dygen	indig	BART-BiHNet-Multitask	0.869300	0.035635
Dygen	Jewish people	BART-Single	0.695314	0.117871
Dygen	Jewish people	BART-Adapter-Single	0.660048	0.091097
Dygen	Jewish people	BART-BiHNet-Single	0.692381	0.126531
Dygen	Jewish people	BART-Adapter-Vanilla	0.859924	0.156352
Dygen	Jewish people	BART-BiHNet+Vanilla	0.770853	0.158664
Dygen	Jewish people	BART-BiHNet+Reg	0.782482	0.129760
Dygen	Jewish people	BART-BiHNet+EWC	0.788038	0.141491
Dygen	Jewish people	BART-Adapter-Multitask	0.819858	0.139384
Dygen	Jewish people	BART-BiHNet-Multitask	0.833787	0.126856
Dygen	Mixed race	BART-Single	0.568220	0.017316
Dygen	Mixed race	BART-Adapter-Single	0.592517	0.017544
Dygen	Mixed race	BART-BiHNet-Single	0.497586	0.014388
Dygen	Mixed race	BART-Adapter-Vanilla	0.699045	0.034146
Dygen	Mixed race	BART-BiHNet+Vanilla	0.586744	0.019444
Dygen	Mixed race	BART-BiHNet+Reg	0.682698	0.028571
Dygen	Mixed race	BART-BiHNet+EWC	0.636702	0.019116
Dygen	Mixed race	BART-Adapter-Multitask	0.694742	0.028807
Dygen	Mixed race	BART-BiHNet-Multitask	0.671599	0.026906
Dygen	Muslims	BART-Single	0.789611	0.106996
Dygen	Muslims	BART-Adapter-Single	0.790257	0.120055
Dygen	Muslims	BART-BiHNet-Single	0.739825	0.125000
Dygen	Muslims	BART-Adapter-Vanilla	0.846611	0.152727
Dygen	Muslims	BART-BiHNet+Vanilla	0.806735	0.122503
Dygen	Muslims	BART-BiHNet+Reg	0.834092	0.191919
Dygen	Muslims	BART-BiHNet+EWC	0.774975	0.142574
Dygen	Muslims	BART-Adapter-Multitask	0.879749	0.168297
Dygen	Muslims	BART-BiHNet-Multitask	0.817724	0.119948
Dygen	Muslim women	BART-Single	0.714734	0.018265
Dygen	Muslim women	BART-Adapter-Single	0.722132	0.021277
Dygen	Muslim women	BART-BiHNet-Single	0.877367	0.023256
Dygen	Muslim women	BART-Adapter-Vanilla	0.686270	0.017143
Dygen	Muslim women	BART-BiHNet+Vanilla	0.619937	0.009756
Dygen	Muslim women	BART-BiHNet+Reg	0.815172	0.015083
Dygen	Muslim women	BART-BiHNet+EWC	0.939060	0.020367
Dygen	Muslim women	BART-Adapter-Multitask	0.908840	0.031250
Dygen	Muslim women	BART-BiHNet-Multitask	0.753292	0.010152
Dygen	Non-whites	BART-Single	0.862599	0.095541

Continued on next page

Continued from previous page

dataset	task	model	few-shot-auc	few-shot-f1
Dygen	Non-whites	BART-Adapter-Single	0.851783	0.078534
Dygen	Non-whites	BART-BiHNet-Single	0.824677	0.070796
Dygen	Non-whites	BART-Adapter-Vanilla	0.827832	0.070640
Dygen	Non-whites	BART-BiHNet+Vanilla	0.862513	0.070764
Dygen	Non-whites	BART-BiHNet+Reg	0.880372	0.077079
Dygen	Non-whites	BART-BiHNet+EWC	0.805565	0.069930
Dygen	Non-whites	BART-Adapter-Multitask	0.888207	0.093923
Dygen	Non-whites	BART-BiHNet-Multitask	0.853555	0.071571
Dygen	Refguees	BART-Single	0.942489	0.223529
Dygen	Refguees	BART-Adapter-Single	0.909316	0.142857
Dygen	Refguees	BART-BiHNet-Single	0.827890	0.063670
Dygen	Refguees	BART-Adapter-Vanilla	0.887150	0.125461
Dygen	Refguees	BART-BiHNet+Vanilla	0.888220	0.091603
Dygen	Refguees	BART-BiHNet+Reg	0.802226	0.082418
Dygen	Refguees	BART-BiHNet+EWC	0.845984	0.080201
Dygen	Refguees	BART-Adapter-Multitask	0.898429	0.143426
Dygen	Refguees	BART-BiHNet-Multitask	0.867457	0.107595
Dygen	support	BART-Single	0.730528	0.023256
Dygen	support	BART-Adapter-Single	0.663866	0.021277
Dygen	support	BART-BiHNet-Single	0.615421	0.012780
Dygen	support	BART-Adapter-Vanilla	0.549388	0.009479
Dygen	support	BART-BiHNet+Vanilla	0.568507	0.012005
Dygen	support	BART-BiHNet+Reg	0.537178	0.010194
Dygen	support	BART-BiHNet+EWC	0.528541	0.011655
Dygen	support	BART-Adapter-Multitask	0.636856	0.017167
Dygen	support	BART-BiHNet-Multitask	0.669362	0.024768
Dygen	threatening	BART-Single	0.875585	0.177650
Dygen	threatening	BART-Adapter-Single	0.836170	0.138889
Dygen	threatening	BART-BiHNet-Single	0.790577	0.108659
Dygen	threatening	BART-Adapter-Vanilla	0.901731	0.130360
Dygen	threatening	BART-BiHNet+Vanilla	0.835296	0.099010
Dygen	threatening	BART-BiHNet+Reg	0.712324	0.081425
Dygen	threatening	BART-BiHNet+EWC	0.864872	0.123077
Dygen	threatening	BART-Adapter-Multitask	0.893550	0.140152
Dygen	threatening	BART-BiHNet-Multitask	0.860865	0.109546
Dygen	Trans people	BART-Single	0.694125	0.134293
Dygen	Trans people	BART-Adapter-Single	0.729872	0.150538
Dygen	Trans people	BART-BiHNet-Single	0.687860	0.119816
Dygen	Trans people	BART-Adapter-Vanilla	0.748769	0.160584
Dygen	Trans people	BART-BiHNet+Vanilla	0.765517	0.127080
Dygen	Trans people	BART-BiHNet+Reg	0.764915	0.123810
Dygen	Trans people	BART-BiHNet+EWC	0.790838	0.161100
Dygen	Trans people	BART-Adapter-Multitask	0.803334	0.166329
Dygen	Trans people	BART-BiHNet-Multitask	0.747644	0.122754
Dygen	Travellers	BART-Single	0.669575	0.021668
Dygen	Travellers	BART-Adapter-Single	0.706848	0.023585
Dygen	Travellers	BART-BiHNet-Single	0.766577	0.032941
Dygen	Travellers	BART-Adapter-Vanilla	0.670805	0.028169
Dygen	Travellers	BART-BiHNet+Vanilla	0.697895	0.026465
Dygen	Travellers	BART-BiHNet+Reg	0.653241	0.020654

Continued on next page

Continued from previous page

dataset	task	model	few-shot-auc	few-shot-f1
Dygen	Travellers	BART-BiHNet+EWC	0.734996	0.027184
Dygen	Travellers	BART-Adapter-Multitask	0.649694	0.026144
Dygen	Travellers	BART-BiHNet-Multitask	0.741318	0.022508
Dygen	Women	BART-Single	0.756641	0.218409
Dygen	Women	BART-Adapter-Single	0.852057	0.308998
Dygen	Women	BART-BiHNet-Single	0.825839	0.273973
Dygen	Women	BART-Adapter-Vanilla	0.841440	0.317797
Dygen	Women	BART-BiHNet+Vanilla	0.834226	0.322457
Dygen	Women	BART-BiHNet+Reg	0.828297	0.278317
Dygen	Women	BART-BiHNet+EWC	0.818255	0.274834
Dygen	Women	BART-Adapter-Multitask	0.858158	0.344423
Dygen	Women	BART-BiHNet-Multitask	0.791889	0.276094
GHC	class for violence	BART-Single	0.641220	0.035088
GHC	class for violence	BART-Adapter-Single	0.631671	0.026230
GHC	class for violence	BART-BiHNet-Single	0.627405	0.026906
GHC	class for violence	BART-Adapter-Vanilla	0.795453	0.042781
GHC	class for violence	BART-BiHNet+Vanilla	0.728225	0.034115
GHC	class for violence	BART-BiHNet+Reg	0.789855	0.047244
GHC	class for violence	BART-BiHNet+EWC	0.757210	0.042827
GHC	class for violence	BART-Adapter-Multitask	0.822064	0.052786
GHC	class for violence	BART-BiHNet-Multitask	0.847850	0.055980
hatecheck	black	BART-Single	0.967115	0.946237
hatecheck	black	BART-Adapter-Single	0.956154	0.868687
hatecheck	black	BART-BiHNet-Single	0.934679	0.831683
hatecheck	black	BART-Adapter-Vanilla	0.944744	0.582781
hatecheck	black	BART-BiHNet+Vanilla	0.956763	0.756303
hatecheck	black	BART-BiHNet+Reg	0.966314	0.671642
hatecheck	black	BART-BiHNet+EWC	0.930929	0.480874
hatecheck	black	BART-Adapter-Multitask	0.928526	0.604317
hatecheck	black	BART-BiHNet-Multitask	0.956154	0.573171
hatecheck	disabled	BART-Single	0.990839	0.836066
hatecheck	disabled	BART-Adapter-Single	0.985898	0.802920
hatecheck	disabled	BART-BiHNet-Single	0.924412	0.571429
hatecheck	disabled	BART-Adapter-Vanilla	0.993782	0.735484
hatecheck	disabled	BART-BiHNet+Vanilla	0.983344	0.666667
hatecheck	disabled	BART-BiHNet+Reg	0.991395	0.741722
hatecheck	disabled	BART-BiHNet+EWC	0.984177	0.750000
hatecheck	disabled	BART-Adapter-Multitask	0.997058	0.881890
hatecheck	disabled	BART-BiHNet-Multitask	0.941039	0.560847
hatecheck	gay	BART-Single	0.972348	0.777778
hatecheck	gay	BART-Adapter-Single	0.956538	0.687500
hatecheck	gay	BART-BiHNet-Single	0.907722	0.537500
hatecheck	gay	BART-Adapter-Vanilla	0.968758	0.739496
hatecheck	gay	BART-BiHNet+Vanilla	0.953200	0.560510
hatecheck	gay	BART-BiHNet+Reg	0.942996	0.578947
hatecheck	gay	BART-BiHNet+EWC	0.918588	0.552632
hatecheck	gay	BART-Adapter-Multitask	0.947909	0.701754
hatecheck	gay	BART-BiHNet-Multitask	0.864985	0.450262
hatecheck	hate	BART-Single	0.701328	0.430678
hatecheck	hate	BART-Adapter-Single	0.717094	0.474286

Continued on next page

Continued from previous page

dataset	task	model	few-shot-auc	few-shot-f1
hatecheck	hate	BART-BiHNet-Single	0.727140	0.569231
hatecheck	hate	BART-Adapter-Vanilla	0.815678	0.795876
hatecheck	hate	BART-BiHNet+Vanilla	0.795384	0.836852
hatecheck	hate	BART-BiHNet+Reg	0.764893	0.836364
hatecheck	hate	BART-BiHNet+EWC	0.820777	0.839552
hatecheck	hate	BART-Adapter-Multitask	0.902120	0.834061
hatecheck	hate	BART-BiHNet-Multitask	0.846102	0.869718
hatecheck	immigrants	BART-Single	0.979479	0.890909
hatecheck	immigrants	BART-Adapter-Single	0.971380	0.857143
hatecheck	immigrants	BART-BiHNet-Single	0.939898	0.708661
hatecheck	immigrants	BART-Adapter-Vanilla	0.932347	0.637037
hatecheck	immigrants	BART-BiHNet+Vanilla	0.968518	0.750000
hatecheck	immigrants	BART-BiHNet+Reg	0.937097	0.634483
hatecheck	immigrants	BART-BiHNet+EWC	0.924857	0.600000
hatecheck	immigrants	BART-Adapter-Multitask	0.968822	0.702290
hatecheck	immigrants	BART-BiHNet-Multitask	0.971897	0.779661
hatecheck	muslims	BART-Single	0.958333	0.714286
hatecheck	muslims	BART-Adapter-Single	0.969806	0.643357
hatecheck	muslims	BART-BiHNet-Single	0.912862	0.558659
hatecheck	muslims	BART-Adapter-Vanilla	0.961359	0.647482
hatecheck	muslims	BART-BiHNet+Vanilla	0.935897	0.620690
hatecheck	muslims	BART-BiHNet+Reg	0.931943	0.656934
hatecheck	muslims	BART-BiHNet+EWC	0.888779	0.523256
hatecheck	muslims	BART-Adapter-Multitask	0.973820	0.717557
hatecheck	muslims	BART-BiHNet-Multitask	0.903157	0.517241
hatecheck	Trans people	BART-Single	0.937442	0.876404
hatecheck	Trans people	BART-Adapter-Single	0.923348	0.716981
hatecheck	Trans people	BART-BiHNet-Single	0.903304	0.645669
hatecheck	Trans people	BART-Adapter-Vanilla	0.935780	0.491018
hatecheck	Trans people	BART-BiHNet+Vanilla	0.916933	0.515723
hatecheck	Trans people	BART-BiHNet+Reg	0.922750	0.557823
hatecheck	Trans people	BART-BiHNet+EWC	0.917531	0.611940
hatecheck	Trans people	BART-Adapter-Multitask	0.933852	0.515723
hatecheck	Trans people	BART-BiHNet-Multitask	0.850020	0.397906
hatecheck	women	BART-Single	0.946348	0.680851
hatecheck	women	BART-Adapter-Single	0.963803	0.857143
hatecheck	women	BART-BiHNet-Single	0.953789	0.891892
hatecheck	women	BART-Adapter-Vanilla	0.928732	0.780488
hatecheck	women	BART-BiHNet+Vanilla	0.955639	0.550000
hatecheck	women	BART-BiHNet+Reg	0.958494	0.839506
hatecheck	women	BART-BiHNet+EWC	0.884371	0.418919
hatecheck	women	BART-Adapter-Multitask	0.949163	0.750000
hatecheck	women	BART-BiHNet-Multitask	0.927204	0.409639
misogyny	-	BART-Single	0.822216	0.329032
misogyny	-	BART-Adapter-Single	0.837551	0.334426
misogyny	-	BART-BiHNet-Single	0.805479	0.322785
misogyny	-	BART-Adapter-Vanilla	0.844372	0.395522
misogyny	-	BART-BiHNet+Vanilla	0.839064	0.382671
misogyny	-	BART-BiHNet+Reg	0.828667	0.335616
misogyny	-	BART-BiHNet+EWC	0.848168	0.372760

Continued on next page

Continued from previous page

dataset	task	model	few-shot-auc	few-shot-f1
misogyny	-	BART-Adapter-Multitask	0.865112	0.396825
misogyny	-	BART-BiHNet-Multitask	0.805919	0.327759
multi	-	BART-Single	0.839205	0.401028
multi	-	BART-Adapter-Single	0.709392	0.259740
multi	-	BART-BiHNet-Single	0.642476	0.196636
multi	-	BART-Adapter-Vanilla	0.617924	0.191589
multi	-	BART-BiHNet+Vanilla	0.614951	0.215269
multi	-	BART-BiHNet+Reg	0.616131	0.191702
multi	-	BART-BiHNet+EWC	0.597265	0.195773
multi	-	BART-Adapter-Multitask	0.674493	0.248244
multi	-	BART-BiHNet-Multitask	0.623469	0.216086
single	-	BART-Single	0.990007	0.852679
single	-	BART-Adapter-Single	0.988204	0.871287
single	-	BART-BiHNet-Single	0.965336	0.679856
single	-	BART-Adapter-Vanilla	0.939223	0.629126
single	-	BART-BiHNet+Vanilla	0.888218	0.508744
single	-	BART-BiHNet+Reg	0.927218	0.629771
single	-	BART-BiHNet+EWC	0.932330	0.634051
single	-	BART-Adapter-Multitask	0.969689	0.716904
single	-	BART-BiHNet-Multitask	0.928502	0.606171
adversarial	-	BART-Single	0.979721	0.837321
adversarial	-	BART-Adapter-Single	0.977043	0.781726
adversarial	-	BART-BiHNet-Single	0.954980	0.670232
adversarial	-	BART-Adapter-Vanilla	0.857171	0.490196
adversarial	-	BART-BiHNet+Vanilla	0.837839	0.439873
adversarial	-	BART-BiHNet+Reg	0.859952	0.511149
adversarial	-	BART-BiHNet+EWC	0.864196	0.520979
adversarial	-	BART-Adapter-Multitask	0.912971	0.607803
adversarial	-	BART-BiHNet-Multitask	0.838634	0.444444
stormfront	-	BART-Single	0.844468	0.805897
stormfront	-	BART-Adapter-Single	0.811555	0.766595
stormfront	-	BART-BiHNet-Single	0.757382	0.709832
stormfront	-	BART-Adapter-Vanilla	0.884122	0.733728
stormfront	-	BART-BiHNet+Vanilla	0.848016	0.756032
stormfront	-	BART-BiHNet+Reg	0.861334	0.776903
stormfront	-	BART-BiHNet+EWC	0.854757	0.792929
stormfront	-	BART-Adapter-Multitask	0.901288	0.810390
stormfront	-	BART-BiHNet-Multitask	0.868593	0.757493
US-election	hateful	BART-Single	0.668330	0.228571
US-election	hateful	BART-Adapter-Single	0.664259	0.232558
US-election	hateful	BART-BiHNet-Single	0.616334	0.224852
US-election	hateful	BART-Adapter-Vanilla	0.761029	0.379747
US-election	hateful	BART-BiHNet+Vanilla	0.744485	0.296875
US-election	hateful	BART-BiHNet+Reg	0.751641	0.357895
US-election	hateful	BART-BiHNet+EWC	0.787684	0.314961
US-election	hateful	BART-Adapter-Multitask	0.781250	0.408602
US-election	hateful	BART-BiHNet-Multitask	0.788209	0.350877

Table 11: AUC and F1 scores for few-shot downstream tasks for the random order experiment