# Detecting harassment and defamation in cyberbullying with emotion-adaptive training


## Abstract

Existing research on detecting cyberbullying incidents on social media has primarily concentrated on harassment and is typically approached as a binary classification task. However, cyberbullying encompasses various forms, such as denigration and harassment, which celebrities frequently face. Furthermore, suitable training data for these diverse forms of cyberbullying remains scarce. In this study, we first develop a celebrity cyberbullying dataset that encompasses two distinct types of incidents: harassment and defamation. We investigate various types of transformer-based models, namely masked (RoBERTa, Bert and DistilBert), replacing (Electra), autoregressive (XLnet), masked&permuted (Mpnet), text-text (T5) and large language models (Llama2 and Llama3) under low source settings. We find that they perform competitively on explicit harassment binary detection, however, their performance is substantially lower on harassment and denigration multi-classification tasks. Therefore, we propose an emotion-adaptive training framework (EAT) that helps transfer knowledge from the domain of emotion detection to the domain of cyberbullying detection to help detect indirect cyberbullying events. EAT consistently improves the average macro F1, precision and recall by 20% in cyberbullying detection tasks across nine transformer-based models under low-resource settings. Our claims are supported by intuitive theoretical insights and extensive experiments.[1]
Warning: This paper contains offensive words, which do not reflect the views of the authors.


## Introduction

Cyberbullying is a serious global issue, which is a specific form of bullying that happens within online environments (Smith et al. 2008). Cyberbullying manifests itself in diverse forms, such as harassment (insults or threats), spreading rumours, impersonation, outing and trickery (sharing someone's confidential information) or exclusion (e.g. from activities) (Peled 2019). Despite the multifaceted nature of cyberbullying, existing research (Agrawal and Awekar 2018a; Muneer and Fati 2020; Kim et al. 2021; Yi and Zubiaga 2023b) has primarily focused on direct forms such as harassment and flaming, with limited exploration of indirect forms like denigration, which is an obstacle for addressing e.g. celebrity cyberbullying. Recent studies demonstrate that cyberbullying targeting influencers and celebrities has become commonplace (Takano et al. 2024). Furthermore, findings from Ouvrein, De Backer, and Vandebosch (2018) suggest that some adolescents do not perceive negative comments against celebrities as cyberbullying, but rather as legitimate, personal opinions. Nonetheless, celebrities often suffer serious consequences, including alcohol and drug addictions, self-blame, and depression. Figure 1 illustrates two distinct forms of cyberbullying towards celebrities, which involve repeated aggressive (direct) and defamatory (indirect) behaviour.

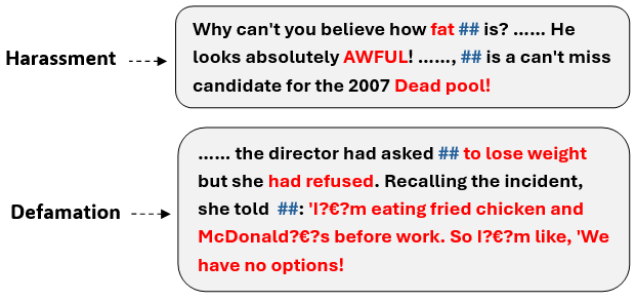


Figure 1: Examples of celebrity cyberbullying, with key terms highlighted and names anonymised as ##.

In this work, we investigate multi-class celebrity cyberbullying detection, tackling three key bottlenecks in existing research: *B1) Lack of suitable datasets*: previous attempts relying on offensive language as a proxy for cyberbullying data collection fail to capture indirect incidents of cyberbullying; *B2) Heterogeneous classes:* existing research on multi-class classification in cyberbullying primarily focuses on discerning the severity levels of direct forms, overlooking the detection of distinct types such as harassment and defamation; *B3) Dearth of research in low-resource settings:* Transformer-based models are widely used in advanced cyberbullying detection research, but there is a notable scarcity of research in the field targeting resource-poor environments, even though these environments are the most common context in cyberbullying studies.

**Proposed Approach:** To address these challenges, we



[1]The data and source are publicly available at https://github.com/Misinformation-emotion/Cyberbullying-emotion

initially develop a new celebrity cyberbullying dataset called HDCyberbullying. This dataset contains two different forms of cyberbullying: harassment and defamation. Then we propose an emotion-adaptive training framework (EAT) under low-resource settings to tackle B2&B3. The approach is straightforward: given the limited data availability, our training emphasis shifts away from the cyberbullying detection dataset. Instead, we focus on acquiring knowledge in the emotion detection domain and subsequently transferring it to the cyberbullying detection task.

To evaluate the effectiveness and robustness of our method, we apply the method to nine transformer-based models: RoBERTa, Bert, DistilBert, Electra, XLnet, T5, Mpnet, Llama2 and Llama3. We conduct a series of experiments and quantitative analyses to understand why our approach can be successful across a broad range of models.

**Contributions:** To the best of our knowledge, our work is the first investigation into the detection of celebrity cyberbullying and multiple classifications of direct and indirect cyberbullying. Our contributions include:

- Our study underscores the transferability from the emotion domain to the domain of cyberbullying detection.
- To encourage diversity in cyberbullying research, we create the first dataset that reflects real-life cyberbullying scenarios involving celebrities.
- We propose the EAT framework to address heterogeneous class detection problems in cyberbullying within resource-poor environments.
- We quantitatively evaluate the efficacy of EAT with nine state-of-the-art models, along with qualitative evaluations to gain an in-depth understanding of why and how EAT achieves competitive and consistent results.

## Related work

Cyberbullying detection is generally defined as a binary classification task determining if an instance constitutes a case of cyberbullying or not. Cyberbullying detection research by Agrawal and Awekar (2018a); Dadvar and Eckert (2018); Yuvaraj et al. (2021); Cheng et al. (2020) leveraged language models, including methods based on GloVe and Word2Vec in combination with deep learning methods such as BLSTM, CNN and attention mechanisms. Using Transformer-based models to identify cyberbullying through multiple user interactions demonstrates that transformer-based models can be strong, and competitive for cyberbullying detection (Gururangan et al. 2020; Yi and Zubiaga 2023b,a). Large language models(LLMs) have demonstrated robust comprehension capabilities (Liu et al. 2023) to human commands, particularly following advancements in reinforcement learning through Prompt Engineering (Ouyang et al. 2022; AlKhamissi et al. 2022). Li et al. (2023) illustrate that ChatGPT exhibits high efficacy in detecting harmful textual content.

Existing research studies have primarily focused on direct forms of cyberbullying, such as harassment and flaming, with limited exploration of indirect cyberbullying, typified by denigration. Handling both overt and implicit forms of cyberbullying simultaneously remains a challenge. However, several works have shown awareness of this issue. For instance, Saengprang and Gadavanij (2021); Maftei, Holman, and Merlici (2022); Scheithauer et al. (2021) have all highlighted the prevalence of harassment and denigration in cyberbullying incidents involving public figures and adolescents. Some research efforts have specifically targeted denigration detection. For example, Sangwan and Bhatia (2020, 2022) employed various feature selection algorithms to enhance the relevance of features for denigration classification. Furthermore, while emotions play a significant role in driving human behaviour, the exploration of their influence on improving cyberbullying detection is still in its early stages (Al-Hashedi et al. 2023). In our study, we perform an in-depth empirical study on the strong connection between emotion and cyberbullying, demonstrating that models can effectively incorporate knowledge from emotion datasets to help identify indirect cyberbullying incidents.

Domain adaptation is the foundation of our approach. It is a type of transfer learning designed to adapt a model trained on a source domain so that it performs effectively on a target domain with a different data distribution. Domain adaptation in the context of cyberbullying detection is an emerging field initiated by Agrawal and Awekar (2018b). They studied the performance of a zero-shot transfer learning approach on three different social platforms (Wikipedia, Twitter and Formspring), training and testing on different platforms. Their study underscored the complexity and challenges of the problem. To tackle this challenge, Yi and Zubiaga (2022) proposed adversarial transformers by integrating unlabeled data from both source and target domains into a unified representation, thereby avoiding platform-specific training. Inspired by this research, we aim to achieve a common space of knowledge transfer between the emotion detection domain and the cyberbullying detection domain.

## HDCyberbullying

In the past decade, NLP researchers made available several cyberbullying datasets for detection tasks (Yi and Zubiaga 2023b), However, existing datasets 1) cover direct cyberbullying but lack coverage of indirect cyberbullying, and 2) do not simultaneously target celebrity harassment and defamation. To address these gaps, we compile HDCyberbullying.

### The definition of celebrity cyberbullying

There is no established definition of celebrity cyberbullying. Based on existing studies (Takano et al. 2022; Karthika 2022; Vogel 2021), we define celebrity cyberbullying as the incidents of repeated harassment and defamation of celebrities by name, where these two terms are defined as follows:

- **Harassment:** including **1) name-calling,** i.e. insulting someone by calling them rude names; **2) offensive abuse,** i.e. comments that cause mental pain through "abusive appearance", "abusive ability", "abusive personality", etc.; **3) hateful comments,** i.e. offensive discourse towards targets based on race, religion, skin colour, sexual identity, gender identity, ethnicity, disability, or national identity.
- **Defamation:** speaking half-truths or lies.

### Selection of comments

According to the above definition, HDCyberbullying is exclusively designed for celebrity cyberbullying. By combining the harassment dataset (surekharamireddy 2021) and the fake news dataset (Pérez-Rosas et al. 2018), both of which are relevant to celebrity disinformation. The harassment dataset (surekharamireddy 2021) notes that many celebrities face backlash and encounter hateful and offensive comments. In our initial analysis, 70% of the samples contained named comments, which is significantly different from the anonymous user attacks prevalent in other cyberbullying datasets. We used the Stanford NER Tagger (Stanford 2022) to identify harmful comments directed at celebrities by selecting those containing celebrities' and other people's names. We did not remove unfamiliar names, believing this would not affect the data quality.

The Defamation Dataset (Pérez-Rosas et al. 2018) focuses on actors, singers, socialites, and politicians, and is collected from web sources to identify naturally occurring false content. The data collection was paired, with one article being legitimate and the other being false. To determine whether a celebrity news article is legitimate or not, claims made in the article were evaluated through gossip-checking websites and cross-referenced with information from other entertainment news sources. Using this method, 500 news articles with an even distribution of fake and legitimate news were collected.

### Annotation

These original datasets already contain labels related to cyberbullying. For example, the Harassment dataset (surekharamireddy 2021) includes six categories: 'Malignant', 'Highly malignant', 'Rude', 'Threat', 'Abuse', and 'Loathe', which are combined into a single 'harassment' tag. Therefore, all samples in the HDCyberbullying dataset fall into one of three categories:**harassment, defamation, and non-cyberbullying**. We manually verify that each tag meets our definition above. Table 1 illustrates examples of annotation.

### Data statistics

The dataset consists of 2,907 reviews, including 250 defamatory comments, 1,204 harassing comments, and 1,453 non-cyberbullying comments. There is a notable imbalance in both text length and category distribution, with defamatory data representing only 8% of the total. In contrast, the average text length is 1985 words, and harassment data accounts for 41%. The average length of harassing text is 455 words. Additionally, the diversity of language features and datasets poses challenges for multi-category tasks.

## Methodology

This section defines the research problem and elucidates the theoretical underpinnings of our solution, thereby laying the foundation for understanding subsequent explanations of EAT.

### Problem definition

**Definition 1: Cyberbullying detection.** A Multiclass classification task determines if each text in $T \in \{T_1, \ldots, T_n\}$ indicates a cyberbullying incident. i.e. $y \in \{0, 1, 2\}$, where $y = 1$ means that a post refers to a case of harassment, $y = 2$ means that a post refers to a case of defamation, and $y = 0$ means that it is not a case of cyberbullying.

**Definition 2: Emotion detection.** A Multiclass classification task determines if each text in $E \in \{E_1, \ldots, E_n\}$ belongs to an emotion group related to cyberbullying incident. i.e. $y \in \{0, 1, 2\}$, where $y = 1$ means that a post expresses one or more emotions indicative of harassment, $y = 2$ means that the emotions of a post most intensely towards defamation. When $y = 0$, it indicates that the detected emotions are unrelated to cyberbullying events.

**Definition 3: Zero-shot classification (ZSC).** In each experiment configuration, datasets belong to the training datasets $s$ or test datasets $t$, which leads to two different input spaces $X_s$ and $X_t$ where $X_s \neq X_t$, and they have different label space $Y_s \neq Y_t$. Moreover, the training dataset's data distribution is $P_s(x, y)$ and the test dataset distribution $P_t(x, y)$, which are different from each other, are both unknown and imbalanced. Zero-shot classification aims to learn a classifier for classifying testing instances belonging to the unseen label $Y_t$.

**Definition 4: Few-shot classification (FSC).** In each experiment configuration, $X_s \sim X_t$ and they have the same label space $Y_s \equiv Y_t$. However, the number of training datasets is too small to help find a data distribution $P_s(x, y) \approx P_t(x, y)$. Therefore, the few-shot classification algorithm is a parameterized optimal strategy for finding $P_t(x, y)$.

**Problem definition:** We assume that the emotion detection domain and the cyberbullying detection domain are related but from different distributions. However, a dataspace $D_e$ in emotion detection datasets can be found to transfer the knowledge from the emotion detection domain to the cyberbullying detection domain. Under this assumption, our task in the study is reducing the disparity across domains and training the emotion detection classifier $C_e$, which can be directly applied to instances from the cyberbullying detection domain (ZSC), on labelled emotion detection data. Or help improve performance of the cyberbullying classifier $C_t$ on low resource settings (FSC).

### Theoretical Analysis

In this section, we will conduct a theoretical analysis of our method. Our analysis is based on (Ben-David et al. 2010) and (Yosinski et al. 2014)'s foundational theoretical research on the adaptability of transfer learning. We explain how these theories support our approach to obtain substantially improved results over state-of-the-art models.

**Data space similarity:** The disparity between the target and source domain can significantly affect the target task. We need high-quality source datasets to train a source classifier with a lower loss to maximise the effectiveness of the transmission. Hence, the initial step for EAT involves carefully selecting the source domain, and aiming to gauge the similarity of the cyberbullying detection dataspace to the emotion detection dataspace.

| Annotation | Posts | Tag |
|---|---|---|
| Offensive abuse | ... my acctions on ## abuse userpage ... i suspect he is a ...sockpuppet | 1 |
| Hateful comments | ...to call ## a moron which is what he is... | 1 |
| Name calling | ...Now,##...the typical right-wing... | 1 |
| Defamation | ...##: Relapse Fears Over Showbiz Deaths... | 2 |
| No-Cyberbullying | ... getting more ## pics if they're available... | 0 |

Table 1: Example annotation in HDCyberbullying, with key terms and names replaced by ## .

**Domain shift:** This part is to deal with the data shift ($P_s(x, y) \neq P_t(x, y)$) between the source domain and the target domain to make them similar ($P_s(x, y) \approx P_t(x, y)$). We can decompose the joint probability distribution as $p(x, y) = p(y)p(x|y)$ or $p(x, y) = p(x)p(y|x)$. The source and the target posteriors are arbitrary $P_s(x|y) \neq P_t(x|y)$. The problem becomes intractable when $P_s(y) \neq P_t(y)$. Thus we make the data distribution shift ($P_s(x) \approx P_t(x)$) first by selecting data with certain groups. Then concept shift makes $P_s(y|x) \approx P_t(y|x)$. This is achieved by grouping and mapping the emotion categories that are closely related to the cyberbullying category.

**A high-quality classifier:** The extent to which knowledge from the source domain can be transferred to the target domain primarily depends on the hypothesis loss in the source domain. The performance of the model is contingent upon the training loss in both domains. If the combined loss is substantial, the model may struggle to perform effectively. Hence, we require a model structure that performs well in both domains. Transformer-based pre-trained models are trained on extensive and diverse corpora. Representations learned by such models demonstrate robust performance across numerous tasks, utilizing datasets of varying sizes and originating from diverse sources. Therefore, we extensively selected pre-trained models of different architectures as the knowledge containers for this study.

## Emotion adaptive training (EAT)

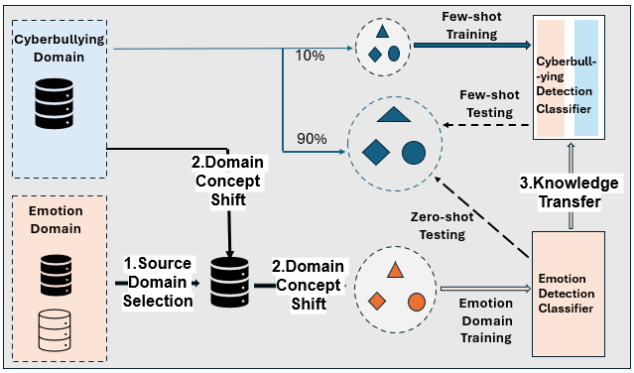


Figure 2: Architecture of EAT

In this section, we detail our simple yet effective domain adaptation approach, which consists of three main phases: source domain selection, domain concept shift and knowledge transfer. The framework is depicted in Figure 2.

**Source domain selection:** This phase aims to perform the distribution shifts ($P_s(x) \approx P_t(x)$) by selecting a single emotion domain as the source domain from multiple potentially helpful emotion datasets. In the past decade, NLP researchers made several datasets available for language-based emotion classification for various domains and applications (Buechel and Hahn 2022; Scherer and Wallbott 1994; Oberländer, Kim, and Klinger 2020; Liu, Osama, and De Andrade 2019; Kajiwara et al. 2021; Demszky et al. 2020). To find a source domain relevant to the target domain and a high-quality dataset sufficient to train a high-quality emotion classifier, we need a large-scale, consistent labelled emotion dataset based on fine-grained taxonomies with proven high-quality annotations and from a single genre (Bostan and Klinger 2018). (Demszky et al. 2020) is the largest manually annotated dataset of 58k English Reddit comments, labelled for 27 emotion categories and demonstrated the high quality of the annotations via Principal Preserved Component Analysis.

To intuitively assess the similarity between the emotion data domain and the cyberbullying data domain, we randomly selected 1000 samples from each domain. We then generated 768-dimensional RoBERTa embeddings and reduced them to 2 dimensions using PCA (Principal Component Analysis). PCA performs dimensionality reduction while preserving as much variance as possible. The resulting, average cosine similarity is 0.993. Figure 3 depicts the dataspace, where blue spots represent the cyberbullying dataset and orange spots represent the emotion dataset. The two domains overlap significantly.

**Domain concept shift:** This phase ensures that $P_s(y|x) \approx P_t(y|x)$, which is achieved by grouping and mapping emotion categories that are closely related to the cyberbullying category. For instance, data instances labelled as 'Anger' or 'Disgust' in the emotion domain are grouped and assigned label 1, corresponding to the cyberbullying domain of 'Harassment'. Figure 4 illustrates this relationship, and is based on prior research identifying anger and disgust as powerful emotions shared by victims and bullies in both physical and electronic forms of bullying (Burgess-Proctor, Patchin, and Hinduja 2009; Lonigro et al. 2015; Wang et al. 2017; Al-Hashedi et al. 2023). Surprise is felt most intensely towards celebrity fake news (Tan and Hsu 2023; Liu et al. 2024). There are many

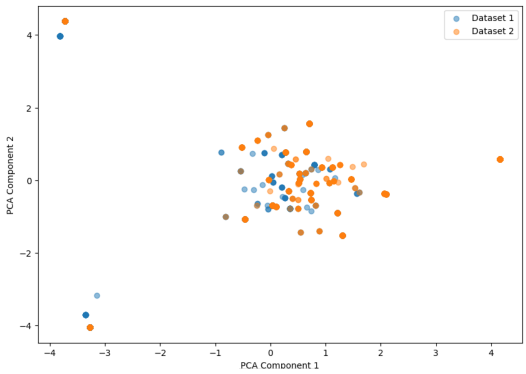


Figure 3: Domain similarity. Orange: Emotion data; Blue: Cyberbullying data.

choices for positive emotions in the emotion dataset. We chose 'gratitude' and 'joy' because the original authors achieved the best results for these two categories using the BERT-based emotion classification model (Demszky et al. 2020). This ensures the maximum potential for high-quality source classifiers from these two categories.

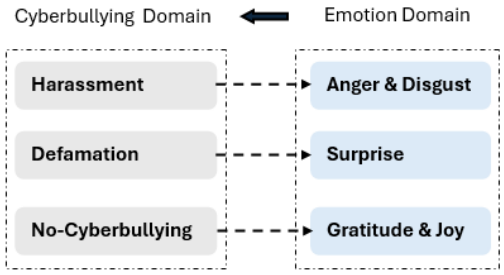


Figure 4: Mapping between Emotion and Cyberbullying domain.

**Knowledge transfer:** In this phase, we consider unsupervised domain adaptation for ZSC and semi-supervised domain adaptation for FSC. Unsupervised domain adaptation fine-tuning pre-trained models on labelled emotion datasets without requiring access to the cyberbullying dataset. Semi-supervised domain adaption based on the model trained by unsupervised domain adaptation do further training on a few cyberbullying data. How much knowledge can be transferred is decided by combining the loss of emotion classifier and the cyberbullying classifier.

## Experiments

This section conducts quantitative evaluation to assess the effectiveness of EAT, complemented by qualitative experiments aimed at providing a comprehensive understanding of the reasons behind EAT's consistent results.

### Transformer-based models

The different pre-training approaches are crucial for the success of downstream tasks. In the experiment, we tested nine transformer-based models with six types of pre-training approaches to examine the broad applicability of the EDA approach.

**Masked:** This approach is to mask some tokens in the input sequence, and the model learns to predict the masked tokens based on the surrounding context which has been shown to have significant advantages for text classification. We test three masked pre-trained models, namely BERT (Devlin et al. 2019), Roberta (Liu et al. 2019) and Distilbert (Sanh et al. 2019).

**Autoregressive:** XLnet (Yang et al. 2019) considering the dependence of masked tokens which missing in the masked approach, based on autoregressive language modelling.

**Masked+Permuted:** Mpnet (Song et al. 2020) inherits the advantages of BERT and XLNet, and leverages the full position information to reduce the discrepancy between pre-training and fine-tuning by combining masked and permuted approaches in a view.

**Replacing:** Electra (Clark et al. 2020) replacing some input tokens with sampled from a small generator network instead of masking the input.

**Text-text:** Text-to-text approach converts text processing problems into a universal text generation structure. T5 (Raffel et al. 2020) will be examined in this experiment as a representative model of this approach.

**LLMs:** Large language models trained on a diverse and extensive dataset composed of publicly available text data. This data includes a wide range of domains, such as books, websites, and other forms of written content, to ensure the model captures a broad spectrum of language. Llama2 and Llama3, which utilize this technique along with extensive human preference data, demonstrate impressive performance on many NLP tasks (Touvron et al. 2023).

### Settings

All pre-trained models are based on Hugging Face's Enterprise Hub. To ensure a fair comparison, all models, except for the Llama2 and Llama3 models, use SimpleTransformer wrapper (Rajapakse 2019). The training hyperparameters recommended by (Sun et al. 2020) are as follows: batch size of 32, learning rate (Adam) of $4e^{-5}$, number of epochs set to 4, and a token length of 400, all trained on an L4 GPU.

### Evaluation

We use three widely adopted evaluation metrics for each class, namely recall, precision and micro-F1. Additionally, we also report macro-averaged recall, precision, and F1 across all classes.

For the baseline, we fine-tune the pre-trained LLM on $10\%$ of the data (291 samples) and then test it on the remaining $90\%$ (2,524 samples). To evaluate the extent to which EAT can help pre-trained models improve generalization, we extracted 6 subsets of training data of different sizes: 72, 140, 210, 400, 700, 1300. Table 2 shows these training settings of each class and the number of testing data.

In the zero-shot experiment, the training dataset is exclusively from the emotion detection domain, with no visibility of the cyberbullying detection dataset. In the few-shot experiment, we use the same fine-tuning and test data as the

| | #Harassment | #Defamation | #No-Cyberbullying | #Emotion | #Testing |
|---|---|---|---|---|---|
| Baseline | 119 | 20 | 152 | 0 | 2,615 |
| Zero-shot | 0 | 0 | 0 | 3700 | 2,615 |
| Few-shot | 119 | 20 | 152 | 3700 | 2,615 |
| 72 | 29 | 5 | 38 | 3700 | 2,834 |
| 140 | 57 | 10 | 73 | 3700 | 2,766 |
| 210 | 86 | 14 | 110 | 3700 | 2,696 |
| 400 | 164 | 27 | 209 | 3700 | 2,506 |
| 700 | 286 | 48 | 366 | 3700 | 2,206 |
| 1300 | 531 | 89 | 680 | 3700 | 1,606 |

Table 2: Data settings for evaluation.

baseline. Each model is run 5 times to report average performance.

## Results

| | | Baseline | | | Zero-shot (EAT) | | | Few-shot (EAT) | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Model | | P | R | F1 | P | R | F1 | P | R | F1 |
| RoBerta | H | 0.81 | 0.91 | 0.85 | 0.75 | 0.94 | 0.83 | 0.84 | 0.95 | **0.89** |
| _base | D | 0 | 0 | 0 | 0.18 | 0.44 | 0.25 | 0.81 | 0.72 | **0.76** |
| | A | 0.58 | 0.55 | 0.56 | 0.60 | 0.61 | 0.57 | 0.85 | 0.83 | **0.84** |
| Bert | H | 0.85 | 0.85 | 0.85 | 0.71 | 0.93 | 0.80 | 0.86 | 0.92 | **0.89** |
| _base | D | 0 | 0 | 0 | 0.17 | 0.10 | 0.12 | 0.73 | 0.66 | **0.69** |
| | A | 0.57 | 0.52 | 0.54 | 0.56 | 0.56 | 0.55 | 0.82 | 0.81 | **0.81** |
| Distilbert | H | 0.87 | 0.86 | 0.87 | 0.69 | 0.93 | 0.79 | 0.86 | 0.92 | **0.89** |
| _base | D | 0 | 0 | 0 | 0.15 | 0.19 | 0.17 | 0.80 | 0.43 | **0.56** |
| | A | 0.54 | 0.59 | 0.56 | 0.56 | 0.56 | 0.55 | 0.84 | 0.74 | **0.77** |
| Mpnet | H | 0.86 | 0.91 | 0.88 | 0.82 | 0.92 | 0.87 | 0.87 | 0.94 | **0.90** |
| _base | D | 0 | 0 | 0 | 0.17 | 0.25 | 0.20 | 0.77 | 0.71 | **0.74** |
| | A | 0.55 | 0.60 | 0.57 | 0.60 | 0.61 | 0.60 | 0.84 | 0.83 | **0.84** |
| Electra | H | 0.90 | 0.37 | 0.52 | 0.67 | 0.90 | 0.70 | 0.86 | 0.89 | **0.87** |
| _small | D | 0 | 0 | 0 | 0 | 0 | 0 | 0.80 | 0.37 | **0.51** |
| | A | 0.49 | 0.44 | 0.42 | 0.45 | 0.48 | 0.44 | 0.82 | 0.71 | **0.74** |
| XLnet | H | 0.87 | 0.91 | 0.89 | 0.73 | 0.92 | 0.82 | 0.87 | 0.91 | **0.89** |
| _base | D | 0.79 | 0.31 | 0.44 | 0.09 | 0.11 | 0.10 | 0.71 | 0.72 | **0.67** |
| | A | 0.82 | 0.69 | 0.72 | 0.53 | 0.53 | 0.52 | 0.82 | 0.80 | **0.81** |
| T5 | H | 0.92 | 0.54 | 0.68 | 0.65 | 0.90 | 0.75 | 0.83 | 0.86 | **0.84** |
| _base | D | 0.49 | 0.10 | 0.17 | 0.18 | 0.26 | 0.21 | 0.30 | 0.97 | **0.45** |
| | A | 0.68 | 0.53 | 0.54 | 0.54 | 0.54 | 0.50 | 0.67 | 0.78 | **0.65** |
| Llama3 | H | 0.71 | 0.70 | 0.72 | 0.91 | 0.85 | 0.87 | 0.90 | 0.91 | **0.91** |
| _8B | D | 0.66 | 0.58 | 0.52 | 0.75 | 0.84 | 0.56 | 0.72 | 0.67 | **0.69** |
| | A | 0.69 | 0.67 | 0.65 | 0.79 | 0.73 | 0.74 | 0.83 | 0.81 | **0.82** |
| Llama2 | H | 0.63 | 0.62 | 0.63 | 0.75 | 0.89 | 0.82 | 0.86 | 0.89 | **0.87** |
| _7B | D | 0.39 | 0.65 | 0.49 | 0.14 | 0.32 | 0.19 | 0.60 | 0.67 | **0.63** |
| | A | 0.56 | 0.61 | 0.57 | 0.57 | 0.57 | 0.55 | 0.78 | 0.80 | **0.78** |
| Overall | | 0.61 | 0.58 | 0.57 | 0.58 | 0.58 | 0.56 | **0.80** | **0.79** | **0.78** |

Table 3: A comparison of models' performance. The best results and models are highlighted in bold. P: Precision; R: Recall; F1: Micro F1;H: harassment; D: defamation; A: Macro-averaging; Overall: Average of A.

Table 3 show that our proposed method EAT can effectively improve the average macro F1, precision and recall by 20%. We report results for both classes: harassment(H) and defamation(D) with Precision, Recall and F1. The following are observed: 1) In case of harsh data imbalance (20 defamation samples for training), our EAT approach wasn't biased to the majority class, and learning meaningful patterns and relationships made the model generalize well when encountering unseen instances from these classes. 2) Mask-based models in small data settings (291 training samples) without EAT fine-tuning are ineffective at identifying defamation. These models use bidirectional context, which may not always capture dependencies, especially in complex language constructs. Moreover, due to the inherent constraints of mask-based models, they perform well with shorter to moderate-length sequences, given their training objective. The average text length of a defamation message is 1985, which poses a challenge. In contrast, XLnet considers the dependence on masked tokens which missing in the masked approach show potential generalizability. 3) T5 and LLaMA models demonstrate excellent scalability, particularly for long text tasks, despite differences in their architectures and training objectives. T5 employs an encoder-decoder architecture and is pre-trained with a span-corruption objective, where random spans of text are masked, and the model learns to reconstruct them. This approach allows T5 to effectively handle context over longer spans of text. In contrast, LLama models use a decoder-only architecture designed for scalability. With large model sizes that capture extensive context, LLama can be fine-tuned on specific tasks or datasets that involve long text. 4) Mpnet is one of the best-performing models across zero-shot and few-shot tasks. Mpnet leverages the dependency among predicted tokens and auxiliary position information to reduce the disparity between the training domain and fine-tuning domain can be demonstrated effectively in the study.

In Figure 5, we can visually compare the performance of EAT on the defamation detection task. RoBERTa, Bert, DistilBert, Mpnet and Electra overlook the minority class with zero on the F1 score. while EAT helps find robust presentations and data patterns leading to optimal performance.

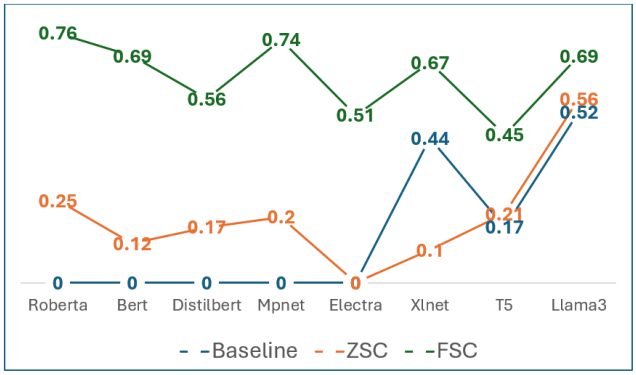


Figure 5: Performance of EAT in the defamation detection task.

## Error Analysis

To gain insights into the models' behaviours, we look at the confusion matrix and manually analyse the classification error samples of the best-performing approaches, i.e. RoBerta_base+EAT. Figure 6a) illustrates that the RoBERTa_base model is capable of detecting direct cyberbullying but struggles to differentiate indirect cyberbullying messages from non-bullying ones in low-resource fine-tuning. Defamation, like other forms of indirect cyberbullying, is ambiguous in its potential to be bullying.

Figure 6b) shows the results when the model is trained

solely on emotion datasets. The model exhibits high accuracy in classifying harassment, indicating that anger and disgust are strongly associated with harassment events. It is also observed that around 60% of non-cyberbullying events express emotions such as joy and gratitude. Interestingly, anger or disgust are also related to non-cyberbullying events. Past studies have found that surprise is most intensely associated with celebrity fake news (Tan and Hsu 2023; Liu et al. 2024). Our experimental results suggest that the relationships between emotions and celebrity defamation incidents are more complex than previously understood. Figure 6c) illustrates that, except for a slight drop in the accuracy of detecting harassment from 95% to 89%, the EAT model's performance improves significantly for the other two classes.

Among the misclassified harassment examples, we find that some messages have an overall tone that is confrontational and assertive, with multiple elements suggesting the speaker is angry and with aggressive attitude. However, these messages are not instances of harassment. From the context, they appear more like defensive statements, as illustrated by sample 1 in Table 4. This indicates that contextual knowledge is essential for the model to accurately distinguish such cases. The observation of inconsistency in sample 2 in Table 4 and the resulting question are typical reactions to something surprising. However, the statement appears to be a factual observation about the difference between the trailer and the final film and does not contain any false information. Factual news can evoke surprise due to the unexpected, unusual, or unprecedented nature of the events or information being reported (Qiu and Golman 2024).

Interestingly, our emotion classification model, EAT, classifies 70% of defamation as joy or gratitude rather than the surprise typically identified in past studies. For example, sample 3 in Table 4 contrasts with traditional cyberbullying research, which often associates such behaviour with a range of negative emotions from both the perpetrator and the victim. In defamation, positive context and emotions can still lead to negative consequences. Therefore, more nuanced emotional indicators need to be explored in future research.

Although anger or disgust are the primary emotions associated with harassment events, some misclassified examples include messages where the perpetrator or victim is not explicitly named. For instance, sample 4 in Table 4 is not predicted as harassment. The meaning is obscured by the use of ambiguous expressions, such as the question "Are you Jewish?" which reflects curiosity about the other person's background. The follow-up question, "How do you see any resemblance between the Holocaust and Armenian deportations?" suggests an attempt to understand the other person's perspective or reasoning. Additionally, the phrase "This may sound like a stupid question, but don't take it the wrong way" indicates an awareness that the question might be perceived as offensive or inappropriate. While these elements show a level of sensitivity and caution, the underlying racist implication remains.

Although the above analysis does not encompass all instances of model misclassification, it elucidates several underlying factors contributing to this phenomenon. Notably, the current intersection of research between the fields of cyberbullying detection and emotion analysis remains incomplete, particularly concerning indirect cyberbullying events. This gap underscores the need for further exploration and research, which may yield more effective detection models. We propose to address these limitations in future studies.

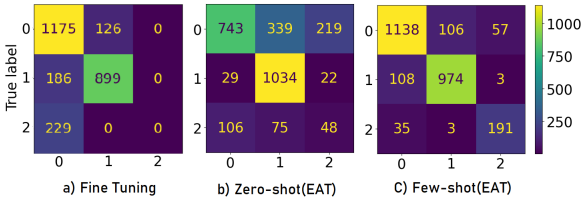


Figure 6: Confusion matrix of EAT. 0: No-cyberbullying; 1: Harassment; 2: Defamation.

## Insights from EAT

**Training loss** Based on the theory discussion in Section , we need to train a source classifier with a lower loss to maximise the effectiveness of transmission. we will specify the discussion to "how many datasets do we need to train a good source classifier to help the target classification task". Figure 7 illustrates the relation between training loss and model performance by using Mpnet+EAT. In general, there is an inverse relationship between emotion classifier training loss and model performance. However, when we continue to feed data, the training loss still decreases and the performance drops as well. Moreover, we observe that, when increasing data input from 2000 to 4000, the training loss is slightly increased. Our conjecture is that, when more data is input, the data diversity increases but the performance is the highest. In conclusion, reducing source classifier training loss is general, but need to consider more factors when a lack of target samples is present. For example, the boundary of source domain data, the quality of data, model structure, etc.

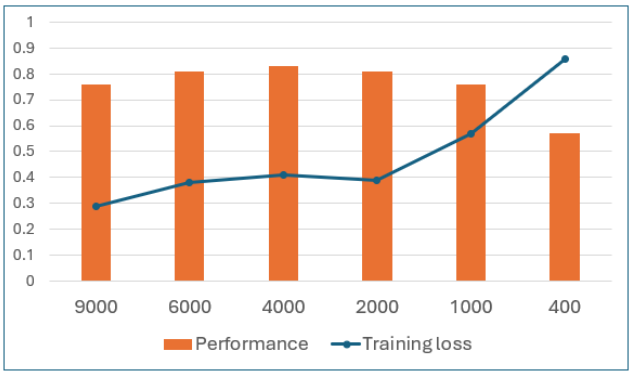


Figure 7: Training loss. The x-axis represents the size of the emotion training data and the y-axis represents the loss of training.

**Knowledge transfer** To visualize the transfer capabilities of EAT, Figure 8 displays the distribution of embeddings generated by three RoBERTa-based models on two domains.

| Index | Posts | Emotion | Cyberbullying | Prediction |
|---|---|---|---|---|
| 1 | ”Do not harass other editors with no basis You have posted a baseless warning/threat on my user page. Do not post any more things on my page. Otherwise, I will complain to the administrators, or even ## if I have to?” | Anger | No | Harassment |
| 2 | Trailer vs. Final Film I noticed that in trailer for the film, ## had two hamster butlers, while in the actual film, he’s the only pet living in the house. What’s up with that? | Surprise | No | Defamation |
| 3 | ##’s New Puppy Helping Her Heal With ‘Love & Support’ After Kidney Transplant With ##’s pooch is making her happier than ever and he’s helping her heal with ‘love and support’ after her kidney transplant, an expert revealed to HL EXCLUSIVELY. Find out why a The dog’s affection is so magical! ##,..... | Joy | Defamation | No |
| 4 | ##,this may sound a bit stupid question but dont get it wrong. are you Jew? if so how are you really thinking there is any resemplence between holocoust and Armeian deportations? | Curiosity | Harassment | No |

Table 4: False Positive and False Negative examples of EAT, with names replaced by ##

We can observe a robust correlation between emotions and cyberbullying and how our emotion-adaptive training framework effectively helps data clustering in both domains. Figure 8b) indicates that despite RoBERTa’s pre-training corpus being sourced from diverse origins, it still encounters challenges when fine-tuning for cyberbullying detection tasks in suboptimal settings. However, Figure 8c) illustrates that our method EAT efficiently transfers class knowledge from emotion detection to aid in the cyberbullying detection task. A reason is our selection of an emotion detection domain similar to the cyberbullying detection domain, as depicted in Figure 8a)

To evaluate the extent to which EAT can help models improve generalization, we extract 6 subsets of training data of different sizes, respectively 72, 140, 210, 400, 700, 1300, which are shown in Table 2 about these training settings of each class and the number of testing data. From Figure 9, we observe that the less data available, the more beneficial EAT becomes. When the training data is approximately 50% of the total data (1300 samples), the performance of the model without EAT fine-tuning is comparable to that of the EAT model, which requires only 10% (210 samples) of the data. EAT enables models to extract more transferable features or representations, enhancing generalization to unseen data.

**Emotion reference** One of the reasons why EAT can be successful is that we have applied concept shift $P_s(y|x) \approx P_t(y|x)$ to make $P_s(x, y) \approx P_t(x, y)$. We choose the emotion label according to past research, intuition and preliminary experiments. To gain insight into the relationship between cyberbullying and emotion, we calculate the likelihood of 28 emotions. From Figure 10, we can observe that

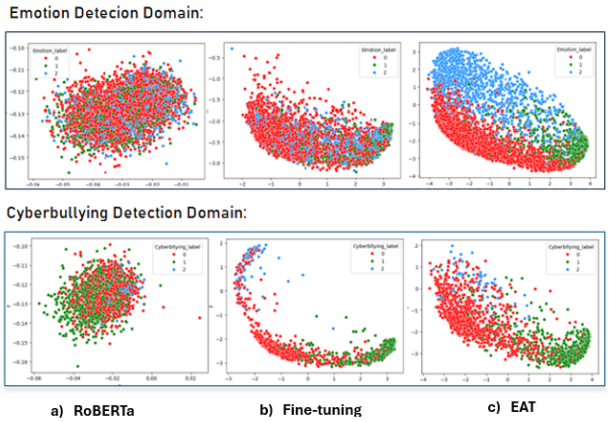


Figure 8: EAT t-SNE. Red: No-Cyberbullying samples, emotion (Gratitude and Joy) samples; Green: Harassment samples, emotion samples (Anger and disgusting); Blue: Defamation samples, emotion samples (Surprise).

the emotions we choose have a high likelihood but not the maximum in each group. Neutral is the most effective for all three types of cyberbullying affecting emotion labels. This shows that although the two data are similar, there is no one-to-one correspondence. In addition, for $P(E|H)$, annoyance has a high likelihood value like anger. For $P(E|D)$ we choose “surprise” according to the past research, but “approval” is more like the emotion the rumour wants to generate, and “admiration” looks more positive attitude related to

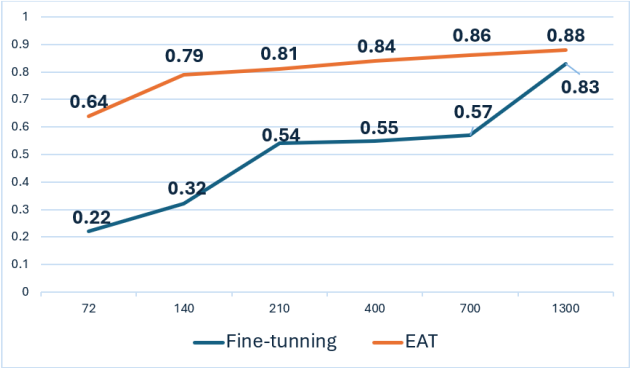


Figure 9: Capacity of transfer. The x-axis represents the size of the training data and the y-axis represents F1.

the no-cyberbullying event. At the same time, we notice all these emotions with high likelihood values didn't belong to (Ekman 1992) or (Plutchik 1980), usually used for emotion detection. So these findings call for further research into the relationship between cyberbullying and emotion.

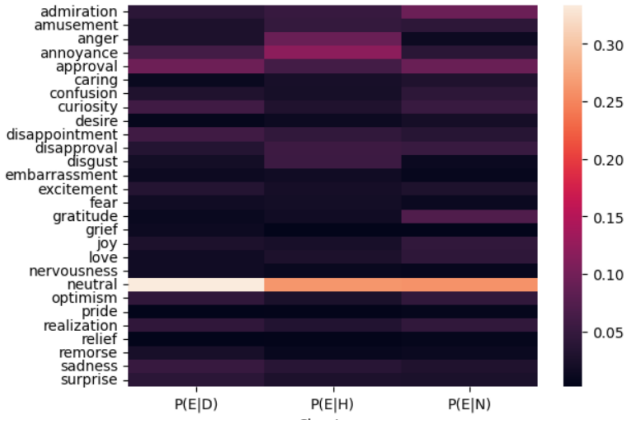


Figure 10: The likelihood of a certain emotion identifying given a class of data in the cyberbullying domain; E: emotion; H: Harassment; D: Defamation; N: No-cyberbullying.

## Conclusion

We investigate a wide range of transformer-based pre-trained models and find that they struggle in low-resource settings for multi-cyberbullying detection tasks, especially with indirect cyberbullying like defamation, as summarized in Table 3. We propose a novel, straightforward yet powerful domain adaptation framework, EAT, to tackle the above issue. Our experiments with EAT reveal that it may be valuable to complement work on ever-larger language models with parallel efforts to identify and use domain- and task-relevant emotion indicators to generalize models. The methods we studied are general enough to be applied to different structure transformer-based models. Our work points to numerous future directions, such as more comprehensive cyberbullying forms detection, better emotion indicator studies, efficient adaptation of large pre-trained language models to distant domains, and how to fine-tune language models after adaptation.

## Limitations

Our proposed EAT model demonstrates state-of-the-art performance in this study. However, our work is not without limitations. Most importantly, the lack of more and broader datasets limits the detection of various forms of cyberbullying beyond harassment and defamation, such as outings and frapping, which are more complicated forms. Furthermore, although our study demonstrates the transferability of emotion domain data to cyberbullying data, the selection of more effective source domain data and emotion indicators requires further research.

## Ethics Statement

Our research aims to support all individuals equally by curbing incidents of celebrity cyberbullying, particularly on social media, without intentionally discriminating against or causing harm to any vulnerable group. Our data comes from publicly available datasets and contains no personal information about each user. However, our research is not without risks, as adversaries involved in cyberbullying incidents could potentially use our findings for nefarious purposes, such as learning how to evade detection. This is not the intended use of our study.

## References

Agrawal, S.; and Awekar, A. 2018a. Deep learning for detecting cyberbullying across multiple social media platforms. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 10772 LNCS(Table 2): 141–153.

Agrawal, S.; and Awekar, A. 2018b. Deep learning for detecting cyberbullying across multiple social media platforms. In *European conference on information retrieval*, 141–153. Springer.

Al-Hashedi, M.; Soon, L.-K.; Goh, H.-N.; Lim, A. H. L.; and Siew, E.-G. 2023. Cyberbullying detection based on emotion. *IEEE Access*.

AlKhamissi, B.; Li, M.; Celikyilmaz, A.; Diab, M.; and Ghazvininejad, M. 2022. A review on language models as knowledge bases. *arXiv preprint arXiv:2204.06031*.

Ben-David, S.; Blitzer, J.; Crammer, K.; Kulesza, A.; Pereira, F.; and Vaughan, J. W. 2010. A theory of learning from different domains. *Machine Learning*, 79(1-2): 151–175.

Bostan, L.-A.-M.; and Klinger, R. 2018. An Analysis of Annotated Corpora for Emotion Classification in Text. In Bender, E. M.; Derczynski, L.; and Isabelle, P., eds., *Proceedings of the 27th International Conference on Computational Linguistics*, 2104–2119. Santa Fe, New Mexico, USA: Association for Computational Linguistics.

Buechel, S.; and Hahn, U. 2022. Emobank: Studying the impact of annotation perspective and representation

format on dimensional emotion analysis. *arXiv preprint arXiv:2205.01996*.

Burgess-Proctor, A.; Patchin, J. W.; and Hinduja, S. 2009. Cyberbullying and online harassment: Reconceptualizing the victimization of adolescent girls. *Female crime victims: Reality reconsidered*, 162: 176.

Cheng, L.; Guo, R.; Candan, K. S.; and Liu, H. 2020. Representation learning for imbalanced cross-domain classification. *Proceedings of the 2020 SIAM International Conference on Data Mining, SDM 2020*, 478–486.

Clark, K.; Luong, M.-T.; Le, Q. V.; and Manning, C. D. 2020. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In *International Conference on Learning Representations*, 1–18.

Dadvar, M.; and Eckert, K. 2018. Cyberbullying detection in social networks using deep learning based models. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 12393 LNCS, 245–255. ISBN 9783030590642.

Demszky, D.; Movshovitz-Attias, D.; Ko, J.; Cowen, A.; Nemade, G.; and Ravi, S. 2020. GoEmotions: A Dataset of Fine-Grained Emotions. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4040–4054. Online: Association for Computational Linguistics.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186.

Ekman, P. 1992. Are there basic emotions?

Gururangan, S.; Marasović, A.; Swayamdipta, S.; Lo, K.; Beltagy, I.; Downey, D.; and Smith, N. A. 2020. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8342–8360.

Kajiwara, T.; Chu, C.; Takemura, N.; Nakashima, Y.; and Nagahara, H. 2021. WRIME: A new dataset for emotional intensity estimation with subjective and objective annotations. In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies*, 2095–2104.

Karthika, C. 2022. Cyberbullying on celebrities: A case study on actress Parvathy. In *AIP Conference Proceedings*, volume 2463. AIP Publishing.

Kim, S.; Razi, A.; Stringhini, G.; Wisniewski, P. J.; and De Choudhury, M. 2021. A human-centered systematic literature review of cyberbullying detection algorithms. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2): 1–34.

Li, L.; Fan, L.; Atreja, S.; and Hemphill, L. 2023. "HOT" ChatGPT: The promise of ChatGPT in detecting and discriminating hateful, offensive, and toxic comments on social media. *ACM Transactions on the Web*.

Liu, C.; Osama, M.; and De Andrade, A. 2019. DENS: A dataset for multi-class emotion analysis. *arXiv preprint arXiv:1910.11769*.

Liu, P.; Yuan, W.; Fu, J.; Jiang, Z.; Hayashi, H.; and Neubig, G. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9): 1–35.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Liu, Z.; Zhang, T.; Yang, K.; Thompson, P.; Yu, Z.; and Ananiadou, S. 2024. Emotion detection for misinformation: A review. *Information Fusion*, 102300.

Lonigro, A.; Schneider, B. H.; Laghi, F.; Baiocco, R.; Pallini, S.; and Brunner, T. 2015. Is cyberbullying related to trait or state anger? *Child Psychiatry & Human Development*, 46: 445–454.

Maftei, A.; Holman, A.-C.; and Merlici, I.-A. 2022. Using fake news as means of cyber-bullying: The link with compulsive internet use and online moral disengagement. *Computers in Human Behavior*, 127: 107032.

Muneer, A.; and Fati, S. M. 2020. A comparative analysis of machine learning techniques for cyberbullying detection on twitter. *Future Internet*, 12(11): 187.

Oberländer, L. A. M.; Kim, E.; and Klinger, R. 2020. GoodNewsEveryone: A corpus of news headlines annotated with emotions, semantic roles, and reader perception. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 1554–1566.

Ouvrein, G.; De Backer, C. J.; and Vandebosch, H. 2018. Online celebrity aggression: A combination of low empathy and high moral disengagement? The relationship between empathy and moral disengagement and adolescents' online celebrity aggression. *Computers in Human Behavior*, 89: 61–69.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.

Peled, Y. 2019. Cyberbullying and its influence on academic, social, and emotional development of undergraduate students. *Heliyon*, 5(3).

Pérez-Rosas, V.; Kleinberg, B.; Lefevre, A.; and Mihalcea, R. 2018. Automatic Detection of Fake News. In *Proceedings of the 27th International Conference on Computational Linguistics*, 3391–3401.

Plutchik, R. 1980. A general psychoevolutionary theory of emotion. In *Theories of emotion*, 3–33. Elsevier.

Qiu, J.; and Golman, R. 2024. Curiosity in news consumption. *Applied Cognitive Psychology*, 38(2): e4195.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140): 1–67.

Rajapakse, T. C. 2019. Simple Transformers. https://github.com/ThilinaRajapakse/simpletransformers.

Saengprang, S.; and Gadavanij, S. 2021. Cyberbullying: The Case of Public Figures. *LEARN Journal: Language Education and Acquisition Research Network*, 14(1): 344–369.

Sangwan, S. R.; and Bhatia, M. 2020. Denigration bullying resolution using wolf search optimized online reputation rumour detection. *Procedia computer science*, 173: 305–314.

Sangwan, S. R.; and Bhatia, M. 2022. D-BullyRumbler: a safety rumble strip to resolve online denigration bullying using a hybrid filter-wrapper approach. *Multimedia Systems*, 28(6): 1987–2003.

Sanh, V.; Debut, L.; Chaumond, J.; and Wolf, T. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Scheithauer, H.; Schultze-Krumbholz, A.; Pfetsch, J.; and Hess, M. 2021. Types of cyberbullying. *The Wiley Blackwell handbook of bullying: A comprehensive and international review of research and intervention*, 1: 120–138.

Scherer, K. R.; and Wallbott, H. G. 1994. Evidence for universality and cultural variation of differential emotion response patterning. *Journal of personality and social psychology*, 66(2): 310.

Smith, P. K.; Mahdavi, J.; Carvalho, M.; Fisher, S.; Russell, S.; and Tippett, N. 2008. Cyberbullying: Its nature and impact in secondary school pupils. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, 49(4): 376–385.

Song, K.; Tan, X.; Qin, T.; Lu, J.; and Liu, T.-Y. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems*, 33: 16857–16867.

Stanford. 2022. StanfordNERTagger4.2.0. https://nlp.stanford.edu/software/. [Online; accessed 16-April-2024].

Sun, C.; Qiu, X.; Xu, Y.; and Huang, X. 2020. How to Fine-Tune BERT for Text Classification? In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 11856 LNAI, 194–206. Springer. ISBN 9783030323806.

surekharamireddy. 2021. Malignant Comment Classification. https://www.kaggle.com/datasets/surekharamireddy/malignant-comment-classification/. [Online; accessed 16-April-2024].

Takano, M.; Taka, F.; Ogiue, C.; and Nagata, N. 2022. Online Harassment of Celebrities and Influencers. *arXiv preprint arXiv:2210.07599*.

Takano, M.; Taka, F.; Ogiue, C.; and Nagata, N. 2024. Online Harassment of Japanese Celebrities and Influencers. *Frontiers in Psychology*, 15: 1386146.

Tan, W.-K.; and Hsu, C. Y. 2023. The application of emotions, sharing motivations, and psychological distance in examining the intention to share COVID-19-related fake news. *Online Information Review*, 47(1): 59–80.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Vogel, E. A. 2021. The State of Online Harassment.

Wang, X.; Yang, L.; Yang, J.; Wang, P.; and Lei, L. 2017. Trait anger and cyberbullying among young adults: A moderated mediation model of moral disengagement and moral identity. *Computers in Human Behavior*, 73: 519–526.

Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R. R.; and Le, Q. V. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Yi, P.; and Zubiaga, A. 2022. Cyberbullying detection across social media platforms via platform-aware adversarial encoding. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, 1430–1434.

Yi, P.; and Zubiaga, A. 2023a. Learning like human annotators: Cyberbullying detection in lengthy social media sessions. In *Proceedings of the ACM Web Conference 2023*, 4095–4103.

Yi, P.; and Zubiaga, A. 2023b. Session-based cyberbullying detection in social media: A survey. *Online Social Networks and Media*, 36: 100250.

Yosinski, J.; Clune, J.; Bengio, Y.; and Lipson, H. 2014. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems*, volume 4, 3320–3328.

Yuvaraj, N.; Srihari, K.; Dhiman, G.; Somasundaram, K.; Sharma, A.; Rajeskannan, S.; Soni, M.; Gaba, G. S.; Alzain, M. A.; and Masud, M. 2021. Nature-Inspired-Based Approach for Automated Cyberbullying Classification on Multimedia Social Networking. *Mathematical Problems in Engineering*, 2021.

## Ethics Checklist

For most authors...

1. Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes, this research question advances science without violating social contracts.
2. Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes, see the Abstract and the Introduction.
3. Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes, see the Methodology section.
4. Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes, see HDCyberbullying.
5. Did you describe the limitations of your work? Yes, see Limitations.
6. Did you discuss any potential negative societal impacts of your work? Yes, see Ethics Statement.

7. Did you discuss any potential misuse of your work? Yes, see Ethics Statement.
8. Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? NA
9. Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes, we read and ensure our paper conforms to them.

Additionally, if your study involves hypotheses testing...

1. Did you clearly state the assumptions underlying all theoretical results? Yes, see Introduction.
2. Have you provided justifications for all theoretical results? Yes, see Results, Error Analysis and Insights from EAT.
3. Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? Yes, see Results
4. Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? Yes, see Experiment
5. Did you address potential biases or limitations in your theoretical framework? Yes, see Theoretical Analysis
6. Have you related your theoretical results to the existing literature in social science? Yes, see Related Work
7. Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? Yes, see Conclusion and limitations

Additionally, if you are including theoretical proofs...

1. Did you state the full set of assumptions of all theoretical results? Yes, see Results
2. Did you include complete proofs of all theoretical results? Yes, see Results

Additionally, if you ran machine learning experiments...

1. Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes, we included source code and related URL.
2. Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes, see Experiments
3. Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? No, Our results are reported as average results after randomly running five times
4. Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes, see Experiments Settings
5. Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes, see Experiment Settings and Results
6. Do you discuss what is "the cost" of misclassification and fault (in)tolerance? NA

Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

1. If your work uses existing assets, did you cite the creators? Yes, we cited them all.
2. Did you mention the license of the assets? No, because it's not stated in the original sources, but we cited.
3. Did you include any new assets in the supplemental material or as a URL? Yes, we included source code and new dataset
4. Did you discuss whether and how consent was obtained from people whose data you're using/curating? NA
5. Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes, we have withheld all personally identifiable information.
6. If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see **?**)? Yes, see HDCyberbullying.
7. If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see **?**)? Yes, we released our datasets.

Additionally, if you used crowdsourcing or conducted research with human subjects...

1. Did you include the full text of instructions given to participants and screenshots? Yes, see HDCyberbullying.
2. Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA
3. Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA
4. Did you discuss how data is stored, shared, and deidentified? NA