Cyberbullying detection across social media platforms via platform-aware adversarial encoding

Peiling Yi, Arkaitz Zubiaga

Queen Mary University of London, UK
{p.yi, a.zubiaga}@qmul.ac.uk

Abstract

Despite the increasing interest in cyberbullying detection, existing efforts have largely been limited to experiments on a single platform and their generalisability across different social media platforms has received less attention. We propose XP-CB, a novel cross-platform framework based on Transformers and adversarial learning. XP-CB can enhance a Transformer leveraging unlabelled data from the source and target platforms to come up with a common representation while preventing platform-specific training. To validate our proposed framework, we experiment on cyberbullying datasets from three different platforms through six cross-platform configurations, showing its effectiveness with both BERT and RoBERTa as the underlying Transformer models.

Introduction

Cyberbullying is a form of bullying that is perpetrated through online devices (Smith et al. 2008). With the growth in usage of digital devices and Internet platforms such as social media, cyberbullying has become a major problem worldwide (Nixon 2014). This has motivated research in cyberbullying detection as the predictive task aiming to identify cyberbullying posts for enabling harm prevention (Rosa et al. 2019).

Despite increasing efforts in furthering research in cyberbullying detection, existing methods have been predominantly investigated in a single social media platform. There is however increasing evidence that classifiers built for and tested on a particular social media platform tend to underperform when applied to new platforms (Yin and Zubiaga 2021), limiting their generalisability. Generalisability of models to other platforms has been barely studied, not least in the context of cyberbullying detection (Mladenović, Ošmjanski, and Stanković 2021). Recent models for contextualised embeddings based on Transformer models, such as BERT (Devlin et al. 2019) and RoBERTa (Liu et al. 2019), are promising alternatives that can provide some generalisability through their ability to transfer knowledge. Still, they have been shown to struggle in situations where there is a big drift from source to target data (Sun et al. 2019).

To increase the potential of Transformer models when applied to a different social media platform, we propose XP-CB, a novel platform-aware adversarial framework for cross-platform cyberbullying detection. By training a classifier on a source platform, for which labelled data is available, we aim to test its ability to generalise to a new target platform, for which labelled data is lacking. To mitigate the effects of platform data shift, the core intuition of our proposed framework is to combine a multi-Transformer embedding alignment strategy with an adversarial network to reconstruct the target Transformer encoder. The target Transformer encoder is forced to map the target input to the source Transformer latent representation space, achieving more similar content representations for both source and target platforms. The classifier trained on the source platform data can be subsequently applied to the target platform. The contributions of this paper can be summarised as follows:

- We propose XP-CB, which is to the best of our knowledge the first framework for cross-platform cyberbullying detection, operationalised by furthering the potential of a Transformer model through the integration of an adversarial network.
- To assess the potential of XP-CB, we perform cyberbullying detection experiments in six cross-platform configurations involving three datasets from Formspring, Twitter and Wikipedia. These platforms present very different characteristics, particularly when it comes to the lenght of the posts.
- We show that XP-CB can achieve state-of-the-art performance by consistently outperforming a competitive model as well as vanilla Transformer models.

Related work

Cyberbullying detection

Early methods to cyberbullying detection relied on rule-based methods (Mahmud, Ahmed, and Khan 2008; Nahar et al. 2014), focused on feature engineering (Hosseinmardi et al. 2015) and used lexica (Dadvar et al. 2012; Van Hee et al. 2018). More recent methods use word embeddings along with deep learning models to build more discriminative models, leading to improved performance (Yuvaraj et al. 2021; Cheng et al. 2020). Still, this research predominantly focuses on tackling cyberbullying on a single platform, which limits the potential of transferring existing mod-

els to new, unseen social media platforms where labelled data is lacking.

Cross-platform cyberbullying detection is still in its infancy, which was pioneered by (Agrawal and Awekar 2018). 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 highlighted the challenging nature of the problem, finding that the three datasets exhibit different forms of cyberbullying with limited feature overlap across platforms. By using a Bidirectional LSTM (BiLSTM) model coupled with lists of swear words to enable transferability, they still showed a substantial performance drop from running experiments within a platform, to switching to experiments across platforms. To our knowledge, the only other work on cross-platform cyberbullying detection is that by Dadvar and Eckert (2020), who further tested the above BiLSTM model on a new platform, YouTube, leading to similar findings and highlighting the need for better models that further generalisability across platforms.

Rather than finding an overlap of features, our aim with XP-CB is to enable Transformer models the capacity of defining a latent feature space that reconciles the differences between the source and target platforms.

Adversarial networks

Adversarial adaptation methods have become increasingly popular for domain adaptation, which seek to minimise the variance between source and target data through an adversarial objective (Tzeng et al. 2017). These methods are motivated by Generative Adversarial Networks (Creswell et al. 2018), which consist of two parts: (1) the Generator is trained to generate synthetic instances in a way that confuses the discriminator; and (2) the Discriminator, responsible in turn for trying to distinguish the samples created by the generator. In the process of adversarial adaptation across divergent data sources, the roles of synthetic instances and real instances can be replaced with training samples and test samples, i.e. in cross-platform experiments, the role of the discriminator becomes that of distinguishing if an instance belongs to the source or target platform (Creswell et al. 2018).

Inspired by this trend, we propose the integration of a Transformer with an adversarial adaptation component for the cross-platform cyberbullying detection task. We build on ADDA (Adversarial Discriminant Domain Adaptation) (Tzeng et al. 2017) as the adversarial component. ADDA is a general network that enables combining a discriminative model, weight sharing, and a GAN loss for effectively training a robust and adaptive Deep Neural Network (DNN).

XP-CB: Model architecture

XP-CB is an end to end framework (see Figure 1), whose components can be trained at different times. The overarching objective of XP-CB is to perform dual alignment, which are operationalised by different components:

- **The embedding alignment** is responsible for injecting different cross-platform fine-tuning strategies into the framework components, which aims to improve the encoder's adaptability to new platforms. It includes three subcomponents: Input Length Optimiser, Hidden States Selector and Adaptive Batch Normalisation classifier.
- **The Adversarial alignment** is responsible for integrating the GAN methodology and the ADDA framework to align the target input representation to the source latent embedding space, which include the Encoder Measurer and Discriminator subcomponents.

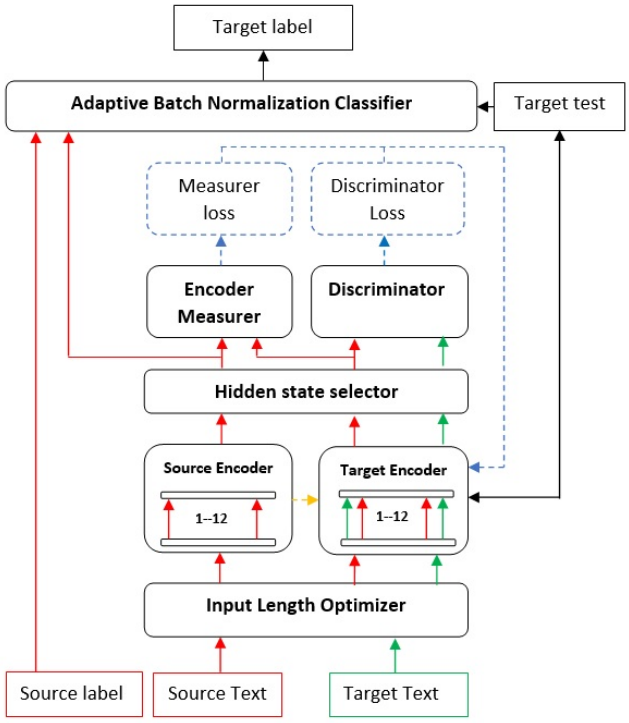

Figure 1: XP-CB Model Architecture. Red: source platform data flow. Green: target platform data flow. Black: test data flow. Blue (dashed): how the loss is fed back to the back propagation algorithm. Orange (dashed): source encoder parameters, used to initialise the target encoder.

Input Length Optimiser. To deal with the different lengths across platforms, a possible approach would be to truncate the input content or take the average input length. However, the text length can vary significantly across social media platforms, which may lead to missing important information and to an increase in the divergence between the inputs. To optimise the input length, we add the Input Length Optimiser component that handles this that iterates through different lengths in search of the optimal value.

Source Encoder & Target Encoder. The two encoders for source and target platforms are based on Transformers which are trained multiple times while being incrementally adapted. First, using the labelled source dataset to train the source encoder. Second, the source encoder will fully or partially share parameters with the target encoder for initialisation. Finally, the source and target encoders will jointly

perform adversarial adaptive training.

Hidden State Selector. The Hidden State Selector assesses the transferability of each layer in the Transformer to obtain the most transferable layers for pre-adversarial training and post-adversarial training.

Discriminator & Encoder Measurer. These two components are combined to form an adversarial network to train the target encoder. The learning goal consists in reconstructing the target encoder to map the target input representation to the source input space, making it difficult for the discriminator to determine which platform the input comes from.

The discriminator consists of two fully-connected layers on top of the encoders. The two-layer feed forward network is designed with Rectified Linear Unit (ReLU) activation and 512 or 3072 hidden sizes for the first layer and Softmax activation for the output layer. We then adopt a supervised loss function from the ADDA framework (Tzeng et al. 2017).

During the experiments, we observed that when there is a big platform data shift, gradient vanishing is common, such as in the case of transferring between Twitter and Wikipedia. To solve this issue, we add the Encoder Measurer component. Its learning goal is to get a similar hypothesis when the target encoder and the source encoder confront the same source datasets. As the loss function, we adopt the Kullback–Leibler divergence (KLD) metric (Eguchi and Copas 2006). These losses (Discriminator loss and Encoder Measurer loss) are then joined to train the target encoder.

Adaptive Batch Normalisation classifier. The Adaptive Batch Normalisation (BN) classifier aims to reduce the distribution difference between the source and the target data by adjusting the dimensionality of input representations from source and target platforms. Similar to the discriminator structure, a two-layer feed forward network is designed by using ReLU activation. We adopt two methods to build the first layer. Reduction consists in reducing the hidden states to 512, while expansion consists in expanding it to 3072. For the output layer, we use Softmax as the activation function. Batch normalisation is added to standardise these inputs and reduce the generalisation error, so as to increase the generalisation ability of the classifier (Li et al. 2017).

Experiments

Datasets

We evaluate XP-CB on three widely-studied cyberbullying datasets[1] from three social media platforms: Formspring (Reynolds, Kontostathis, and Edwards 2011), Twitter (Waseem and Hovy 2016) and Wikipedia (Wulczyn, Thain, and Dixon 2017). Where datasets provide finer-grained labels for types of cyberbullying, we collapse them into a cyberbullying label. The text length of each platform (see Table 1) varies greatly from a maximum length of 38 to 2,846; so does the cyberbullying ratio ranging from 0.08 to 0.32.

[1]We restrict to cyberbullying datasets avoiding conflation with related phenomena such as hateful / toxic / abusive content.

Table 1: Dataset statistics.

	Formspring	Twitter	Wikipedia
#Posts	12,773	16,090	115,864
Max Length	1099	38	2832
Cyberbullying Ratio	0.08	0.32	0.11

Experiment Setup

We set up experiments in line with previous work (Agrawal and Awekar 2018; Dadvar and Eckert 2020). We conduct six cross-platform configurations for our experiments by defining all six possible combinations of source-target dataset pairs. We focus on zero-shot settings, where the model doesn't see any labelled instances of the target platform. For a fair comparison with previous work, we adopt the same approach to mitigate the class imbalance by over-sampling the training data from the bullying class thrice.

Transformer models. We test XP-CB with two different Transformer models: BERT_base (uncased) and RoBERTa_base. We use the hyper-parameters recommended by (Sun et al. 2019); Batch size: 16; Learning rate (Adam): 2e -5; Number of epochs: 4.

Baseline models. We compare our models with three competitive baseline models: (1) the cross-platform BiLSTM model with attention by Agrawal and Awekar (2018), (2) BERT_base (uncased) and (3) RoBERTa_base.

Results

Table 2 shows the Macro-averaged F1 (Macro-F1) scores of all models under study, including the state-of-the-art model by Agrawal and Awekar (2018) (A&A) and the baseline Transformer models, BERT and RoBERTa.

	Baselines			XP-CB	
Source→Target	A&A	BERT	RoBERTa	**-BERT**	**-Roberta**
In-platform					
TW → TW	0.93	0.95	0.86	–	–
WP → WP	0.87	0.86	0.88	–	–
FS → FS	0.91	0.88	0.87	–	–
Average	0.903	0.897	0.870	–	–
Cross-platform					
FS → TW	0.03	0.43	0.46	**0.58**	**0.61**
WP → TW	0.28	0.47	0.51	**0.53**	**0.56**
FS → WP	0.35	0.74	0.78	**0.81**	**0.82**
TW → WP	0.10	0.54	0.56	**0.60**	**0.60**
TW → FS	0.07	0.63	0.66	**0.71**	**0.71**
WP → FS	0.58	0.78	0.78	**0.88**	**0.86**
Average	0.235	0.598	0.625	**0.685**	**0.693**

Table 2: In-platform and cross-platform classification results. **FS**: Formspring; **WP**: Wikipedia; **TW**: Twitter.

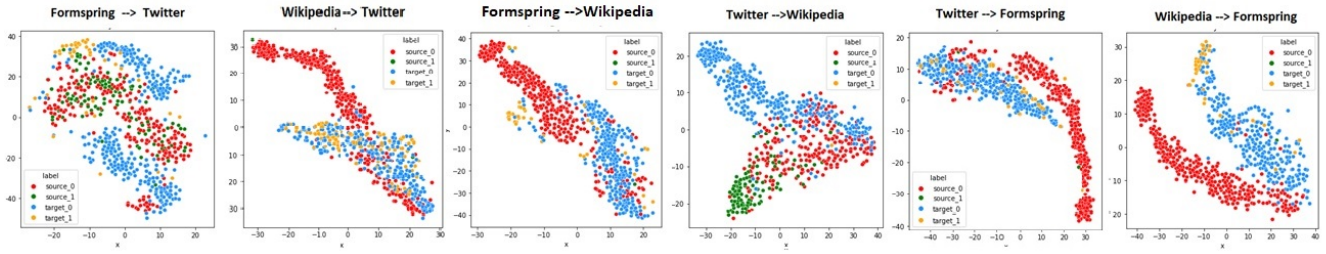

Figure 2: BERT t-SNE. Red:source negative;Green:source positive; Blue:target negative; Yellow: target positive.

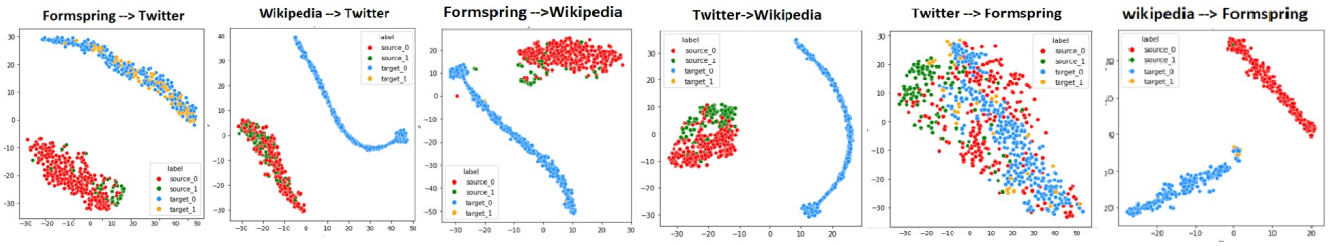

Figure 3: XP-CB t-SNE. Red:source negative;Green:source positive; Blue:target negative; Yellow:target positive.

A look at the in-platform classification results shows the strong potential of the A&A model, achieving slightly better average performance than Transformer models BERT and RoBERTa, despite this improvement not being consistent for all datasets. While BERT and RoBERTa models perform best for Twitter and Wikipedia respectively, it is the A&A model that achieves the highest performance on Formspring. The high performance scores of these three models however drops substantially when we look at cross-platform experiments, with absolute performance drops of 67% (A&A), 30% (BERT) and 25% (RoBERTa) when we look at average performances. While RoBERTa demonstrates to be the best of the three models for cross-platform transfer, its performance is still seriously impacted.

We observe that the proposed XP-CB framework boosts this performance in all six cross-platform configurations, demonstrating its ability to further the cross-platform transferability of both BERT and RoBERTa encoders. XP-CB provides an absolute improvement of 9% when we use BERT as the underlying encoder and an absolute improvement of 7% when we use RoBERTa. These improvements are consistent across all six configurations, where the differences between XP-CB-BERT and XP-CB-RoBERTa are generally more marginal. We observe better overall performance when crossing between Formspring and Wikipedia in either direction, potentially due to the lengthier posts in both cases. Performance is lower for configurations involving Twitter, where the length is much shorter.

To assess the effectiveness of XP-CB in inferring representations that reconcile source and target platforms, we visualise the resulting embeddings for the six cross-platform configurations by using tSNE (t-Distributed Stochastic Neighbour Embedding) (Van der Maaten and Hinton 2008). Figure 2 displays the embeddings generated by the BERT_base model.[2] In some of the configurations (FS $\rightarrow$ TW, WP $\rightarrow$ TW and TW $\rightarrow$ WP), we can observe that the data points of different clusters in the source and target platforms are mixed together, which shows that a model trained on the source platform labelled data using only BERT is not enough for the target platform classification. Regarding the other three cross-platform configurations (FS $\rightarrow$ WP, TW $\rightarrow$ FS, WP $\rightarrow$ FS), the data points have begun to move closer to their own clusters. Although the boundaries of each group are not so clear, this shows that BERT has started to have some platform awareness.

Figure 3 shows the XP-CB embeddings. We can observe a clearer separation of classes with respect to the BERT embeddings, which demonstrates the increased platform-awareness of XP-CB. Especially on Wikipedia and Formspring, samples originating from different platforms are spatially separated. Along with the improved performance scores, visualisation of embeddings demonstrates the potential of XP-CB to bring the representations of source and target social media platforms closer to each other.

Conclusion

We have proposed XP-CB, a novel framework for cyberbullying detection in settings hitherto largely overlooked, i.e. across different social media platforms through zero-shot settings. Building on Transformer models BERT and RoBERTa, our framework couples its fine-tuning capacity with adversarial learning to enable cross-platform transfer. Through experiments on six cross-platform configurations, our study demonstrates the consistent effectiveness of XP-CB to outperform competitive baselines, including the state-of-the-art cross-platform cyberbullying detection model.

[2]We focus these visualisations on the BERT embeddings, rather than the RoBERTa embeddings, due to the limited space.

References

Agrawal, S.; and Awekar, A. 2018. Deep learning for detecting cyberbullying across multiple social media platforms. In *Proceedings of the European Conference on Information Retrieval*, 141–153. Springer.

Cheng, L.; Guo, R.; Candan, K. S.; and Liu, H. 2020. Representation learning for imbalanced cross-domain classification. In *Proceedings of the 2020 SIAM international conference on data mining*, 478–486. SIAM.

Creswell, A.; White, T.; Dumoulin, V.; Arulkumaran, K.; Sengupta, B.; and Bharath, A. A. 2018. Generative Adversarial Networks: An Overview. *IEEE Signal Processing Magazine*, 35(1): 53–65.

Dadvar, M.; and Eckert, K. 2020. Cyberbullying detection in social networks using deep learning based models. In *International Conference on Big Data Analytics and Knowledge Discovery*, 245–255. Springer.

Dadvar, M.; Ordelman, R.; De Jong, F.; and Trieschnigg, D. 2012. Improved cyberbullying detection using gender information. In *Dutch-Belgian Information Retrieval Workshop, DIR 2012*, 23–26.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*, 4171–4186.

Eguchi, S.; and Copas, J. 2006. Interpreting kullback–leibler divergence with the neyman–pearson lemma. *Journal of Multivariate Analysis*, 97(9): 2034–2040.

Hosseinmardi, H.; Mattson, S. A.; Rafiq, R. I.; Han, R.; Lv, Q.; and Mishra, S. 2015. Analyzing labeled cyberbullying incidents on the instagram social network. In *International conference on social informatics*, 49–66. Springer.

Li, Y.; Wang, N.; Shi, J.; Liu, J.; and Hou, X. 2017. Revisiting batch normalization for practical domain adaptation. *5th International Conference on Learning Representations, ICLR 2017 - Workshop Track Proceedings*, 1–12.

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*.

Mahmud, A.; Ahmed, K. Z.; and Khan, M. 2008. Detecting flames and insults in text. In *Proceedings of 6th International Conference on Natural Language Processing*, 1–10.

Mladenović, M.; Ošmjanski, V.; and Stanković, S. V. 2021. Cyber-aggression, cyberbullying, and cyber-grooming: a survey and research challenges. *ACM Computing Surveys (CSUR)*, 54(1): 1–42.

Nahar, V.; Al-Maskari, S.; Li, X.; and Pang, C. 2014. Semi-supervised learning for cyberbullying detection in social networks. In *Australasian Database Conference*, 160–171. Springer.

Nixon, C. L. 2014. Current perspectives: the impact of cyberbullying on adolescent health. *Adolescent health, medicine and therapeutics*, 5: 143.

Reynolds, K.; Kontostathis, A.; and Edwards, L. 2011. Using machine learning to detect cyberbullying. In *Proceedings of ICMLA workshops*, volume 2, 241–244. IEEE.

Rosa, H.; Pereira, N.; Ribeiro, R.; Ferreira, P. C.; Carvalho, J. P.; Oliveira, S.; Coheur, L.; Paulino, P.; Simão, A. V.; and Trancoso, I. 2019. Automatic cyberbullying detection: A systematic review. *Computers in Human Behavior*, 93: 333–345.

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*, 49(4): 376–385.

Sun, C.; Qiu, X.; Xu, Y.; and Huang, X. 2019. How to fine-tune bert for text classification? In *China National Conference on Chinese Computational Linguistics*, 194–206. Springer.

Tzeng, E.; Hoffman, J.; Saenko, K.; and Darrell, T. 2017. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7167–7176.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Van Hee, C.; Jacobs, G.; Emmery, C.; Desmet, B.; Lefever, E.; Verhoeven, B.; De Pauw, G.; Daelemans, W.; and Hoste, V. 2018. Automatic detection of cyberbullying in social media text. *PloS one*, 13(10): e0203794.

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

Wulczyn, E.; Thain, N.; and Dixon, L. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the World Wide Web conference*, 1391–1399.

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

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.