Offensive Language Detection on Video Live Streaming Chat

Zhiwei Gao, **Shuntaro Yada**, **Shoko Wakamiya**, and **Eiji Aramaki**

Nara Institute of Science and Technology, Japan
{gao.zhiwei.fw1,s-yada,wakamiya,aramaki}@is.naist.jp

Abstract

This paper presents a prototype of a chat room that detects offensive expressions in a video live streaming chat in real time. Focusing on Twitch, one of the most popular live streaming platforms, we created a dataset for the task of detecting offensive expressions. We collected 2,000 chat posts across four popular game titles with genre diversity (e.g., competitive, violent, peaceful). To make use of the similarity in offensive expressions among different social media platforms, we adopted state-of-the-art models trained on offensive expressions from Twitter for our Twitch data (i.e., transfer learning). We investigated two similarity measurements to predict the transferability, textual similarity, and game-genre similarity. Our results show that the transfer of features from social media to live streaming is effective. However, the two measurements show less correlation in the transferability prediction.

1 Introduction

Posting offensive comments is a common problem of abusive behavior not only on representative social media platforms (e.g., *Twitter*) but also on other growing platforms[1]. For example, offensive expressions in live streaming chat could appear more frequently because users communicate with others in real time with less introspection during live streaming. In our preliminary study using crowdsourcing, seven out of 100 users of live streaming platforms answered that they had sent offensive expressions, while ten users answered that they had been the target of offensive behaviors. Tackling this problem on social media can provide more protection to users and more inspection methods for service providers.

Offensive language detection on *Twitter* has already attracted much attention due to Twitter's large user base and linguistic idiosyncrasies. For example, SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (Zampieri *et al.*, 2019b) was organized as a shared task for offensive language detection on *Twitter*. The task consisted of three sub-tasks: (a) offensive language identification, (b) categorization of offense types, and (c) offense target identification. To provide more diversified characteristics for research on offensive language detection, it is also crucial to focus on video live streaming platforms that share similar features to *Twitter*, such as *Twitch*[2], which is a popular live streaming platform mainly for gamers.

This paper introduces a chat room that detects offensive expressions in real time and filters them out, as illustrated in Figure 1. To detect offensive expressions, we apply a deep learning approach. Modern neural architectures for natural language processing are generally highly effective when provided with a large amount of labeled training data (Dai *et al.*, 2019). However, large volumes of labeled data are difficult to obtain, especially for video live streaming, as it is an emerging platform. Therefore, we apply a transfer learning approach to solve this problem. Domain adaptation, a sub-field of transfer learning, aims to learn a transferable representation from a source domain and apply it to a target domain (Dai *et al.*, 2019). We propose the use of (1) an existing large labeled social media dataset from *Twitter*

[1] https://cnet.co/3dSW6mH

[2] https://www.twitch.tv

Proceedings of the 28th International Conference on Computational Linguistics, pages 1936–1940
Barcelona, Spain (Online), December 8-13, 2020

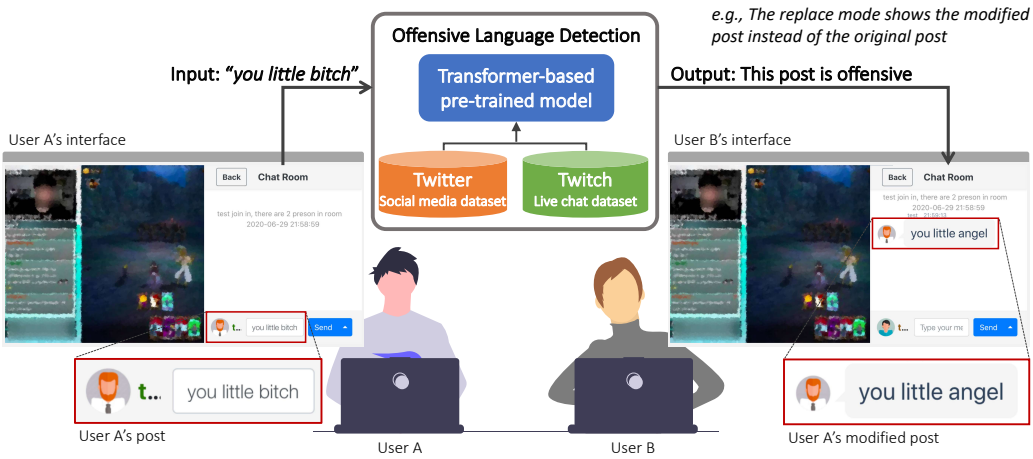


Figure 1: Proposed video live streaming chat. Given a message, the system detects whether it contains offensive expressions. If it does, the masked/replaced message is presented to the other users.

and (2) combinations of a large dataset and small datasets (live chat datasets from *Twitch*) as the source domain to demonstrate the effectiveness of transferring the obtained features to the target domain. To compare the impact of different small datasets on transferability, we assumed that the more similar the small datasets are, the higher the accuracy of transferability. We considered the textual similarity and game-genre similarity to measure transferability. In our previous paper (Gao *et al.*, 2020), we presented a discussion on the feasibility of both measurements. Then, we trained these data on transformer-based pre-trained models and chose the best-performing method to construct the chat room. To address a variety of user needs, our proposed chat room has three modes: hide, replace (Figure 1), and alert. We selected four games from *Twitch* to collect data and experimented with the effects of combination of various small datasets.

2 Offensive Language Detection

2.1 Datasets

We used the existing Offensive Language Identification Dataset (OLID) (Zampieri *et al.*, 2019a) as the social media dataset and newly created the live chat dataset.

The OLID dataset was used in SemEval-2019 Task 6 (Zampieri *et al.*, 2019b). The organizers collected tweets including specific keywords and annotated them hierarchically regarding the offensive language, offense type, and offense target. One of the sub-tasks, sub-task A, involves offensive language detection and provided 13,240 tweets as training data and 860 tweets as test data.

For the live chat dataset, we collected posts from the chat rooms of the following four games in January 2020; moreover, we collected keywords of the games from the game review sites Metacritic [3] and GamePressure[4] to measure the game-genre similarity (see Section 2.2):

- *Fortnite* (LC_{FN}): a popular survival game (keywords: action, shooter, TPP, tactical, zombie, sandbox, survival, multiplayers, cooperation, crafting)
- *Grand Theft Auto V* (LC_{GT}): an open-world game known for its violent content (keywords: action, FPP, TPP, vehicles, gangster, sandbox, shooter, modern, adventure, open-world)
- *Hearthstone* (LC_{HS}): a digital collectible card game (keywords: logic, fantasy, card game, 2 players, strategy, miscellaneous, turn-based)
- *Minecraft* (LC_{MC}): a sandbox, Lego-like game (keywords: adventure, FPP, TPP, fantasy, sandbox, RPG, elements, survival, multiplayers, cooperation, crafting, action, 3D)

We finally obtained 2,000 posts from the four game titles and annotated them as offensive or not offensive. We employed Cohen's kappa (Cohen, 1960) to measure inter-coder agreement and achieved a value of 0.77, which shows a substantial agreement between our coders.

[3] https://www.metacritic.com
[4] https://www.gamepressure.com

	Training set					TVcC	Overlap coffifient	Precision	Recall	F1-score
	Twitter	+LC_{FN}	+LC_{GT}	+LC_{HS}	+LC_{MC}					
LC_{FN}	✓					-	-	0.766	0.832	0.792
	✓		✓			0.363	0.400	0.784	0.802	0.793
	✓			✓		0.327	0	0.773	0.808	0.788
	✓				✓	0.349	0.700	0.805	0.791	0.798
	✓		✓	✓		0.472	0.400	0.782	0.794	0.788
	✓		✓		✓	0.475	0.800	0.822	0.787	**0.802**
	✓			✓	✓	0.456	0.700	0.800	0.786	0.792
	✓		✓	✓	✓	**0.537**	0.800	0.806	0.774	0.788
LC_{GT}	✓					-	-	0.686	0.780	0.715
	✓	✓				0.349	0.400	0.712	0.781	0.738
	✓			✓		0.306	0	0.717	0.791	0.745
	✓				✓	0.321	0.500	0.719	0.752	0.734
	✓	✓		✓		0.446	0.400	0.722	0.780	0.745
	✓	✓			✓	0.442	0.600	0.734	0.745	0.739
	✓			✓	✓	0.426	0.500	0.751	0.769	**0.759**
	✓	✓		✓	✓	**0.504**	0.600	0.736	0.761	0.747
LC_{HS}	✓					-	-	0.729	0.778	0.749
	✓	✓				0.315	0	0.749	0.768	0.758
	✓		✓			0.307	0	0.764	0.770	**0.768**
	✓				✓	0.305	0.143	0.778	0.749	0.762
	✓	✓	✓			0.412	0	0.773	0.763	0.767
	✓	✓			✓	0.408	0.143	0.794	0.742	0.764
	✓		✓		✓	0.409	0.143	0.790	0.740	0.761
	✓	✓	✓		✓	**0.470**	0.143	0.789	0.741	0.762
LC_{MC}	✓					-	-	0.611	0.777	0.632
	✓	✓				0.335	0.700	0.620	0.765	0.648
	✓		✓			0.320	0.500	0.630	0.767	0.660
	✓			✓		0.304	0.143	0.627	0.772	0.657
	✓	✓	✓			0.427	0.692	0.632	0.771	0.664
	✓	✓		✓		0.427	0.615	0.630	0.763	0.660
	✓		✓	✓		0.423	0.462	0.633	0.756	0.664
	✓	✓	✓	✓		**0.485**	0.769	0.639	0.770	**0.672**

Table 1: Results of RoBERTa model using live chat datasets. The left column shows the test set. In the training set columns, the training set used in the current test set is checked. We used "*Twitter* + *X*" to denote our training sets, where Twitter is the baseline, and "*X*" refers to its combinations with LC_{FN}, LC_{GT}, LC_{HS}, and LC_{MC}. The TVcC column is the textual similarity of chat posts, and the overlap coefficient column is the game-genre similarity.

2.2 Model and Similarity Measures

We adopted the following transformer-based pre-trained models: BERT (Devlin *et al.*, 2019), RoBERTa (Liu *et al.*, 2019), and XLNet (Yang *et al.*, 2019), which have performed well on many natural language processing tasks. We trained each model with 5 epochs and a batch size of 64. After training on the social media dataset, RoBERTa achieved the best performance (0.795 in F1-score). Therefore, we selected RoBERTa as our model.

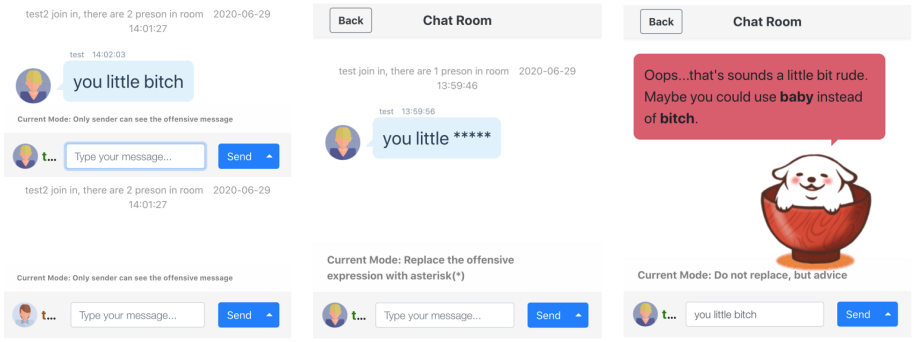
For the similarity measures, we adopted the following approaches:

Textual similarity: Because similar expressions in chat posts contribute to transfer learning, we employed Target Vocabulary Covered (TVcC) (Dai *et al.*, 2019) to measure the textual similarity among our live chat datasets. TVcC indicates the percentage of target vocabulary that is also presented in the source datasets, and it only considers content words (nouns, verbs, and adjectives). It is calculated as $TVcC\left(D_{\text{S}}, D_{\text{T}}\right) = \frac{\left|V_{\text{D}_{\text{S}}} \cap V_{\text{D}_{\text{T}}}\right|}{\left|V_{\text{D}_{\text{T}}}\right|}$, where V_{D_S} and V_{D_T} are sets of unique tokens in the source and target datasets, respectively.

Game-genre similarity: Under the assumption that similar target (game) content would result in similar offensive expressions, we employed an overlap coefficient of the target review keywords as the similarity in target content, calculated as $overlap\left(R_{\text{A}}, R_{\text{B}}\right) = \frac{|R_{\text{A}} \cap R_{\text{B}}|}{min(|R_{\text{A}}|, |R_{\text{B}}|)}$, where R_A and R_B are the keywords for two different target reviews. A higher overlap coefficient indicates a higher similarity between the two targets.

2.3 Performance

We evaluated the performance of the models based on precision, recall, and macro F1-score. We report the mean values of F1-scores over five repetitive experiments. Table 1 lists the results of the best performing model (i.e., RoBERTa) on our live chat datasets. The first row for each live chat dataset is the baseline, which only uses the social media dataset as the training set. To reduce the high similarity caused by excessive word coverage, we did not add the social media dataset when comparing TVcC and



(a) Hide mode (b) Replace mode (c) Alert mode

Figure 2: Three variations of the implementation of the proposed offensive detection

overlap coefficient and only compared the similarities between each live chat dataset. However, Table 1 indicates that even when we combined all the *Twitch* datasets, we may not obtain the best performance (F1-score). For instance, for LC_{HS}, the best F1-score was achieved only for the combination of *Twitter* + LC_{GT}. Therefore, the two similarity measurements of TVcC and overlap coefficient do not predict transferability as well.

3 Demonstration System

We created an online chat room for offensive expression detection. The best model, RoBERTa, and the training set with the best results over all experiments, the combination of *Twitter* + LC_{GT} + LC_{MC} (0.802 in F1-score), were used to construct the chat room. To make full use of this model, we equipped our chat room with three modes to handle different scenarios and needs. We used sockets for the chat room communication function. The chat room is available online [5].

- **Hide Mode:** When an offensive post is detected, this mode does not send the post to the socket channel; it displays it only on the sender's client, as shown in Figure 2(a).
- **Replace Mode:** When an offensive post is detected, this mode shows a modified post in which offensive expressions are masked or replaced.
  After identifying offensive words by segmenting the input sentence into word granularity, they are masked with asterisks or replaced with inoffensive candidates based on the part of speech (POS) of these words, as shown in Figure 2(b). We first determined a few inoffensive words and then collected our candidates by selecting the synonyms of these words.
- **Alert Mode:** In this mode, if offensive expressions are detected in a post, the sender is alerted before the post is sent to the socket channel. A virtual character[6] gently urges the user to rephrase the offensive expressions, as shown in Figure 2(c). Offensive expressions are identified in the same way as in replace mode.

4 Conclusion

This paper presented a chat room for video live streaming platforms that detects offensive expressions and filters them out via three modes (hide, replace, and alert). We employed transfer learning from *Twitter* to video live streaming posts on *Twitch*. Our results demonstrate that the transfer learning approach improves performance on the offensive language detection task for video live streaming. Moreover, we investigated two measurements to quantify transferability. One limitation of this research is that the comparison of game-genre (game content) similarity is still challenging. In future work, we will collect more data on a wider variety of games. We will also consider assessing the similarity of the datasets based on embedded vectors rather than surface textual information, such as word occurrence, as used in this research. The idea of updating existing training data for new data (new games) has much room for future study.

[5] https://aoi.naist.jp/chatRoom/
[6] We adopted Wankoromochi (http://wankoromochi.com).

References

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, **20**(1), 37–46.

Dai, X., Karimi, S., Hachey, B., and Paris, C. (2019). Using Similarity Measures to Select Pretraining Data for NER. In *NAACL-HLT*.

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

Gao, Z., Yada, S., Wakamiya, S., and Aramaki, E. (2020). A preliminary analysis of offensive language transferability from social media to video live streaming. *Proceedings of the Annual Conference of JSAI*, **JSAI2020**, 1K3ES204–1K3ES204.

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*, **abs/1907.11692**.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J. G., Salakhutdinov, R., and Le, Q. V. (2019). XLNet: Generalized Autoregressive Pretraining for Language Understanding. In *NeurIPS*.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., and Kumar, R. (2019a). Predicting the Type and Target of Offensive Posts in Social Media. In *NAACL-HLT*.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., and Kumar, R. (2019b). SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). In *SemEval@NAACL-HLT*.