Online Social Networks and Media 27 (2022) 100180

Contents lists available at ScienceDirect

Online Social Networks and Media

journal homepage: www.elsevier.com/locate/osnem

Understanding and identifying the use of emotes in toxic chat on Twitch

Jaeheon Kim [a,b,d], Donghee Yvette Wohn [c,*], Meeyoung Cha [d,a,*]

[a] School of Computing, Korea Advanced Institute of Science and Technology, Daehak-ro 291, Yuseong-gu, Daejeon, 34141, South Korea
[b] Hyperconnect, Yeongdong-daero 517, Gangnam-gu, Seoul, 06164, South Korea
[c] New Jersey Institute of Technology, 323 Martin Luther King Jr Blvd., Newark, NJ 07103, United States
[d] Data Science Group, Institute for Basic Science, Expo-ro 55, Yuseong-gu, Daejeon, 34126, South Korea

ARTICLE INFO

Keywords:
Visual toxic chat
Usages
Live streaming
Emotes
Detection
Algorithmic moderation
Twitch

ABSTRACT

The latest advances in NLP (natural language processing) have led to the launch of the much needed machine-driven toxic chat detection. Nevertheless, people continuously find new forms of hateful expressions that are easily identified by humans, but not by machines. One such common expression is the mix of text and emotes, a type of visual toxic chat that is increasingly used to evade algorithmic moderation and a trend that is an understudied aspect of the problem of online toxicity. This research analyzes chat conversations from the popular streaming platform Twitch to understand the varied types of visual toxic chat. Emotes were sometimes used to replace a letter, seek attention, or for emotional expression. We created a labeled dataset that contains 29,721 cases of emotes replacing letters. Based on the dataset, we built a neural network classifier and identified visual toxic chat that would otherwise be undetected through traditional methods and caught an additional 1.3% examples of toxic chat out of 15 million chat utterances.

1. Introduction

Live video casting services have become a popular Internet destination [1]. Twitch is one such service used to stream online games and has expanded into various topics like music, cooking, and in-real-life [2]. It has become a prominent and influential social media platform with a disproportionately young user base (e.g., 16-to −24 years and 25-to −34 years make up 41% and 32% in 2019) [3]. Unlike traditional TV broadcasts where content is highly editorialized, Internet streaming can contain inappropriate scenes and uncensored interaction among users [4]. In particular, viewers can participate in the broadcast via publicly visible chats in real-time, which leads to a challenge in regulating and monitoring massive amounts of live streams and the accompanying live conversations [5].

Toxic chat, a hateful speech or an offensive language via chat, is known to have a negative effect on communication and emotion [6–8]. Twitch channels commonly use two main methods to fight toxic chat: human moderators and AI [9,10]. Human moderators provide sophisticated support and can understand the subtle context of hateful comments. Nonetheless, this approach is not scalable; it is not easy to monitor the explosive amount of chats over extended hours [11], and moderation comes with many emotional tolls [10]. AI is used in moderation algorithms in the Twitch system as well as third-party chatbots that can prevent certain words from being typed or delete chats that match the predefined rule set on known toxic expressions [9]. However, these chatbots are not able to catch a nuanced tone or misspellings that are common tactics used to thwart algorithms [12]. Some chat users are increasingly exploiting AI's weakness by substituting letters with emotes and producing conversations that can only be understood by humans [13]. As increased calls for open data in our research community [14,15], our dataset will be released for further discussion.

In Twitch, chat is the way that enables communication between broadcasters and viewers. It is textual, but it is also used in combination with visual aids. In addition to the emoji, there are visual aids used only within the Twitch platform, which are called *emotes*. Although they provide a diversity of expression and support richer expression by utilizing visual effects, the problem is that these emotes are sometimes used for toxic chats, making them difficult to detect. Sometimes viewers use emotes to attack others or encrypt text metaphorically.

The first image in Fig. 1 shows an example of a toxic conversation where two consonants are replaced with the face emotes of a black streamer. The image on the right shows a more nuanced and subtle example, where a KFC chicken emote that was initially launched as a promotion for the franchise appears next to face emotes of black streamers. This combination of two benign emotes now suggests a racist stereotype and has been flagged as 'toxic' [16]. While these cases are easy to identify by humans, they are non-trivial for machines to detect.

* Corresponding authors.
E-mail addresses: jayden.k@hpcnt.com (J. Kim), Wohn@njit.edu (D.Y. Wohn), mcha@ibs.re.kr (M. Cha).

https://doi.org/10.1016/j.osnem.2021.100180
Received 31 December 2020; Received in revised form 27 September 2021; Accepted 8 October 2021
Available online 15 November 2021

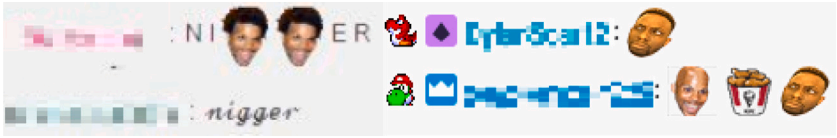


Fig. 1. Example of using emotes of a Black streamer's face to replace letters in a word that would otherwise be automatically banned by the system (left) and a mix of fried chicken and a different Black streamers' face emotes that is offensive in context (right).

In this paper, we conduct two studies to understand the use of emotes in toxic chat well. In study 1, we explore Twitch chat and emotes to find how the emotes are used with toxic chat. In study 2, we target one specific situation, replacing a letter with an emote, analyze it more deeply, and suggest a solution. This research contributes to a better understanding of chat conversations containing such visual toxic chats. We crawled a large amount of chat data and extracted community-specific emotes. Then, we hired crowdsourced workers to label cases where emotes are substituting letters. We built a neural network model that restores the original word from the incomplete text and emotes based on the labeled dataset.

This model has two key components: First is the symbol or letter-level bidirectional long short-term memory (bi-LSTM) structure that can learn time-series data. Here, the model learns the relationship between words and can handle unordinary cases. Second is the selection component that identifies possible restorations from the list of community-specific words, followed by general words. Selection component predicts the appropriate word among the Twitch corpus. The model's final restored words were compared against a popular hate speech library, Hatesonar [13], and confirmed a 1.28% improvement in hate speech detection.

Our major contributions are as follows:

(1) We identify different usage cases of emotes used in toxic chat: replacing words, replacing letters, community membership, attention-seeking, visual enhancement, emotional expressions, and figurative expression of words.
(2) We build the gathered Twitch corpus and prediction dataset. The 11,485 cases of crowdsourced labels, where an emote is substituting a letter(s), will be a useful dataset for future research that tries to understand the hidden context in chat conversations.
(3) To detect toxic chat that mixes text and emotes, we propose a new model that predicts the original word. The sequence-based bi-directional language model is used to guess latent letters. This model also gives higher weights to community-specific expressions rather than the general word corpus.
(4) The classifier identified 196,448 additional hateful expressions from 15 million chat conversations that were tested, which would have gone undetected with existing tools. This improvement over conventional algorithms is significant, given the harm toxic chat brings.

Hateful expressions are expanding at a rapid rate. This research examined one emerging type that is a mix of text and emotes. While the proposed LSTM-based model could gain additional improvement compared to an existing detector, the current work serves as a starting point for adopting flexible neural models that outperform dictionary-based detection. We also found that emotes often replace vowel letters with consonants, likely because such forms are more comfortable to understand. These insights could help build future NLP-based models. More advanced structures (e.g., generative models) could better handle the growing toxic chat vocabulary in the future.

2. Background

2.1. Online streaming service

Twitch (www.twitch.tv), an online streaming service acquired by Amazon, broadcasts various subjects like gaming, cooking, chatting, music, and art via the Internet [17]. The platform has grown massively popular to serve 15 million daily active users, 1 million average viewers at any point, 9 million content streamers, and 44 billion watching time per month by 2018 [18].

Studies point out that Twitch's unique interfaces that allow viewers to participate in the live stream via chats, donating, and sending emotes or emojis have made the service popular [19]. Hence the audience enjoys a unique experience of viewing the streaming content itself and interacting with the streamer and other viewers in real-time. Sjöblom [5] analyzed the reason for its popularity based on 1000 user surveys and identified (1) the burst of epic moments that occur during the broadcast, (2) information about new games and their strategies, and (3) social satisfaction obtained from communicating directly with streamers, as significant drivers for success. The passive nature of watching gameplay rather than playing one also seemed to contribute to the routine use of Twitch.

Hamilton [2] analyzed Twitch channels as participatory communities. A broadcast channel made by one person creates a kind of small society. It is divided into moderators that manage the channel together and viewers who participate in the community. This small community, centered on broadcasting content, continues to develop, and broadcasting contents include communication with viewers that occur in it and games. Broadcasters generally use a visual way through the screen and an auditory style through voice for discussion, and viewers receive this signal and exchange opinions in text form named chats.

2.2. Chats analysis

Platforms provide a wide range of tools to fight hateful expression [20,21]. The most common form is a user-report system. Facebook enforces a set of banned words to users such that any post containing those words will not appear on other people's timeline [22]. Users can customize this list, and if a user sees undetected harassment, they may "hide" a post or request to show less of future posts of similar types. Twitter employs a constant check to identify accounts that produce hate speech and asks for re-verification or suspend those accounts. Individuals can report other people's tweets or block them if they see harassment [23].

On Reddit, each discussion board or subreddit has administrators who can prevent other participants who write harmful content or remove posts that violate the community rule [24]. Subreddits that were polluted too much that the human manager could no longer manage were deleted.[1] On YouTube, channel owners cannot gain revenue from reported videos, and any videos that are flagged as harmful by many audiences get deleted [25]. The service also employs algorithms to remove content promoting violence or hatred against individuals of a certain age, disability, ethnicity, gender, nationality, race, religion [26]. On Instagram, detection further targets visually harassing content [27]. Wikipedia relies on volunteers and moderators to fight against hate speech.

On Twitch, channel owners can directly manage the chat content by eliminating inappropriate messages and banning users who engage in toxic behavior. Channel owners can hire human moderators or

[1] Removing harassing subreddits, https://bit.ly/1FeFw7n.

moderation chatbots to mitigate the heavy workload of administration [9,10]. Chatbots can detect excessive usage of word capitalization, banned words, hyperlinks to external websites, spam, and repeated messages [28]. Detected harassers can be banned or temporarily timed out (e.g., 10–30 min). Nevertheless, it is a challenge to monitor live content continually.

In this paper, we focused on the Twitch platform for two reasons. First, compared to other platforms, few known sanction algorithms exist, and only simple methods such as rule-based are used. Second, other platforms are usually creating comments for already edited images, photographs, etc. However, Twitch is a form in which viewers participate in chats, not channel owners, and it is challenging to manage toxic chats because they are shown simultaneously as real-time broadcasting.

3. Study 1: Case study on toxic chat with emotes

Our study tackles to detect the harmful use of emotes, with a focus on fighting hateful expressions. Our first research question aims to characterize how emotes are used together with text, what types appear, and the tendency to understand the general tendency of emoting usage.

RQ1: How are emotes used in toxic chat and what are their various forms?

3.1. Related works for study 1: Emotes and emojis

Emojis are called *emotes* on Twitch. They are graphical digital icons in various shapes such as faces, animals, and objects like emojis. Emotes outnumber emojis by thousands. They allow expressing emotion without words, clearly and concisely, in the fast-paced chat platform [29]. They are an instrument to replace the lost nuance of gesture and tone of voice that pure text lacks [30]. Emotes play a critical role in defining identity and spreading community around Twitch; they represent group membership and express social waves [31]. Some emotes are created to show support toward streamers. While most emotes are free, some are stream-specific and unlockable perks for those who pay and subscribe to a channel. The need for channel-specific emotes has led to its growing diversity [32].

The meaning of an emote is versatile. Depending on the person who sees it, its purpose and sentiment may change. Miller [33] studied the use of emojis and found that the meaning of text could change by the insertion of particular emojis. Sometimes, text with happy sentiment can be accepted negatively (i.e., irony). The use of emojis in irony and sarcasm is not new [34], and the same goes for emotes. On Twitch, certain emotes are known to be associated with discouraging online culture, such as trolling. For instance, the emote 'kappa' that represents the face of one of the Twitch employees is used in sarcasm (see Rank 5 in Table 2). Studies have developed word embedding models such as emoji2vec [35], and similarly, new studies have proposed detection models to classify the various emote types [36]. Some research has suggested deep learning models to infer the meaning of emoji accompanying text [37].

3.2. Dataset

For data collection, we identified top-100 streamers from the public lists like TwitchMetrics[2] and examined their recent video logs. We manually inspected the country each streamer resided in and identified those streamers in English-speaking countries. Any non-individual accounts (such as e-sports brand accounts) were excluded for consistency in analysis. For the chosen 54 streamers, we gathered logs from the most recent 100 videos and their accompanying chats via the npm JavaScript package.[3] The data collection took a three-month time from March 2018. We consider each utterance (i.e., all input separated by a user entering the enter key) as the unit of conversational analysis and call this *chat*. In total, we collected 103,182,314 chats in 4333 videos from 54 channels. This finding corresponds to, on average, 23,813 chats per video. Key statistics are summarized in Table 1.

To examine toxic chat, we employed the Hatesonar library [13] that is based on a comprehensive set of words listed on Hatebase.org, as well as other 25K labeled tweets. The reason is two-fold. First, it is hard to synchronize subjective standards for hate speech. Our goal was to identify toxic chats rather than select the proper criteria for hate speech. Second, Twitch conversations are quite short and massive in volume, so creating a balanced library for detecting hate expression takes much time and cost. While Hatesonar classifies content into three categories: hate speech, offensive language, and normal, we combine all content flagged as either hate speech or offensive language as 'toxic chat' in this research. Hatesonar detected 3,946,401 chats (3.8% of all chats) in our data to correspond as toxic chats. For example, comments like '`GOD DAMN IT`', and '`YO WTF WHY YOU RACIST BRUH <emote>`' were flagged. Compared to this large number, we found a disproportionately small fraction of chats had been deleted by streamers, chatbots, or channel moderators, taking up 7141 or fewer than 0.1% of all utterances. Since the original chat message is no longer accessible online, we do not know whether the deleted messages are related to hateful speech.

Hatesonar cannot capture the new types of toxic chat that mixes text and emotes, as shown in Fig. 1. Emotes are frequently used in chats, appearing in 33.9% on average. To get a sense of what kinds of emotes look like in general conversations and hate expressions, we show the top 10 ranked emotes in Table 2. The top-ranked emotes mostly depicted faces; 47 out of the generally popular top-50 emotes were faces, including 27 human faces and 20 non-human faces such as cartoon characters and robots. Hateful utterances classified by Hatesonar also included similar emotes. The rank order was slightly different; the top-2 and top-3 emotes that accompanied toxic chat were top-7 and top-6 in general chats. Top-7 emote in hateful expressions was not in the top-10 list of public conversations. Nonetheless, it is remarkable to observe that emotes used in toxic chat are not fundamentally different from the most popular list; the LUL emote (short for *Lame Uncomfortable Laugh*) was top-1 in both of these lists.

3.3. Usage cases of emotes in live chat

We explored through examples to see how emotes are used on live stream chats. Out of about 100 million chats, one thousand two hundred random chats, including emotes were extracted. These chats were sampled indifferent to toxic chats. These chats were then qualitatively analyzed by two researchers who went through the corpus together and identified themes using a grounded theory approach that involved open coding. With the unit of analysis as one word that contains an emote, researchers labeled the data with codes that were generated based on the way in which emotes were used to disguise hateful language. Below, we describe key usages of emoted chat that we observed. Emotes can appear in toxic chat (e.g., replacing alphabets, adding racist context), and they can also be used to express emotion or aid visual context.

3.3.1. Replacement of an entire word

The first use of emotes was when the emotes were used to replace an entire word based on the visual image shown in the emote, representing 4.5% of all emoted chat usage. In innocent situations, for example, emotes of hearts would be used to replace the word "love". When used for toxic chat, however, the images depicted in the emotes were more symbolic but made clear about what word was being represented by the emote due to the context of the sentence. The Trihard emote frequently appears as a substitute for Black people, as shown in the example in Fig. 2.

[2] https://www.twitchmetrics.net/channels/popularity
[3] https://www.npmjs.com/package/twitch-chatlog

Table 1
Summary of the data collected.

Total streamers	Total videos	Total chat messages	Deleted chats (percent)	Unique chat participants	Avg chats per video	Chat with emotes	Hatesonar flagged [13]
54	4,333	103,182,314	7,141 (0.007%)	15,779,760	23,813.14	33.9%	3.82%

Table 2
Top frequently used emotes in all chats and by those chats flagged by Hatesonar [13].

Type	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5	Rank 6	Rank 7	Rank 8	Rank 9	Rank 10
All chats	LUL	ninjaPon	PogChamp	TriHard	Kappa	cmonBruh	DansGame	BibleThump	:)	WutFace
Hatesonar flagged	LUL	DansGame	cmonBruh	:)	TriHard	PogChamp	moon2A	4Head	WutFace	Kappa

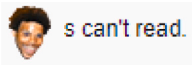


Fig. 2. A popular streamer Trihard's face emote is used most likely to refer to Black people.

Fig. 3. The letter O has been replaced with round-shaped emotes and two M's have been used to misspell 'homo'.

Fig. 4. Emotes replacing the letter E.

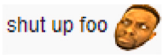


Fig. 5. The emote chosen to replace the final letter could also be an expression of exasperation.

3.3.2. Replacing a letter with an emote

The second usage example is an emote replacing a letter, representing 8.6% of all emoted chat usage. Vowels were commonly replaced by emotes because the word could still be deduced without the missing vowel. This method is an easy way to bypass blocklists for known banned words such as profanity, racist slurs, and sexist words. There were three common examples of this letter replacement. The first case is when the letters were replaced with a random emote that did not provide any context. The second case is when the shape of the picture in the emote implied the missing letter; for example, a round image is used to replace the letter `o`, as shown in Fig. 3 and thereby bypassing algorithmic moderation.

In the third case, the emote used would provide further nuance to the word. In Fig. 4, this example has an emote of an angry Asian girl replacing the letter `e`, but it could also indicate the writer's negative attitude. A weeb is a derogatory term for someone who is not Japanese but overly passionate about superficial aspects of Japanese culture, mostly based on consuming anime and manga.

When emotes were used to replace letters, there were examples where it was unclear if the emote's meaning had added meaning on top of the emote being used to replace a letter. For instance, in Fig. 5, the emote is replacing the letter `l`, but the name of the emote is 'cmonBruh' (come on, bro), so the use of this emote could also indicate exasperation.

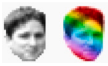


Fig. 6. Kappa (left) and KappaPride (right).

Fig. 7. Attention seeking use of emotes.

3.3.3. Community-specific emotes

Certain emotes were unique to the platform and had a special meaning that would be difficult to understand outside of the context of Twitch culture, like PogChamp's use to express excitement (Table 2, Rank 2) or Kappa's use for sarcasm (Table 2, Rank 5) [38]. Streamers can also have their personal emotes, and sometimes these took on a life of their own — coming to bear meaning unintended by the emote designer. Negative examples include 'TriHard', which happened to be used as an accompaniment to derogatory comments about or toward Black people, and 'KappaPride,' a rainbow version of Kappa, often used to accompany mean remarks about LGBTQ people. In these cases, the emotes did not have standalone meaning and would only make sense as an accompaniment to other chats. This represented 12.2% of all emoted chat usage (see Fig. 6).

3.3.4. Emotes for attention seeking

People also used emotes to visually enhance their text by placing an emote between every letter, word, or phrase so that their comment in chat would be obvious, as demonstrated in Fig. 7. Such emote usage is similar to the use of capital letters. These made up 15.7% of the data. Another method of attention-seeking was to spam emotes to take up a lot of space in the chat. However, Emote spamming was rarely used for toxic chat and was more of a way to indicate participation in a group or a type 'crowdspeak' as described in [11]. In these usage cases, the emotes themselves were usually not offensive but used to accentuate offensive comments.

3.3.5. Visual enhancement of text

Emotes were also used as pictorial representations of text, for example, having a fire emote next to text that describes something burning or an emote of a dog accompanying a comment about a dog. This represented 7.2% of all emoted chat usage. We did not see any toxic chat examples in these cases among the 1200 samples that were examined.

3.3.6. Emotes as pictures of words or emotion expression

The final use case was an emote that had literal text written inside the image. This was the most frequent use of emotes, accounting for 46.8% of the data. These emotes made it extremely clear what meaning they convey; the emotes often had eye-catching fonts, colors, or accompanying images that enhanced the text. Regular examples included "Hype", "Let's go" or "#1". We did not find any toxic chat examples in this category. Emotes of facial expressions were also used to accompany text (e.g., sad, happy, angry faces) — this was the most common use of emotes.

4. Study 2: Toxic chat detection

Study 1 identified that toxic chat is prevalent (i.e., 3.82% by Hatesonar) on Twitch. Large portion of the chat included emotes (33.9%), and usage cases of emotes were various. Because emotes are large in number and used in various ways, toxic chats with emotes cannot be detected by machines easily. Many of the emote usage cases would require an understanding of the context that would be difficult for computers to identify. However, the circumstances in which emotes were used to substitute letters could potentially be automatically detected, although they are beyond the scope of the current language models to handle. To automatically detect this new type of visual toxic chat, one needs to collect many ground truth labels and build a classifier, which is the second contribution of this research. Below, we review the recent accomplishments in toxic chat detectors and explain the method for constructing labels for the classifier.

RQ2: How can we convert emote substitution to the original text for toxic chat detection?

4.1. Related works for study 2: NLP literature on hate speech detection

Detecting cyberbullying, hate speech, toxic comments, and abusive language has been an active research topic within the NLP community. Bender and Friedman [39] emphasized the importance of proper data statements in NLP datasets that might incur bias and ethical issues. According to a comprehensive survey in [40], new methods utilized bag-of-words and N-grams for the task. Bag-of-Words is a method of digitizing the frequency of words without considering their orders. N-gram extends the word count concept by considering N consecutive words in a chain and their joint occurrences. To apply these concepts, one needs an extensive collection of data. One research studied comments from 2 million distinct online users via N-grams to detect hate speech and showed the consistent performance of detection [41].

The next series of research utilizes language models that are based on machine learning and deep learning [42,43]. For example, embedding techniques vectorize each word via examining its surrounding words, rather than merely counting the word frequency. This method can further enhance the quality of the analysis [44]. For example, Djuric [45] adopts distributed sentence embedding from Doc2Vec algorithm [42] to instill the semantic information for detecting hate speech. Recurrent neural network (RNN) and the long short term memory (LSTM) structures are also particularly well suited in handling text sequences [46]. These structures can reveal the contextual information from text sequences and show higher detection performance than those methods that look for specific patterns [47]. For instance, a word generalization technique has been used in hate speech detection, where clustering multiple words used in everyday hate speech was shown to improve the detection rate [48]. This work tried to build a classifier to recognize word clusters of positive and negative sentences. In addition to improving performance, research is actively conducted to find corner cases of hate speech [49].

Even now, research on the detection of hate speech is actively underway. A variety of attempts have been made this year, including research that combines BERT [50], the state-of-the-art in the field of natural language processing, and improvement of SVM, a traditional method [51]. Meanwhile, Gordon [52] warned against naive use of classification labels in deep learning. This work emphasized that the proportion of the population that the classifier agrees with is more important than that of the ground truth label. Also, research on a dataset that facilitates model training [53] and qualitative analysis on the phenomenon of hate speech are also active [54].

A generative adversarial network (GAN) is a technology for creating plausible objects such as drawings, images, and text. Filling an empty part of the data is done by learning features from massive amounts of data. In particular, MaskGan is a model that fills the empty part of a text sentence by examining words in common sentences [55]. Techniques like this can be used to analogize missing letters in a word sequence.

4.2. Labeling preparation

Emotes have unique names, like LUL and PogChamp, and the Twitch chat interface renders the corresponding emote images when their names are typed in chat messages. For instance, the emote-annotated expression for `NIGGER` in Fig. 1 would have been typed as '`NI TriHard TriHard ER`' by the user. Given our purpose is to build a flexible model that can restore original words from various emote-substituted chat conversations, we collected many instances of "emoted" words from chat conversations and asked people to guess what the unique words would have been.

From 103 million chats in Table 1, we pre-processed the data in the following steps. First, we removed all messages written by chatbots. As mentioned earlier, streamers used various chatbots to interact with viewers and moderate the chat room. Chatbots can be customized, but often their behaviors are recognizable because of repeated actions. To identify chatbots, we targeted chat accounts that post repetitive messages through data alignment and tested whether the poster is using some macro and hence, is a chatbot via the Twitch API. This ratio corresponded to 2.16% of all data. Second, we removed text-only or emotes-only chats because our goal is to analyze conversations involving both emotes and text, corresponding to 66.9% and 18.3% of all data, respectively. Third, we removed chats where the same emote appears three consecutive times. This step was necessary because no English words contain the same letters three or more consecutive times. Therefore, if emotes were used more than three times continuously within the chat, it is excluded. This step corresponded to 14.8% of all data. Fourth, we removed duplicated chats or chats of identical content and length. Of the total data, 3,552,397 (3.44%) of the conversations were selected in this manner.

Before launching the crowdsourcing task, a pilot test was run to gain insights for labeling. We randomly chose 600 emote-annotated chats and hired 18 internal group annotators to label these chats. At least three annotators marked each chat utterance, classifying the emote usage into either: substitution, non-substitution, and do not know. For the substitution cases, annotators made their best guesses and provided which alphabets might have been replaced. We have asked, "How familiar are we with Twitch as a preliminary question?" to associate the quality of labeling with their familiarity with Twitch.

This pilot test revealed several findings. First, many restored words were not necessarily appropriate dictionary words because people commonly use abbreviations, slang, and memes. This finding calls for a challenge in utilizing standard word dictionaries in the model. Second, familiarity with Twitch was critical in the restoration quality. Labelers, who have never used Twitch, naturally could not guess community-specific terms. One such example is Fig. 8, which led to inconsistent labels between Twitch users and non-users. Twitch users translated this chat as `DOTA`, a popular multiplayer online battle game, whereas non-users translated the same conversation as `DATA`. If a model were to utilize a dictionary matching scheme, we would have missed the community-specific context. This finding calls for the importance of knowing the community culture in the labeling process. Third, labelers

Table 3
The most frequent usage of emotes as substitutions for letters and their examples.

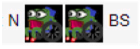
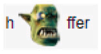
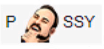
Rank	Top-1	Top-2	Top-3	Top-4	Top-5	Top-6
Substitution	o	a	e	u	i	s
Ratio	60.77%	17.94%	5.07%	3.93%	3.41%	1.03%
Example1						
Example2						
Example3						

Fig. 8. Conflicting labels by Twitch users, who translated this chat as 'DOTA' that is a popular multiplayer game, and by non-Twitch users, who translated the same chat as 'DATA.'

took disproportionately different amounts of time to complete the task; Twitch users took on average 25 min to label 200 chats, whereas non-users took on average 59 min. This finding suggests the different levels of burdens and skills required in the task.

4.3. Large-scale labeling through Amazon MTurk

We conducted larger-scale labeling with the Amazon Mechanical Turk (MTurk) platform following the pilot test insights. Given emotes are specific to Twitch, we wanted to decide how familiar participants need to be with the streaming platform to generate a meaningful label. We created an assignment (HIT) on the emote restoration task over 20 individuals and asked participants' familiarity with Twitch (i.e., "Are you a frequent user of Twitch?"). Then the participants were asked to restore 'emoted' text into all text words. When we compared the speed at which individuals could restore emoted words, it took 25 min to restore 200 emoted words for Twitch users, whereas it took 59 min to do so for participants not familiar with Twitch. Based on this pre-test, we decided to hire only users familiar with Twitch and paid participants 1 USD for restoring 50 emoted words, which on average took 6 min, yielding an hourly rate of 10 USD. There was no constraint on participants related to age, gender, race, or socioeconomic status. The HIT recruited English speaking participants who were familiar with Twitch.

We hired 3000 annotators who were each assigned 50 chat utterances to label, yielding a total of 150,000 decisions. Each chat utterance was given to at least three users to break a tie when needed. All 3000 users were familiar with Twitch, and they passed a screener that asked them a question about their knowledge of Twitch culture.[4] The task was to provide labels for an emote-annotated word; turkers could choose (1) an emote used to express emotion without any substitutions or (2) an emote is substituting specific letter(s). We have also asked them to make their best guesses on the original words in the latter case. Overall, the agreement of the turkers was relatively low. We could identify 11,485 cases of all chats where at least two turkers identified an emote as an identical substitution.

Table 3 shows the most frequently substituted letters from the MTurk labeling task. The most commonly replaced letters were vowels: `o`, `a`, `e`, `u`, and `i`, which are the top-5, then followed by `s`, `l`, `g`, `r`, and `n`. Table 4 displays the top emotes used to replace these alphabets, implying that any emotes can appear in substitution. Note that the substitution ratio is predominant for `o` and `a`, taking up 60.77% and 17.94% each, and decreases rapidly afterwards. We speculate on two reasons why vowels are more commonly replaced by emotes. One reason is the shape: emotes often have round boundaries, and many represent a face. This naturally fits vowels like `o` (e.g., see the Top-2 ranked emote for replacing `o` in Table 4). Another reason is the easiness of prediction. Substituted words need to be understood by other community members, and vowels have fewer branches than consonants, making them easier to guess. This preferred usage of vowel replacement is useful for building a classifier.

Table 4
Distribution of frequently substituted letters with the associated emotes.

Alphabet	Top-1	Top-2	Top-3	Top-4	Top-5
O	 26.33%	 10.21%	 5.73%	 4.09%	 3.84%
A	 46.98%	 4.77%	 3.88%	 2.92%	 2.03%
E	 35.68%	 6.81%	 5.40%	 3.52%	 2.58%
U	 46.65%	 4.29%	 2.14%	 2.14%	 1.61%
I	 29.11%	 8.23%	6.01%	3.16%	2.85%
S	26.32%	6.32%	4.21%	4.21%	2.11%

The MTurk task yielded 11,485 labels, which is not enough to build a neural network model, nor all of them are related to toxic chat. We hence augmented the data collection process by gathering an extensive collection of toxic word dictionaries and then synthetically replacing a single letter of any chosen word with a emote. In doing so, we may utilize the frequency distribution of the substituted letters and emotes appearing in our Twitch chat data. Five data sources were used on toxic chat: (1) a collection of swear words from a free online repository,[5] (2) a group of bad words from a CMU lab,[6] (3) the list of blocked words related to profanity by FrontGate,[7] (4) the list of Google black-listed words by Free Web Headers,[8] and (5) the list of blocked words suggested to YouTube moderators by Free Web Headers.[9] From the compiled collection of toxic words from these resources, we generated an additional 18,236 emoted-annotated toxic words that will be used for training a neural model. We plan to release this training data (i.e., labels of 29,721 emoted-words) and the codes for the neural model.

[4] We asked the meaning of the 'Kappa' emote, and only those who gave a correct answer (i.e., sarcasm) were invited as annotators.

[5] http://www.bannedwordlist.com/.
[6] https://www.cs.cmu.edu/biglou/resources/bad-words.txt.
[7] https://bit.ly/2lEbeyK.
[8] https://tinyurl.com/y2jk3tee.
[9] https://tinyurl.com/yxb4kmxg.

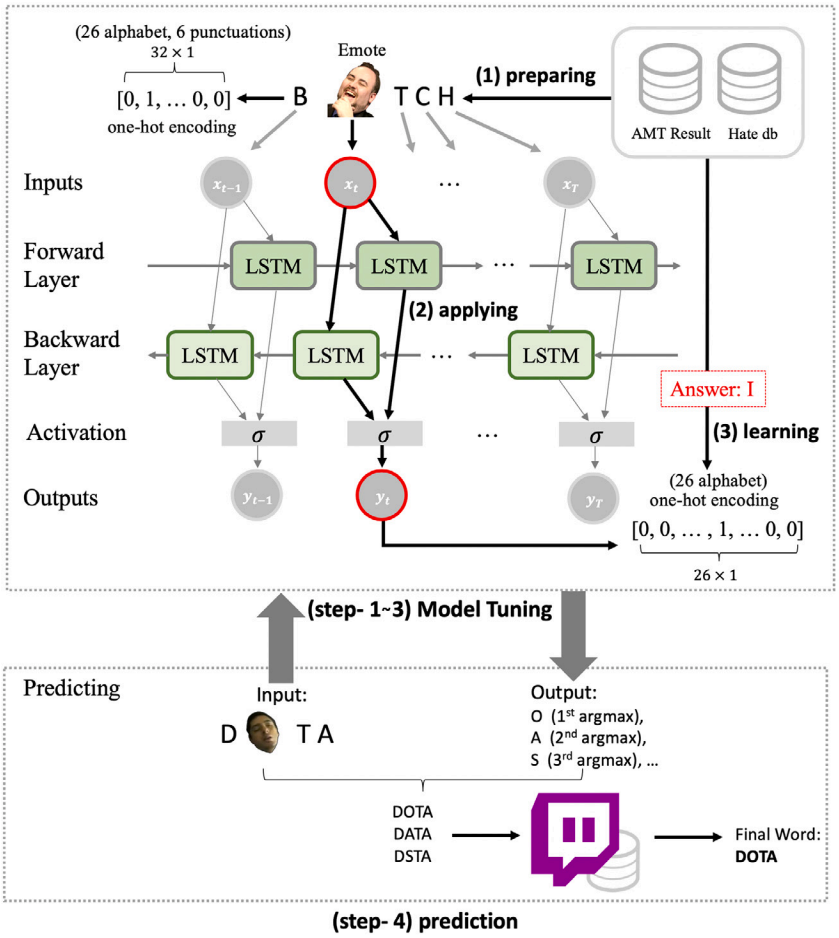


Fig. 9. Diagram for the bidirectional LSTM model.

4.4. Model description

Words are not a random mix of letters but have common forms that consist of a sequence of vowels and consonants that can be learned from massive data. This characteristic aligns well with the LSTM structure of the RNN-based model because the same model is used to learn from time-series data. We also adopt this neural network structure to take a sequence of letters as an input and predict the next letters to complete a word. For example, for the word `apple`, a network can guess what will follow `a-p-p-l` is likely a letter `e`. Slang and memes, which frequently occur in toxic chat, do not appear in the dictionary. The LSTM model that learns the relationship between words can handle these unordinary cases while the existing rule-based restoration is not. Besides, if sufficient data are collected in the future, the emotes' graphical feature will be used for the restoration model.

To ensure that we can treat emotes that appear next to a word rather than replacing a letter in it, we consider a bidirectional LSTM (Bi-LSTM) model that considers input in both forward and backward directions [56], in Fig. 9. Bi-LSTM model is the model that employs two LSTM layers: the forward layer and the backward layer. The main difference between these two layers is the direction of the layer. The forward layer aggregates the hidden states of each letter input from the start of the sentence, while the backward layer conducts the same process from the end of the sentence. In other words, the forward direction flow enables the model to preserve information from the past, whereas the backward direction flow enables to preserve information from the future. In Bi-LSTM, the joint consideration allows the model to consider information both from the past and future uniquely. This setting is essential when we want to distinguish letter replacement from both postfix or prefix. Because the aim is to restore the original word from "emote" words (i.e., part of the letter is substituted by a visual emote), we used the labeled dataset gathered from the AMT labeling in the model training process.

The considered NLP model operates in four steps. The first three steps describe the parameter tuning phase, with an example of the word `BITCH` where `I` is replaced with an emote. In Step-1, this input word is separated into letters, and they are encoded into one-hot vectors of 32 ×1 size, including 26 alphabets and six punctuation marks. Emotes are also encoded as a vector of the same size. In Step-2, each of the encoded vectors is used in the LSTM cells that scan the word in both forward and backward directions. The forward and backward connection in the BiLSTM structure helps predict what the emote's original letter would have been, as when it knows the position of emote in the word, through the letters used before and after. Formally, let us denote the input text as a list of vector $\{x_1, \ldots x_T\}$, whose element is a one-hot encoded vector representing every single letter.[10] The input vector is put into the BiLSTM layer, which is composed of two LSTM networks: forward $h_t^f = LSTM(x_t, h_{t-1}^f)$ and backward layer $h_t^b = LSTM(x_t, h_{t+1}^b)$. The outputs from forward and backward networks are concatenated as $h_t = [h_t^f, h_t^b]$. Finally, learning occurs in Step-3, where the output `I` that is missing is prepared as a 26 ×1 one-hot encoding vector. BiLSTM embedding h_t is forwarded to dense layer with softmax function for final prediction, and the model parameters are tuned via cross-entropy loss so that their response matches the outcome, the missing letter `I`.

[10] For emotes, we instead put zero vector **0** for masking.

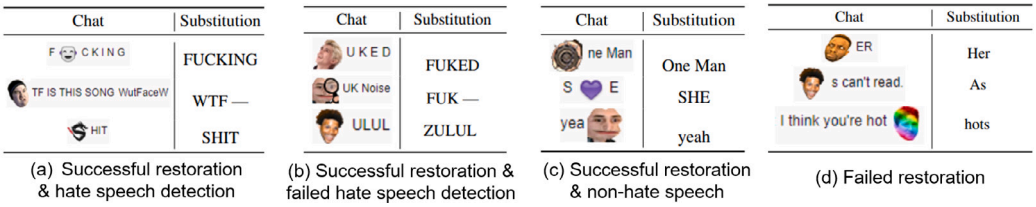


Fig. 10. Four types of detection examples.

Table 5
Toxic chat ratio difference after applying the model.

Total chats with emotes	Hatesonar flagged	LSTM model + Hatesonar flagged
15,290,530	694,970 (4.55%)	891,418 (5.83%)

Once the model is trained, it is now ready to be used for prediction in Step-4. What is unique about this phase is that we employ a community-specific corpus from Twitch chats to find the guessed output more relevant in its context. We have built a Twitch corpus from the 100 million chat conversations collected earlier, where each row contains a word with its rank order based on the frequency. By giving higher weight to words that appear more frequently on Twitch, we may avoid trivial predictions like `data` instead of the community-specific predictions like `dota` in case of the example in Fig. 8. The benefits of using community-specific corpus are two-fold. One is that the model will give a higher probability of restoring community-specific words like "DOTA" than general words. Another is that it could keep intentional grammar errors and slang usages that are common in chats. Twitch chat data contain many intentional misspelling or grammatical errors, and keeping this subtle nuance may be important in the restoration task. Hence, the comprehensive Twitch corpus enables the model to predict the original words accurately even if the words are grammatically incorrect (see Fig. 2).

4.5. Prediction result

The efficiency of the LSTM model was tested on real data. Out of the total 100 million data, only those chat utterances containing both emote and text were considered, resulting in 15,290,530 chats. We first applied Hatesonar to check what fraction of the data became flagged as toxic chat. Then, we restored emoted words by the LSTM model and applied Hatesonar to see how many new chats are now considered toxic. Table 5 displays the result. Overall, 694,970 instances or 4.55% of the emote-containing chats were flagged as hateful. In contrast, the LSTM model could identify an additional 196,448 chat instances that were newly flagged as hateful speech — an increase of 1.28% in detection.

Investigating further the replacement instances, we found four categories of outputs. These four kinds are illustrated with examples in Fig. 10.

- **Successful restoration & detection:** This is when the LSTM model successfully replaced the substituted letters, and Hatesonar flagged the original word as toxic chat. Examples of such kind include words like `WTF` and `SHIT`. These are the instances that motivated this research.
- **Successful restoration & failed detection:** Some of the instances were correctly translated by our LSTM model and judged as toxic to human eyes. However, they were not flagged by Hatesonar because of intentional misspelling, slangs, memes, abbreviated expressions, or play-on-words. Examples include `FUKED`, `FUK`, and `ZULUL`. The last example, in particular, is a racist word that is not widely used but would be recognized by minorities.[11] While Hatesonar is based on exact-matching of dictionary words, instances belonging to this category may be handled better by sophisticated toxic chat detectors.
- **Successful restoration & non-toxic chat:** The third category of words included successfully restored words, which do not include toxic chat. It is natural to find these cases because the mix of text and emotes is not restricted to toxic chat.
- **Failed restoration:** The last category is instances where the LSTM model seems to have failed in finding an appropriate replacement. The first example of `ER` could represent an angry astonishment instead of `HER`. In the second example, the Trihard face emote was replaced by the letter `A`, although the context seems to suggest that the emote is referring to `Black` — an entire word, instead of a letter. The final example also shows a likely misplacement of the emote. These cases depend on the context of chats, which is fundamentally challenging for machines to handle.

Regarding the model's practical use, we may consider how quickly toxic chat can be detected. Under a setting of Intel(R) Xeon(R) CPU E5-2620 v4 @ 2.10 GHz and 13 GB memory, the speed at which emoted-chats can be translated is 78.164 chats per second. This speed is faster than the rate at which typical chats occur on Twitch. In our 100 million-scale data, the quickest conversation peak was marked at 14 chats per second. This can be well covered by standard computing systems, although such high speed would make an average user challenging to comprehend the complete content. According to one research [11], Twitch conversations that are considered a "massive-yet-readable scale" mark around 3 to 4 chat utterances per second. These numbers indicate that the LSTM-based model can efficiently handle the real-time restoration of emoted chats.

The above analysis, however, does not indicate what fraction of the additionally flagged chats in Table 5 by the LSTM model are indeed toxic. It simply recalls that the candidate toxicity targets can be enlarged by the restoration method. To examine the actual rate of hate speech detected from this enlarged pool, we randomly chose 2000 likely hate comments that were newly discovered based on our model and tested their toxicity level with a third hate speech detection tool, the Google Perspective API. Perspective API is the product of a collaborative research effort by Jigsaw and Google's Counter Abuse Technology team [20]. This evaluation revealed that, out of 2000 newly identified hate speech based on our model, over 60% or 1203 chats were determined to contain some degree of hateful emotions, measured by Google Perspective.

To augment the analysis, we also consider the fact that Google Perspective may not be a gold standard. As an additional evaluation, we here resorted to manual labeling. This manual labeling took place by the first author randomly selecting 400 chats containing both texts and emotes from Twitch corpus and labeling whether they contain hate speech. Any ambiguous cases were discussed with the other authors and resolved. Note that the subset is specifically targeting the emoted conversation. Nonetheless, Hatesonar may still detect hate speech if independent words would contain any of the hate words. Our restoration

[11] https://knowyourmeme.com/memes/zulul

Table 6
Performance comparison after applying restoration model based on a manually labeled set of 400 chats that utilizes combined text and mote words.

	Accuracy	Pecision	Recall	f1-score
Baseline (Hatesonar)	0.693	0.86	0.644	0.737
Restoration + Hatesonar	0.703	0.87	0.652	0.745

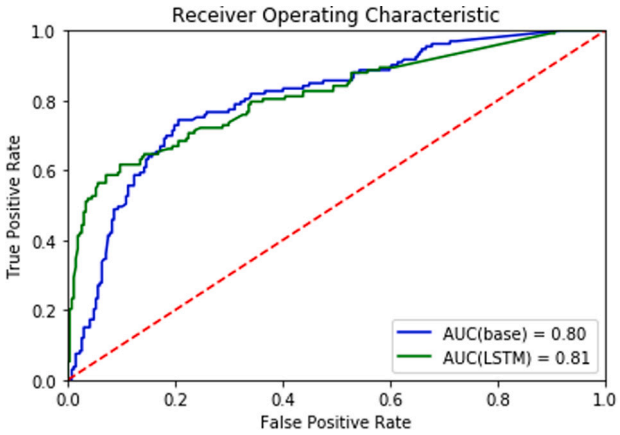


Fig. 11. ROC curve comparison of two models based manually labeled dataset.

model would be beneficial if the hate speech can be detected from the "emoted" text.

Table 6 shows the results in terms of accuracy, precision, recall, and F1-score based on the manually labeled dataset. Fig. 11 shows the AUROC curve based on the same comparison, which is a graph showing the performance of a classification model at all classification thresholds. Although the increase in performance gain is not substantial in both the table and the figure, the restoration model could detect a new type of toxic conversation. Examples of hate speech that our LSTM model could newly identify are given in Fig. 10.

5. Discussion

5.1. Implications

One of the surprising findings from this study is that the kinds of emotes frequently used in toxic chat are not fundamentally different from the list of popular emotes, as demonstrated in Table 2. This finding implies that an emote's meaning may change dramatically, depending on the context it is used in, i.e., a smile can turn a smirk by its accompanying text. This finding also indicates that while several researchers have proposed using emojis to infer sentiments' valence, the same kind of rigid definition should not apply to understand Twitch conversations.

In terms of the types of toxic chat, we observed a pattern of collectivism in Study 1 [57]; people utilize famous emotes depicting faces (e.g., Black streamer, Kappa with a rainbow background) to refer to minority groups in toxic chat. Increasing the use of these popular emotes in a negative context leads to unintended consequences. For instance, exposure to casual racism denoted by the Trihard and CmonBruh emotes might make viewers less sensitive to reporting such toxic chats and may further associate negative connotations to those emotes.

In Study 2, the proposed classifier could newly identify 1.28% additional instances. While this ratio is small, they account for massive amounts of hateful expressions detected in real-time — 196,448 new cases among the 15 million chat utterances. Had Twitch channels hired human moderators to identify these, it would have taken a substantial amount of time and effort. For example, it would take more than 200,000 h to examine 100 million chat utterances. Techniques to efficiently handle visual toxic chats and suggest labels for human administrators to confirm would reduce the real-time moderation burden.

Nevertheless, not all emote translations were successful. The failed detection and failed restoration examples listed in Fig. 10 indicate that machines alone will not be able to detect visual toxic chat completely; certain cases require a deep understanding of the online community and sophisticated handling of broken language that humans so far excel in recognizing than machines.

The technique to translate emote-substituted words is helpful for toxic chat detection and channel moderators who want to grasp the meaning quickly to monitor live conversations. Given the growing trend of adding visual cues in text-based chats, this technique can be applied to other kinds of user-generated data, such as social media streams and news comments.

The internal group study had identified several main challenges. While not explained above, we encountered various kinds of failed restoration cases. One of them was due to the use of third-party emotes, which could not be appropriately displayed. Therefore, this study was limited to those emotes that could be displayed in the default Twitch platform. Another challenge arose when multiple translations were possible. While our work utilized the Twitch corpus to give higher weights to community-specific expressions, the emoted chat's true meaning is left open for dispute. Examining user-level characteristics or sentence-level translations (instead of word-level translations) might provide more profound insight into this problem.

The labeling preparation step also revealed that familiarity with the Twitch platform is critical to infer emote-substituted chats' correct meaning. Those labelers who did not know the Twitch culture took a long time to complete the task, and their labels were not as accurate as those familiar with the system. Therefore, we ensured that all participants in the Mechanical Turk were well-aware of popular emotes before participating in the labeling task. Systems that suffer from toxic content may utilize this finding in hiring high-quality labelers who recognize community-specific language — one may envision a gamified reputation system that encourages advanced-level users to label or report toxic behaviors.

One of the interesting facts that this study has shown is that emote substitution is not only a phenomenon of toxic chat. The proposed restore module can be effetely used for various tasks conducted on social media, especially for those with diverse writing style, e.g., rumor detection [58,59], stance detection [60,61], and opinion summarization [62], etc. On Twitch, emotes are the most popular non-verbal communication. On other platforms, different kinds of emoticons, emoji, and GIFs are used.

Our key contribution is not Twitch-specific; our hate speech detection approach through reverse-engineering of words containing pictures is a unique method that could be used on other platforms since text substitution is a growing trend across various Internet communities. Even if the specific algorithm that we used may not work well on other platforms, our general approach is not platform-specific. Building a customized label dataset was critical to our research. The labelers who had little understanding of Twitch were substantially less accurate in their coding process and took a long time. Thus, researchers who would want to use this approach in other social media contexts should also be sensitive about that specific community's norms and culture.

Finally, in terms of design implications, we may envision a live chat interface that visualizes the level of toxic chat. Since streamers and moderators decide what kinds of behavior are permitted, the level of toxicity may vary from one chat room to another. Some chat rooms may allow some degree of toxicity, while others may be less susceptible to toxicity. The technique to restore and translate emoted words can help measure toxicity levels more accurately and contribute to such a system.

5.2. Limitation and future work

The idea to predict a missing letter in a word is simple yet powerful. This method could detect new toxic chat instances with high time efficiency. We could process hate words from public sources, including direct crowd-sourced data, and use them as learning data.

Current toxic chat detection techniques face two challenges. First, the nature of "offense" is subjective in that individuals have different levels of susceptibility [63]. Sap [64] noted that biases in labeling might occur due to such a difference in user perception. Second, supervised learning approaches require a massive amount of labels that are costly to generate. We used the labels from crowd-sourcing and augmented the dataset with common hate words to solve this problem. The compiled collection of emoted chats and their labels with original words will be shared for research purposes.

Our data indicated that adding an emote is common, accounting for one-third of all chats. Emotes make conversations epic and exciting; it is a light way of expressing emotion without words in the fast-paced chat platform. However, it is also notable that some hateful comments also accompany emotes. Given that emojis play similar functions, this research's findings could apply to other social media platforms. We also expect that the process of restoring the replacement of letters and the replacement of words can be carried out in the future (see discussion of failed restoration cases in Fig. 10). Even if a particular emote represents race or social group, it can be found that the group is degraded or mocked.

Future studies may explore a classifier that detects emotes representing a word (rather than a letter) via the use of sophisticated language models like BERT [65] and XLNet [66]. In such a case, a flexible embedding of each emote can be generated while utilizing the understanding of language which is pre-trained in the existing model parameters. Such "fine-tuning" may be helpful, compared to starting from a model with zero knowledge. Fine-tuning in deep learning involves using weights of a previous deep learning algorithm for programming another similar deep learning process. NLP models that are pre-trained on other online short text data, such as the Twitter corpus, may be fine-tuned and used for the context of Twitch. Another exciting direction would be to visually treat all chat input and apply CNN-based filters to interpret emotes that mimic letters. One may extend the study as in [35], which examined emojis as visual aids to text.

When it comes to data collection, we limited our observations to live chat on top-100 streamers, given they produce a large number of chat messages. It would be an important future direction to detect toxicity in chats in streams of marginalized creators who sometimes become targets of hate raids. One may test hate detection models' efficacy and understand the possible bias due to localized language usage.

6. Conclusion

This research analyzed toxic chats on the online streaming platform Twitch and examined emerging visual toxic chats that involve a combination of text and emotes. Such moderations are hard to detect using conventional text-only hate speech classifiers. We found that the top emotes are a similar set, for both random chats and toxic chats. We found six common use cases of emotes in toxic chat: (1) emotion replaces the entire word, (2) emotion replaces a specific letter, (3) Twitch or community-specific meaning, (4) attention-seeking (an act of attracting attention by putting an emotes in the space between words), (5) visual enhancement of text (pictorial representations of writing), and (6) emotes as pictures of words or emotional expression.

Among the various use cases, we targeted restoring words where an emote replaced a single letter. We employed crowdsourcing to obtain many labels describing potential answers for which letter would have been replaced by an emote. We also augmented this labeled dataset by utilizing other hate-related public resources. Together, 29,721 emoted words were collected, which was used as a training dataset for the model. The classifier was based on a bidirectional LSTM structure, efficient in handling sequence data such as words. In the end, our neural classifier could flag 196,448 (1.28%) new hateful chat messages against the existing Hatesonar library out of 15 million data.

Hateful expressions are expanding rapidly, and Internet users continuously create new forms of toxic expression that can bypass algorithmic moderation. This research examined one emerging type where emotes are replacing text in a word. The bidirectional LSTM structure is intuitive and can effectively identify tactics like visual toxic chat. This method is flexible in handling new combinations of text and emotes, and it also applies to domains outside Twitch to treat emojis, memes, and emoticons. While efficient, we do not argue that this proposed model act on behalf of human moderators. Instead, we acknowledge there can be many other creative variations of emote usage in hate speech that future studies can tackle and better estimate the value-added of the model in detecting toxicity. Our research also indicates that understanding the subtle context is critical in detecting toxic chat, and models presented in this study can best be used as assisting tools.

Despite the positive detection result, we recognize the systematic challenge in building language models in that many Internet users are adopting non-standard writing styles. Some systems intentionally modify the spelling, as well-known in leet (or leetspeak), a modified spelling system primarily on the Internet. A popular example is `n00b` to represent newbies. Certain toxic chats that mix leet with emotes would be challenging to detect unless one knows the language is written. Given the gradual popularity of systems like this, future NLP techniques would need to be more comprehensive in analyzing social media posts aided with encryption and slang.

CRediT authorship contribution statement

Jaeheon Kim: Collected the data, Analyzed the data, Writing – original draft. **Donghee Yvette Wohn:** Conceived the research, Analyzed the data, Writing – original draft. **Meeyoung Cha:** Conceived the research, Analyzed the data, Writing – original draft.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the Institute for Basic Science (IBS-R029-C2) and the National Research Foundation, South Korea (No. NRF-2017R1E1A1A01076400) funded by the Ministry of Science and ICT in Korea. We thank Lucy Feng for help in the labeling task, and this work was done while the first author was at KAIST. All authors approved the final version of the manuscript.

References

[1] Frank Bentley, Danielle Lottridge, Understanding mass-market mobile TV behaviors in the streaming era, in: Proc. of the ACM CHI, 2019.
[2] William A. Hamilton, Oliver Garretson, Andruid Kerne, Streaming on Twitch: fostering participatory communities of play within live mixed media, in: Proc. of the ACM CHI, 2014.
[3] Distribution of Twitch users worldwide as of 2nd quarter 2019, by age group, 2019, Available at https://www.statista.com/statistics/634057/twitch-user-age-worldwide/.
[4] Donghee Yvette Wohn, Guo Freeman, Caitlin McLaughlin, Explaining viewers' emotional, instrumental, and financial support provision for live streamers, in: Proc. of the ACM CHI, 2018.
[5] Max Sjöblom, Juho Hamari, Why do people watch others play video games? An empirical study on the motivations of Twitch users, Comput. Hum. Behav. 75 (2017) 985–996.

[6] Ryan D. King, Gretchen M. Sutton, High times for hate crimes: Explaining the temporal clustering of hate-motivated offending, Criminology 51 (4) (2013) 871–894.

[7] Clay Calvert, Hate speech and its harms: A communication theory perspective, J. Commun. 47 (1) (1997) 4–19.

[8] Susan Villani, Impact of media on children and adolescents: A 10-year review of the research, J. Am. Acad. Child Adolesc. Psychiatry 40 (4) (2001) 392–401.

[9] Joseph Seering, Robert Kraut, Laura Dabbish, Shaping pro and anti-social behavior on Twitch through moderation and example-setting, in: Proc. of the ACM CSCW, 2017.

[10] Donghee Yvette Wohn, volunteer moderators in Twitch micro communities: How they get involved, the roles they play, and the emotional labor they experience, in: Proc. of the ACM CHI, 2019.

[11] Colin Ford, Dan Gardner, Leah Elaine Horgan, Calvin Liu, Bonnie Nardi, Jordan Rickman, et al., Chat speed op pogchamp: Practices of coherence in massive Twitch chat, in: Proc. of the ACM CHI Extended Abstracts, 2017.

[12] William Warner, Julia Hirschberg, Detecting hate speech on the world wide web, in: Proceedings of the Second Workshop on Language in Social Media, Association for Computational Linguistics, 2012, pp. 19–26.

[13] Thomas Davidson, Dana Warmsley, Michael Macy, Ingmar Weber, Automated hate speech detection and the problem of offensive language, in: Proc. of the ICWSM, 2017.

[14] Lewis L. Chuang, Ulrike Pfeil, Transparency and openness promotion guidelines for HCI, in: Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems, 2018, pp. 1–4.

[15] Matthew Kay, Steve Haroz, Shion Guha, Pierre Dragicevic, Chat Wacharamanotham, Moving transparent statistics forward at CHI, in: Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems, 2017, pp. 534–541.

[16] Julia Alexander, Abuse of KFC emote on Twitch leads to more conversations about toxic chat culture, 2018, Available at https://www.polygon.com/2018/3/26/17163582/kfc-emote-twitch-trihex-forsen-trihard-xqc.

[17] Karine Pires, Gwendal Simon, YouTube live and Twitch: a tour of user-generated live streaming systems, in: Proceedings of the 6th ACM Multimedia Systems Conference, ACM, 2015, pp. 225–230.

[18] 25 Useful Twitch statistics for influencer marketing managers, 2018, Available at https://influencermarketinghub.com/25-useful-twitch-statistics/.

[19] Hendrik Storstein Spilker, Kristine Ask, Martin Hansen, The new practices and infrastructures of participation: how the popularity of Twitch. tv challenges old and new ideas about television viewing, Inf. Commun. Soc. (2018) 1–16.

[20] Google perspective API, 2019, Available at http://www.perspectiveapi.com/.

[21] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, Yi Chang, Abusive language detection in online user content, in: Proceedings of the 25th International Conference on World Wide Web, International World Wide Web Conferences Steering Committee, 2016, pp. 145–153.

[22] J. Nathan Matias, Preventing harassment and increasing group participation through social norms in 2,190 online science discussions, Proc. Natl. Acad. Sci. USA 116 (20) (2019) 9785–9789.

[23] Kunal Relia, Zhengyi Li, Stephanie H. Cook, Rumi Chunara, Race, ethnicity and national origin-based discrimination in social media and hate crimes across 100 US cities, in: Proceedings of the International AAAI Conference on Web and Social Media, Vol. 13, 2019, pp. 417–427.

[24] Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, Eric Gilbert, The internet's hidden rules: An empirical study of reddit norm violations at micro, meso, and macro scales, in: Proc. of the ACM CSCW, 2018.

[25] Shan Jiang, Ronald E. Robertson, Christo Wilson, Bias misperceived: The role of partisanship and misinformation in YouTube comment moderation, in: Proceedings of the International AAAI Conference on Web and Social Media, Vol. 13, 2019, pp. 278–289.

[26] James Banks, Regulating hate speech online, Int. Rev. Law Comput. Technol. 24 (3) (2010) 233–239.

[27] Homa Hosseinmardi, Sabrina Arredondo Mattson, Rahat Ibn Rafiq, Richard Han, Qin Lv, Shivakant Mishra, Detection of cyberbullying incidents on the instagram social network, 2015, arXiv preprint arXiv:1503.03909.

[28] Joseph Seering, Michal Luria, Geoff Kaufman, Jessica Hammer, Beyond dyadic interactions: Considering chatbots as community members, in: Proc. of the ACM CHI, 2019.

[29] Azadeh Nematzadeh, Giovanni Luca Ciampaglia, Yong-Yeol Ahn, Alessandro Flammini, Information overload in group communication: From conversation to cacophony in the Twitch chat, 2016, arXiv preprint arXiv:1610.06497.

[30] J. Park, V. Barash, C. Fink, M. Cha, Emoticon style: Interpreting differences in emoticons across cultures, in: Proc. of the ICWSM, 2013.

[31] Nikola Ljubešić, Darja Fišer, A global analysis of emoji usage, in: Proceedings of the 10th Web As Corpus Workshop, 2016, pp. 82–89.

[32] Tim Highfield, Emoji hashtags//hashtag emoji: Of platforms, visual affect, and discursive flexibility, First Monday 23 (9) (2018).

[33] Hannah Jean Miller, Jacob Thebault-Spieker, Shuo Chang, Isaac Johnson, Loren Terveen, Brent Hecht, "Blissfully happy" or "ready to fight": varying interpretations of emoji, in: Proc. of the ICWSM, 2016.

[34] Jayashree Subramanian, Varun Sridharan, Kai Shu, Huan Liu, Exploiting emojis for sarcasm detection, in: Social, Cultural, and Behavioral Modeling, 2019, pp. 70–80.

[35] B Eisner, T Rocktäschel, I Augenstein, M Bošnjak, S Riedel, emoji2vec: Learning Emoji Representations from their Description, in: Proc. of the ACL, 2016.

[36] Francesco Barbieri, Luis Espinosa Anke, Miguel Ballesteros, Juan Soler, Horacio Saggion, Towards the understanding of gaming audiences by modeling Twitch emotes, in: Proc. of the Workshop on Noisy User-Generated Text, 2017.

[37] Bjarke Felbo, Alan Mislove, Anders Sø gaard, Iyad Rahwan, Sune Lehmann, Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm, in: Proc. of the ACL, 2017.

[38] Jedrzej Olejniczak, A linguistic study of language variety used on Twitch. tv: desriptive and corpus-based approaches, Redefining Community in Intercult. Context (2015) 329–344.

[39] Emily M. Bender, Batya Friedman, Data statements for natural language processing: Toward mitigating system bias and enabling better science, Trans. Assoc. Comput. Linguist. 6 (2018) 587–604.

[40] Anna Schmidt, Michael Wiegand, A survey on hate speech detection using natural language processing, in: Proc. of the International Workshop on Natural Language Processing for Social Media, SocialNLP, 2017.

[41] Ying Chen, Yilu Zhou, Sencun Zhu, Heng Xu, Detecting offensive language in social media to protect adolescent online safety, in: Proc. of the International Conference on Privacy, Security, Risk and Trust, 2012.

[42] Quoc Le, Tomas Mikolov, Distributed representations of sentences and documents, in: Proc. of the ICML, 2014.

[43] Björn Gambäck, Utpal Kumar Sikdar, Using convolutional neural networks to classify hate-speech, in: Proceedings of the First Workshop on Abusive Language Online, 2017, pp. 85–90.

[44] Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean, Efficient estimation of word representations in vector space, 2013, arXiv preprint arXiv:1301.3781.

[45] Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, Narayan Bhamidipati, Hate speech detection with comment embeddings, in: Proceedings of the 24th International Conference on World Wide Web, 2015, pp. 29–30.

[46] Yashar Mehdad, Joel Tetreault, Do characters abuse more than words? in: Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 2016, pp. 299–303.

[47] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, Vasudeva Varma, Deep learning for hate speech detection in tweets, in: Proceedings of the 26th International Conference on World Wide Web Companion, International World Wide Web Conferences Steering Committee, 2017, pp. 759–760.

[48] Amr Mousa, Björn Schuller, Contextual bidirectional long short-term memory recurrent neural network language models: A generative approach to sentiment analysis, in: Proc. of the European Chapter of the Association for Computational Linguistics, EACL, 2017.

[49] Karthik Dinakar, Birago Jones, Catherine Havasi, Henry Lieberman, Rosalind Picard, Common sense reasoning for detection, prevention, and mitigation of cyberbullying, ACM Trans. Interact. Intell. Syst. 2 (3) (2012) 18.

[50] Marzieh Mozafari, Reza Farahbakhsh, Noel Crespi, A BERT-based transfer learning approach for hate speech detection in online social media, 2019, arXiv preprint arXiv:1910.12574.

[51] Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, Ophir Frieder, Hate speech detection: Challenges and solutions, PLoS One 14 (8) (2019).

[52] Mitchell L. Gordon, Kaitlyn Zhou, Kayur Patel, Tatsunori Hashimoto, Michael S. Bernstein, The disagreement deconvolution: Bringing machine learning performance metrics in line with reality, in: Conference on Human Factors in Computing Systems, 2021, pp. 388:1–388:14.

[53] Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, William Yang Wang, A benchmark dataset for learning to intervene in online hate speech, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP, 2019, pp. 4757–4766.

[54] Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, Dit-Yan Yeung, Multilingual and multi-aspect hate speech analysis, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP, 2019, pp. 4667–4676.

[55] William Fedus, Ian Goodfellow, Andrew M. Dai, MaskGAN: Better text generation via filling in the _____, in: Proc. of the ICLR, 2018.

[56] Alex Graves, Jürgen Schmidhuber, Framewise phoneme classification with bidirectional LSTM and other neural network architectures, Neural Netw. 18 (5–6) (2005) 602–610.

[57] Harry C. Triandis, Individualism and Collectivism, Routledge, 2018.

[58] Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J Jansen, Kam-Fai Wong, Meeyoung Cha, Detecting rumors from microblogs with recurrent neural networks, in: Proceedings of the International Joint Conferences on Artificial Intelligence, IJCAI, 2016, pp. 3818–3824.

[59] Jing Ma, Wei Gao, Kam-Fai Wong, Detect rumors in microblog posts using propagation structure via kernel learning, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2017, pp. 708–717.

[60] Michal Lukasik, Trevor Cohn, Kalina Bontcheva, Classifying tweet level judgements of rumours in social media, 2015, arXiv preprint arXiv:1506.00468.

[61] Michal Lukasik, PK Srijith, Duy Vu, Kalina Bontcheva, Arkaitz Zubiaga, Trevor Cohn, Hawkes processes for continuous time sequence classification: an application to rumour stance classification in twitter, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Vol. 2, 2016, pp. 393–398.

[62] Xinfan Meng, Furu Wei, Xiaohua Liu, Ming Zhou, Sujian Li, Houfeng Wang, Entity-centric topic-oriented opinion summarization in twitter, in: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2012, pp. 379–387.

[63] Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, Ruihong Huang, Sarcasm as contrast between a positive sentiment and negative situation, in: Proc. of the EMNLP, 2013.

[64] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, Noah A Smith, The risk of racial bias in hate speech detection, in: Proc. of the ACL, 2019.

[65] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2018, arXiv preprint arXiv:1810.04805.

[66] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, Quoc V Le, XLNet: Generalized autoregressive pretraining for language understanding, 2019, arXiv preprint arXiv:1906.08237.