Proceedings of the Eighteenth International AAAI Conference on Web and Social Media (ICWSM 2024)

Games and Beyond: Analyzing the Bullet Chats of Esports Livestreaming

Yukun Jiang, Xinyue Shen, Rui Wen, Zeyang Sha, Junjie Chu, Yugeng Liu,
Michael Backes, Yang Zhang

CISPA Helmholtz Center for Information Security

Abstract

Esports, short for electronic sports, is a form of competition using video games and has attracted more than 530 million audiences worldwide. To watch esports, people utilize online livestreaming platforms. Recently, a novel interaction method, namely "bullet chats," has been introduced on these platforms. Different from conventional comments, bullet chats are scrolling comments posted by audiences that are synchronized to the livestreaming timeline, enabling audiences to share and communicate their immediate perspectives. The real-time nature of bullet chats, therefore, brings a new perspective to esports analysis. In this paper, we conduct the first empirical study on the bullet chats for esports, focusing on one of the most popular video games, i.e., League of Legends (LoL). Specifically, we collect 21 million bullet chats of LoL from Jan. 2023 to Mar. 2023 across two mainstream platforms (Bilibili and Huya). By performing quantitative analysis, we reveal how the quantity and toxicity of bullet chats distribute (and change) w.r.t. three aspects, i.e., the season, the team, and the match. Our findings show that teams with higher rankings tend to attract a greater quantity of bullet chats, and these chats are often characterized by a higher degree of toxicity. We then utilize topic modeling to identify topics among bullet chats. Interestingly, we find that a considerable portion of topics (14.14% on Bilibili and 22.94% on Huya) discuss themes beyond the game, including genders, entertainment stars, non-esports athletes, and so on. Besides, by further modeling topics on toxic bullet chats, we find hateful speech targeting different social groups, ranging from professions, regions, etc. To the best of our knowledge, this work is the first measurement of bullet chats on esports livestreaming. We believe our study can shed light on esports research from the perspective of bullet chats.

1 Introduction

Electronic sports, known as *esports*, is a form of competition using video games and has gained increasing popularity with the emergence of streaming media (Hamari and Sjöblom 2017; Telefonica 2023). As of 2022, there are approximately 3.09B esports players and more than 530M esports audiences globally (Howarth 2023; Gough 2023b). Given such a large user base, recent years have witnessed an increased interest in the field of esports research.

Figure 1: An illustration for bullet chats of League of Legends livestreaming.

To watch esports, one major means is through online livestreaming platforms, such as Twitch[1] in Western as well as Bilibili[2] and Huya[3] in China (Gough 2023a). Different from the Western world, the audiences on Chinese livestream platforms have a specific method, namely "bullet chats," to enable audiences to interact with each other. As illustrated in Figure 1, bullet chats consist of scrolling comments posted by audiences that appear on the screen for others to see during the livestreaming. This feature provides a real-time social interaction for audiences to share their immediate perspectives and engage in live communication, which also introduces a novel perspective for esports analysis.[4] By 2023, bullet chats have been integrated into the majority of Chinese esports livestreaming, resulting in more than 10 billion bullet chats. However, to the best of our knowledge, there has been no comprehensive analysis of esports from the perspective of bullet chats.

Previous studies on esports primarily focus on game players via analyzing their social structures, emotions, and behaviors (Kou and Gui 2014, 2020). For instance, (Aguerri, Santistebanand, and Miró-Llinares 2023) find that 70% of matches (a round of the game) in an online video game are affected by toxic behaviors, like quitting the match midway,

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

[2]https://www.bilibili.com/.

[3]https://www.huya.com/.

[4]https://www.goldthread2.com/culture/crazy-way-people-watch-videos-china-whole-subculture-its-own/article/2157123.

hate speech, etc. Such behaviors not only contravene the terms of service for the game[5] but also detrimentally affect the overall user experience. Besides, esports-related discussions in traditional social media, like Facebook, Reddit, and YouTube comment sections, have also been studied (Torres-Toukoumidis 2022; Lambert, Rajagopal, and Chandrasekharan 2022). In this work, we aim to make the first attempt at analyzing esports from the perspective of bullet chats. More precisely, we focus on three research questions given below:

RQ1 How do the season, teams, and matches affect the bullet chats?

RQ2 Can current toxicity detectors identify toxic bullet chats? How toxic are the bullet chats, and how do they change over time?

RQ3 What (toxic) themes do the bullet chats include?

Methodology. In this paper, we set out to study bullet chats in esports with a focus on League of Legends (LoL) in China, where LoL is one of the most popular esports games in the world, and China has the largest esports market and most active bullet chat users (Kou and Gui 2020; Gough 2022, 2023a; He et al. 2021). Concretely, during an entire game season of LoL, we collect 12,902,940 and 8,105,173 bullet chats from the official LoL livestreaming rooms on two mainstream platforms, Bilibili and Huya, respectively.

To address **RQ1**, we utilize the Frequency of Bullet Chats (FBC) as the metric to measure how frequently users send bullet chats within a specific time interval. By varying the time interval, we explore how FBC changes with respect to three aspects, i.e., the season, team, and match (Section 4). For **RQ2**, we first evaluate existing methods and our fine-tuned model for detecting toxic bullet chats of esports in Chinese, then we use the best-performed one (i.e., our fine-tuned model) to measure the toxicity of the aforementioned three aspects (Section 5). Regarding **RQ3**, we build topic models on bullet chats to explore what exact (toxic) themes are contained in each platform (Section 6).

Main Findings. Overall, we make the following findings:

- In both platforms, teams with higher rankings are prone to obtain a greater quantity of bullet chats. Besides, the beginning and end of a match commonly attract more bullet chats. This indicates esports audiences across different platforms share similar interests while watching games (RQ1).
- Traditional methods face significant challenges in identifying toxic bullet chats, with an F1-Score below 0.1. Hence, we fine-tune a pre-trained model for our needs and achieve an F1-Score of 0.66. This highlights the difficulty in addressing toxic content in this unique online social context (RQ2). We are committed to sharing our toxicity detection model for future research.[6]
- Bullet chats on Huya are 77% more toxic than those on Bilibili, highlighting a significant disparity in the interaction patterns between users on the two platforms (RQ2).
- Bullet chats not only revolve around in-game topics like players and teams but also touch upon societal topics such as gender, entertainment figures, and even issues affecting specific social groups. Furthermore, various target groups are observed in toxic bullet chats, including gender, professions, regions, and so on (RQ3).

Contributions. Our contributions are threefold. First, we begin to study esports (an industry serving billions of people) from the perspective of livestreaming bullet chats, a novel as well as real-time social interaction. Then, we conduct quantitative and longitudinal measurements on millions of real-world bullet chats across three aspects, i.e., the season, teams, and matches, to understand how the (toxic) bullet chats are distributed and related to esports. We find that the quantity and toxicity of bullet chats are related to teams, which is suggestive for advertisers to place ads. Finally, we systematically investigate the themes of bullet chats. We find that the discussion scope of bullet chats has gone far beyond esports and commonly comprises sexism, hate speech, stereotypes, etc. Given the extensive user base and the instantaneous nature of bullet chats, toxic bullet chats can be quickly viewed by a substantial audience within a short timeframe, facilitating the spread of toxic rhetoric on social media. This work could serve as a basis for understanding the ecosystem of bullet chats in esports livestreaming and a guide for the industry to consider better mechanisms for bullet chat governance.

Ethical Considerations. We emphasize that we rely entirely on publicly available data shared on video platforms and do not collect user information. Meanwhile, all the manual annotations are performed by the authors of this work. Hence, our work is not considered human subject research by the Ethical Board Committee.

Disclaimer. The bullet chat content can be considered unsafe. We include them in the paper to better illustrate the peculiarities of this problem. Reader discretion is advised.

2 Preliminaries

Esports. Esports, short for electronic sports, usually takes an organized form of multiplayer video game competitions, especially multiplayer online battle arena (MOBA) games such as Defense of the Ancients (DotA) and League of Legends (LoL). The origins of esports can be traced back to competitive in-person arcade video game contests that took place during the 1970s, and it has grown rapidly in recent years with billions of participants (Howarth 2023; Gough 2023b). By redefining sports, gaming, and entertainment, fostering community, and promoting well-being, esports has significantly influenced society and culture(Torres-Toukoumidis 2022).

League of Legends (LoL). League of Legends (LoL) is a MOBA game published by Riot Games[7] in 2009. In each match, two teams of five players fight against each other, and the condition for victory is to destroy the opponent's base.

[5] https://www.riotgames.com/en/terms-of-service.

[6] https://github.com/Ashbringer0926/Toxicity-Detection.

[7] https://www.riotgames.com.

LoL has become one of the most popular esports games, with around 180 million active players and millions of viewers (Kamberovic 2023; Li 2017). For instance, during the 2019 LoL World Championship finals, over 100 million viewers were drawn to the event (Webb 2019). Additionally, in October 2021, there were a recorded 208 million hours of LoL game viewing on Twitch.[8]

There are, in total, 9 professional leagues for LoL around the world. The LPL, shorted for League of Legends Pro League, is the top-level professional league owned by Tencent for League of Legends in China, established in 2013. Among all regions, China has the most (around 75 million) LoL players, as well as approximately 100 million LoL audiences (Kamberovic 2023; Wikipedia 2023a). LPL has two splits each year: the Spring Split and the Summer Split. Each split has a regular season where all teams perform single round robin and playoffs for the top eight teams in the regular season. In this paper, we focus on the regular season to cover each team completely. During the regular season, there are 17 teams (Figure 4's y-axis concludes each team's name and ranking), 136 BO3 (the acronym for best of 3, where the winner must win two out of three matches to win the contest), and 328 separate matches. During each match, there are three noticeable time points as follows:

- **BP (Ban and Pick) start time.** At the very beginning of each match, both teams need to ban and pick (BP) the characters (legends) used in the match. We record the BP start time as the origin of each match time.
- **Fight start time.** After the BP, players from both sides will officially enter the fight. We record the time they enter the fight as the fight start time.
- **Match end time.** When one team defeats the other (i.e., destroys the opponent's base), the match ends. We record this time as the match end time.

Bullet Chatting. Bullet chatting, also known as danmu, is a comment system and even a social mechanism, allowing users to post moving comments, namely bullet chats, onto a playing video that is synchronized to the video timeline (Huang et al. 2023; Wikipedia 2023c). As shown in Figure 1, bullet chatting allows audiences to post their comments on the screen during livestreaming. Compared to traditional comment systems like YouTube comment section, bullet chatting could provide a better experience for users to interact in real-time (Huang et al. 2023). Thus, the majority of video platforms in China have implemented bullet chatting as a vital way of commenting (Zhang and Cassany 2020; Mei 2021; Huang et al. 2023), and bullet chats have become a new method for online users to express their feelings in a more timely, straightforward, and dense manner. Due to the popularity and real-time nature of bullet chats, through esports livestreaming, numerous bullet chats are posted by audiences (see Table 1 as an example).

3 Data Collection

To understand the ecosystem of bullet chats in LoL, we first collect bullet chats during the regular season of LPL 2023 from Bilibili and Huya, two most popular livestreaming platforms in China with the (exclusive) livestreaming rights of LPL 2023.[9] Table 1 summarizes our collected datasets.

Bilibili. As one of the major streaming platforms in China, Bilibili provides various video services such as livestreaming and movie clips, and it has achieved more than 44 billion e-sports video views.[10] We choose to include Bilibili as it is famous for introducing bullet chatting as a form of commenting on its videos and it is an official platform for LPL livestreaming. Thus, it is likely to attract a considerable amount of bullet chat users. For each match day, we begin to crawl bullet chats before the BP starts and stop crawling after the end of all matches on this day. We do not actively request any additional web server content and thus will not cause additional load to the platform. Besides, for each match, the authors are online to record the accurate BP start, fight start, and match end times to filter out bullet chats outside matches. In total, we collect 12,902,940 bullet chats from January 14, 2023, to March 26, 2023, across the entire game season. We keep 6,931,433 filtered bullet chats in our Bilibili dataset, where all bullet chats fall within the BP start time and match end time of a certain match.

Huya. Huya is a mainstream livestreaming platform in China, with 150 million monthly active users (Wikipedia 2023b). We choose to include Huya because it mainly focuses on video game livestreaming, and it is also an official LPL livestreaming platform. Similar to Bilibili, we crawl all bullet chats from the livestreaming room. Since the livestreaming of the match is synchronized on different platforms, the recorded times in Bilibili could also be used to filter bullet chats in Huya. In total, 8,105,173 bullet chats from January 14, 2023, to March 26, 2023, are collected, and 3,928,780 bullet chats are kept after filtering.

Livestreaming Room. Because the duration of a match cannot be predicted in advance, there may be two matches taking place at the same time. When this happens, the two platforms will enable another livestreaming room called the side room for the later-started match, and the previous livestreaming room is called the main room. Once the side room is activated, it will work consistently until the end of the day's game. When the game played in the main room ends, the game in the side room will be played simultaneously in both the main and side rooms. Hence, a single match may be played in the main or the side room only, or it may be played in both rooms. In Section 4, we further analyze how this livestreaming room allocation mechanism affects the number of bullet chats for matches and teams. For its (potential) effects on the toxicity and topics, we leave them as future work.

4 Frequency Analysis

We first analyze how bullet chats are distributed during the livestreaming of the regular season of LPL 2023 by perform-

[8]https://twitchtracker.com/games/21779.

[9]https://www.sportspromedia.com/news/league-of-legends-pro-league-huya-broadcast-rights-china.

[10]https://www.sportspromedia.com/news/league-of-legends-bilibili-riot-games-streaming-rights-china-esports.

Dataset	# Bullet Chats	Filtered # Bullet Chats	# BO3	# Matches	# Teams	Time Range
Bilibili Huya	12,902,940 8,105,173	6,931,433 3,928,780	136	328	17	2023.01.14-2023.03.26

Table 1: Overview of the Bilibili and Huya Datasets.

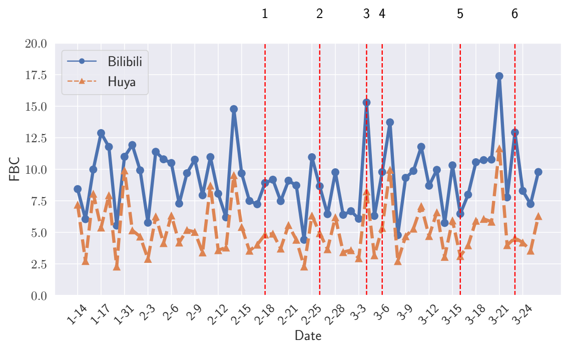


Figure 2: The FBC across the regular season of LPL 2023 on two platforms. The red dotted lines refer to hot events in Table 2. For simplicity, we omit the year (2023) on the x-axis.

NO.	Peak Day	Event (Source)
1	2023-02-18	Game player Bin in BLG is interviewed exclusively. (Baidu)
2	2023-02-26	A discussion about what ranking T1 (a team in League of Legends Champions Korea) can get if they play in LPL. (Baidu) Team OMG ends team TT's four-BO3 winning streak. (Baidu) ChatGPT gives the best lineup of LPL. (Douyin) Team WBG will face LNG in the LPL Spring Split. (Douyin)
3	2023-03-04	Team EDG defeats LNG with a record of 2-0 in the crucial battle. (Baidu) The former world championship team FPX ranks bottom in the 2023 LPL. (Baidu)
4	2023-03-06	Game player Uzi suggests that LPL should hold a super week. (Douyin) An analysis on why does commentator Jide Wang not commentate on LPL games anymore is posted. (Douyin)
5	2023-03-16	The top ten players for the 10th anniversary of the LPL are being voted on. (Douyin) The 2023 LPL Spring Split Standings are released. (Douyin)
6	2023-03-23	The list of top ten LPL players is released. (Douyin) The 2023 LPL Spring Split Standings are released. (Douyin)

Table 2: Hot events annotated on Figure 2.

ing quantitative analysis with respect to the entire season, each team, and each match.

Teams and matches that receive more attention are likely to receive more bullet chats. However, since each match may have a different length of time, directly analyzing the total number of bullet chats of each match or team is inappropriate, as a long match with low attention may get more bullet chats than a short match with high attention. Hence, We introduce a metric named frequency of bullet chats (FBC) that

$$\text{FBC} = \frac{\text{\# Bullet Chats between } t_a \text{ and } t_b}{\Delta_{a,b}},$$

where time interval $\Delta_{a,b} = t_b - t_a$, t_a and t_b denotes two different timestamp that holds $t_b > t_a$. This metric shows how frequently (# per second) users send bullet chats within a given time interval.

Season. Figure 2 shows the FBC of each day across the whole regular season. It is worth noting that there could be two or three BO3s within the same day. Thus, later, we will analyze each BO3 separately. From Figure 2, we see that Bilibili owns more bullet chats than Huya, and these two platforms' FBC is likely to change similarly with a Pearson correlation coefficient of 0.85 ($p < 0.01$). Furthermore, we look at the search index of "LPL" from the two platforms of Baidu[11] and Douyin[12], from January 14, 2023 to March 26, 2023 (i.e., the regular season date span of the LPL 2023). Table 2 summarizes the hot events that occur on the days when LPL's search index peaks, and we also mark the days in Figure 2. The search index and FBC usually do not peak on the same days, which indicates that the distribution of bullet chats cannot be represented by other mainstream media. Furthermore, we perform BERTopic (Grootendorst 2022) for bullet chats on each peak day in Table 2. For each peak day, among the top-20 topics, we find no topic related to the hot events of the day, which means the topics of bullet chats also cannot be represented by other mainstream media.

[11]https://www.baidu.com.

[12]https://www.douyin.com, a Chinese version of TikTok.

Team. Recall that the livestreaming for a team's matches could be played in different rooms. Figure 3 shows the percentages of livestreaming time and # of bullet chats in different livestreaming rooms, where all 17 teams and two platforms are considered. We observe that 1) some teams' (e.g., EDG with 0.70%) games are rarely played only in the side room, but some teams are different (e.g., WE with 19.47%), and 2) the speed of sending bullet chats in the side room is significantly lower than that in the main room (as for the side room, the percentage of the number of bullet chats is

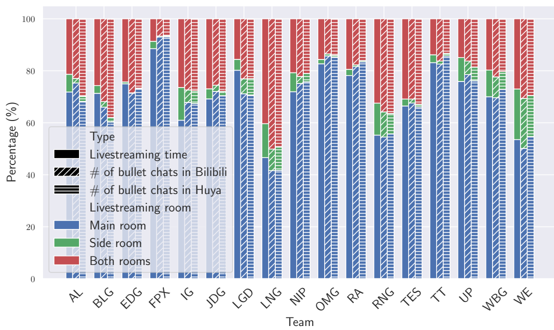


Figure 3: The percentages of livestreaming time and # of bullet chats in different livestreaming rooms on two platforms for the whole regular season.

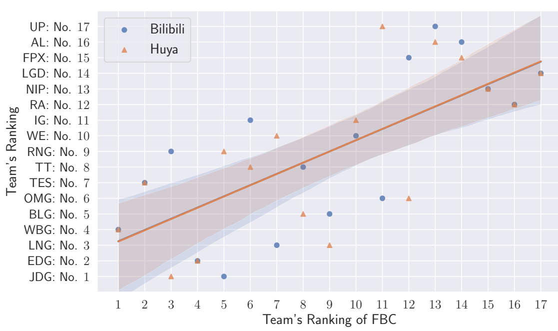
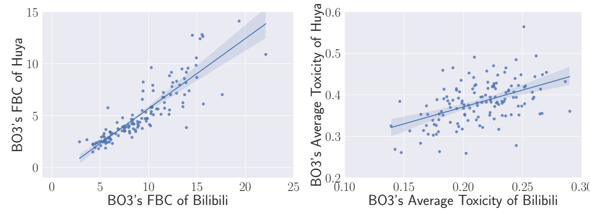


Figure 4: Each team's ranking of the FBC on two platforms and ranking in the regular season of LPL 2023.

generally lower than that of livestreaming time). To assess whether these differences affect the number of bullet chats received by each team, we calculate the relationship between teams' percentage of livestreaming time in the side room and teams' FBC during the regular season. On two considered platforms, we obtain the Pearson correlation coefficients of -0.05 on Bilibili ($p = 0.86$) and 0.01 on Huya ($p = 0.96$), demonstrating that this livestreaming room allocation mechanism does not affect the number of bullet chats received by the team. However, due to observation 2), we still appeal to livestreaming platforms to consider more effective mechanisms to eliminate the disparities in the number of bullet chats caused by differences in live streaming rooms.

Specifically, in Bilibili, WBG receives 1,244,144 bullet chats, representing the highest count amongst all participating teams. Conversely, team RA recorded 602,756 bullet chats, which is less than half of the bullet chats received by WBG. In Huya, WBG also serves as the team with the most bullet chats (745,411), and RA is the team with the fewest (301,872) bullet chats. Due to the page limit, we do not show the number of bullet chats for each team in this paper. Figure 4 shows the relation between each team's ranking in the regular season of LPL 2023 and the ranking of FBC on two platforms, where the lines are fitted linear regression lines. For both platforms, we could find a vivid relation that a team



(a) FBC

(b) Average Toxicity

Figure 5: The FBC and average toxicity of each BO3 on two platforms.

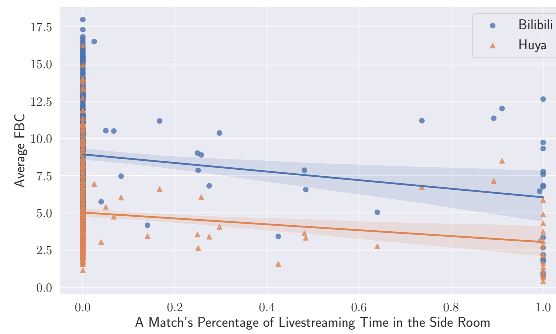


Figure 6: Each match's percentage of livestreaming time in the side room and average FBC on two platforms.

with a higher FBC is more likely to have a higher ranking in the game.

BO3 / Match. Recall that there is more than one BO3 in each game day. For each BO3, we show how frequently bullet chats are sent on two platforms in Figure 5a, where each dot represents a BO3 and the line is a fitted linear regression line. It is obvious that a BO3 that receives high attention on one platform (high FBC value) will also receive high attention on another platform. Among all BO3s on Bilibili, the match EDG vs. LNG on March 3, 2023, owns the largest FBC. On Huya, the match JDG vs. WBG on February 11, 2023, owns the largest FBC. Since all the aforementioned four teams are the top 4 teams this season, this could also support our finding that teams with higher rankings attract more bullet chats.

Moreover, we study how FBC changes within each match. Because there could be two or three matches in each BO3, we further measure how FBC changes in terms of each match. Due to the aforementioned livestreaming room allocation mechanism, a match could be played in different livestreaming rooms, and the side room receives relatively fewer bullet chats than the main room within the same time duration (see Figure 3). To assess how this mechanism affects the FBC of matches, we plot each match's percentage of livestreaming time in the side room and the average FBC in Figure 6. We note that, for both platforms, when the percentage in the side room increases, the average FBC of matches decreases with Pearson coefficients of -0.19 on Bilibili ($p < 0.01$) and -0.16 on Huya ($p < 0.01$). On av-

Method	Acc	Prec	Recall	F1
Perspective API	0.63	**1.00**	0.01	0.02
COLD	0.62	0.75	0.03	0.06
Fine-tuned COLD	**0.74**	0.67	**0.65**	**0.66**

Table 3: Performance of toxicity detectors.

Platform	# Samples	# Toxic	% Toxic	Mean	Var
Bilibili	346,560	69,412	20.03%	0.22	0.13
Huya	196,437	75,084	38.22%	0.39	0.20

Table 4: Statistics of toxicity on sampled datasets.

erage, the FBC of games only played in the main room is $1.68\times$ and $2.02\times$ that of games played in the side room. This observation further illustrates that a better livestreaming room allocation mechanism is needed. When comparing the average FBC of matches on two platforms, we derive a Pearson coefficient of nearly 1.0 ($p < 0.01$), which suggests that the differences in FBC caused by the existing livestreaming room allocation mechanism do not affect our further comparison of the two platforms.

The duration of different matches varies greatly, thus we first filter out matches with top 5% and bottom 5% lengths, then perform a KMeans clustering (Kanungo et al. 2002) with $K = 3$, where the number of clusters K is determined by the Elbow method (Bholowalia and Kumar 2014), to divide all matches into three disjoint groups based on the length of matches. However, the matches in the same group still have different time lengths, we further divide each match into one hundred time slots evenly and compute the average FBC of each time slot among all games in the same group. As shown in Figure 7, the FBC increases gradually from the BP start time, then drops rapidly at the fight start time. Thereafter, FBC gradually rises in fluctuations, and the highest peak appears in the final stage of the match. This shows that the more anxious stage the match is in, the more bullet chats users send. When analyzing the relationship between the two platforms, we see that Bilibili generally has a larger FBC than Huya, but during intense time periods (such as before the end of the match), Huya's FBC may surge to more than Bilibili's FBC. Moreover, the FBC of the two platforms seems to have a similar trend of change. For the Pearson correlation coefficient between the FBC of two platforms, the groups in Figure 7 receive a value of 0.82, 0.87, and 0.88, respectively, all with a p-value lower than 0.01.

5 Toxicity Analysis

To investigate how bullet chats shape the esports discussion ecosystem further, in this section, we perform quantitative analysis focusing on the toxicity of bullet chats among the season, teams, and single matches, indicating how toxic these bullet chats are.

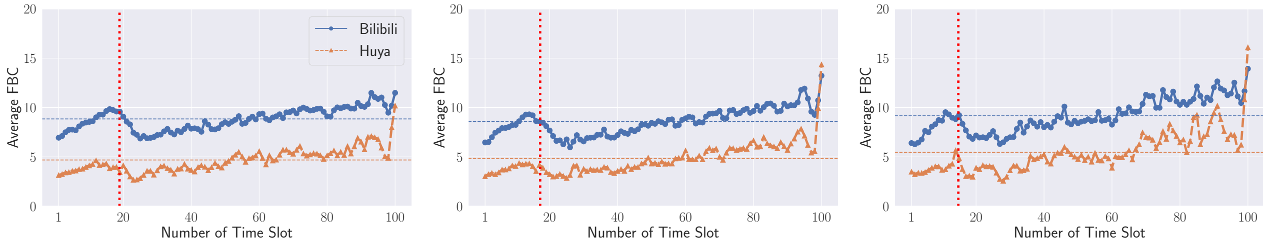
To identify the toxicity of bullet chats for LPL, we evaluate three classifiers, Google's Perspective API[13] and COLD (Deng et al. 2022), as well as our fine-tuned COLD. First, we randomly sample 1,200 bullet chats (600 from each platform) and let the authors annotate them (0 for non-toxic and 1 for toxic). Before annotation, all authors are told to annotate based on the definition that a toxic comment is rude, disrespectful, or unreasonable that makes you leave a discussion.[14] Each bullet chat is labeled twice by different authors, and we get Cohen's kappa coefficient of 0.76, which indicates substantial agreement between our authors. Then, we decide on the label of bullet chats with controversy through discussion. In our annotated 1,200 samples, 473 and 727 for toxic and non-toxic bullet chats, respectively. We further randomly divide them into a training dataset with 960 (80%) samples and a test dataset with 240 (20%) samples. Specifically, for Perspective API, we use the attribute TOXICITY and follow its instructions[15] to set the threshold to 0.7. For COLD and fine-tuned COLD, we apply a Softmax layer on the 2-dimension output and compute the element-wise product with vector $[0, 1]$, to get the final toxicity value (from 0 to 1) and the prediction (with a threshold of 0.5). In fine-tuning, we train the pre-trained COLD on our training dataset with learning rate of 1^{-5} for 5 epochs. Table 3 shows the performance of three classifiers on the test dataset. Both Perspective API and COLD have very low recall, which means many toxic comments are missed. Though they have decent precision, the limited value of F1-Score indicates that these two methods cannot meet our needs. Among all evaluated classifiers, our fine-tuned COLD stands out with the highest accuracy, recall, and F1 score. Hence, we utilize our fine-tuned COLD model to report the toxicity of bullet chats.

Season. By randomly selecting 5% samples from two datasets and filtering out blank samples, we get 346,560 and 196,437 bullet chats from Bilibili and Huya, respectively. Table 4 summarizes statistics of the toxicity value on sampled datasets. We find that, in Bilibili, among all bullet chats, about 20.02% of them are toxic, which is greatly lower than Huya (38.22%). Though Bilibili has more bullet chats than Huya, the number of toxic bullets in Huya is even more than in Bilibili. On average, there are 21,132 and 11,978 bullet chats for a match on Bilibili and Huya, which means when watching a match, an audience is likely to see more than **4,000** toxic comments. We further show the average toxicity for bullet chats on different days in Figure 8. We observe that the average toxicity for the two platforms is moderately similar with a Pearson correlation coefficient of 0.35 ($p < 0.01$), as the two platforms may reach peaks on the same day (e.g., 2-25). However, we also find some clear differences between the two platforms, indicating that different platforms have different toxicity distributions of bullet chats. For instance, considering the days with average toxicity greater than the mean (that is, above the horizontal dotted line in the figure 8) as high toxicity days, most (18 out of 31) of Bilibili's high toxicity days are distributed in the first half of the season. Nevertheless, for Huya, the majority (18 out of 29) are

[13] https://perspectiveapi.com.

[14] https://developers.perspectiveapi.com/s.

[15] https://developers.perspectiveapi.com/s/about-the-api-score.



(a) Group 1 (b) Group 2 (c) Group 3

Figure 7: The average FBC of matches on two platforms, where matches are categorized into three groups. The vertical red dotted line is the fight start time, and the horizontal line represents the mean value of the y-axis of the corresponding platform.

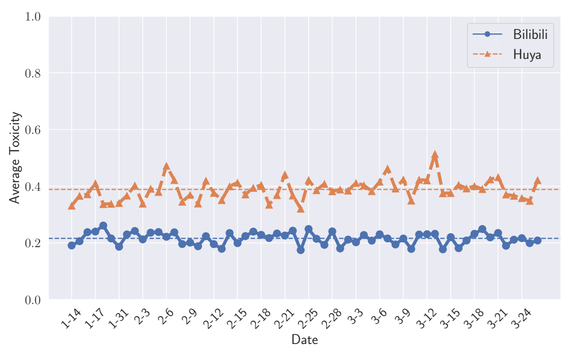


Figure 8: The average toxicity across the regular season of LPL 2023 on two platforms. The horizontal line represents the mean value of the y-axis of the corresponding platform. We omit the year (2023) on the x-axis scale for simplicity.

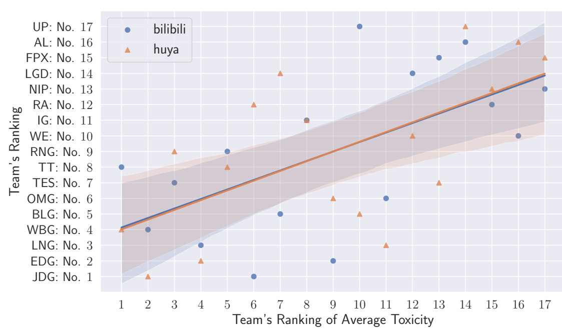


Figure 9: Each team's ranking of the average toxicity value on two platforms and ranking in the regular season of LPL 2023.

spread over the second half of the season.

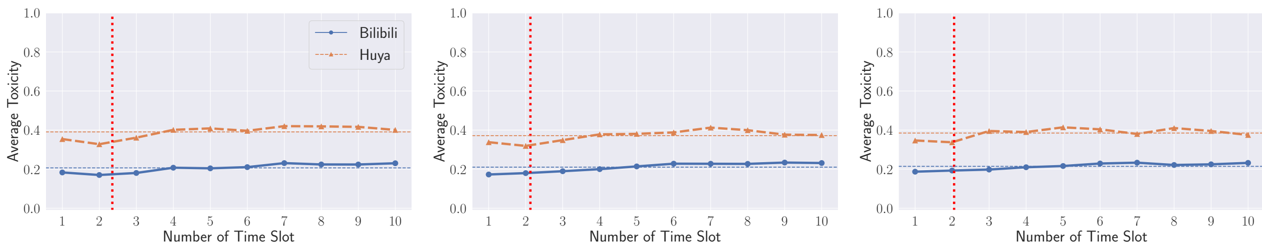
Team. We also randomly sample 5% samples from each team's bullet chats on two platforms, and feed them into our fine-tuned COLD to predict toxicity. Figure 9 plots each team's ranking of average toxicity on two platforms and ranking in this season, where the lines are fitted linear regression lines. Among all teams, in Bilibili, TT (0.24) and NIP (0.19) receive the highest and lowest average toxicity values, respectively. While on Huya, WBG (0.44) and FPX (0.35) are the two teams with the highest and lowest average toxicity values. It is worth noting that both TT and WBG are in the top half of the LPL's rankings, while NIP and FPX are in the bottom half. To show the relation between each team's ranking of this season and ranking of toxicity value, we plot two fitted lines in Figure 9. It indicates that, in both platforms, the team with a higher ranking in the season, is also more likely to have a higher average toxicity value.

BO3 / Match. First, we consider the average toxicity of each BO3 on two platforms, which is computed on 5% randomly sampled bullet chats for every BO3. As shown in Figure 5b, the fitted linear regression line reveals a positive relation between the average toxicity of BO3s on different platforms. In addition, the toxicity of different BO3s on the same platform differs. On Bilibili, the lowest average toxicity among all BO3s is 0.13 (AL vs. WE on February 17, 2023), but the highest reaches 0.30 (TT vs. OMG on February 17, 2023). On Huya, the BO3 WBG vs. OMG on March 13, 2023, receives the highest average toxicity of 0.56, which is 2.29$\times$ of the lowest one (FPX vs. TES on March 23, 2023).

Following the group setting in Figure 7, for each match in all groups, we sample 5% bullet chats in our evaluation. Also, due to the different lengths of matches and the decreased number of samples, for each group, we divide each match into ten time slots evenly and report the average toxicity of each time slot in Figure 10. We observe from both platforms that the value of toxicity basically shows an upward trend from the beginning to the end of the match, and it reaches the maximum value in the seventh or eighth time slot. Besides, the toxicity value can be differentiated in the middle stage of the match (the fifth and sixth time slots). Through this middle stage, the toxicity value grows from below to above average. Furthermore, in groups 1 and 2, the average toxicities of the two platforms have a strong positive correlation with a Pearson correlation coefficient of 0.87 and 0.88 ($p < 0.01$). Nevertheless, the coefficient in group 3 is 0.66 ($p = 0.04$), which indicates a weaker correlation.

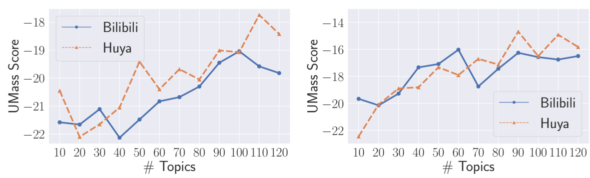
6 Topic Analysis

The former analysis uncovers how bullet chats are distributed among different platforms and how toxic these com-



(a) Group 1 (b) Group 2 (c) Group 3

Figure 10: The average toxicity of matches on two platforms, where matches are categorized into three groups. The vertical red dotted line is the fight start time, and the horizontal line represents the mean value of the y-axis of the corresponding platform.



(a) Sampled datasets (b) Sampled toxic datasets

Figure 11: UMass scores on different topics.

ments are. In this section, we dive into the content of bullet chats, by modeling what topics and toxic topics are discussed.

6.1 Topics on Games and Beyond

To understand the topics inside the bullet chats, we perform topic modeling for each platform. Following the previous study (Lu, Mei, and Zhai 2011), we randomly sample 5% bullet chats and leverage BERTopic (Grootendorst 2022) to extract topics from them with the default settings. The model produces 5,119 and 2,872 topics for Bilibili and Huya, respectively. Then, we hierarchically reduce the topic numbers to smaller values ranging from 10 to 120 with a stepping size of 10. We take the UMass score (Mimno et al. 2011) as the coherence measure for models with different topic numbers, where a higher value indicates a better model (Shen et al. 2022). For Bilibili and Huya, we observe that the UMass score reaches its highest value when the number of topics is set to 100 and 110, respectively, as shown in Figure 11. Therefore, we consider these numbers as the final choice for the number of topics.

By manually inspecting these topics, we find that these topics can be primarily divided into two scopes, i.e., games and beyond games. We then perform open coding on these topics and corresponding bullet chats. Concretely, two coders with specialized knowledge of LoL first independently code all topics based on the corresponding bullet chats. Subsequently, the resulting codes were discussed and refined to agree on a final code book. Then, the two coders independently coded the topics again and resolved the conflicts in the discussion. The final codebook (Table 5) includes ten codes/themes, namely game assets, players, etc., across two scopes. Our results show a good inter-coder agreement (kappa = 0.86 and 0.87 on Bilibili and Huya). Note that the bullet chats in a topic may have multiple themes; if so, we count it into all related themes. Besides, we include any topics that are not in the above themes into the "Others" theme.

Topics on Bilibili. As shown in Table 5, for Bilibili, the majority of topics within bullet chats (90.90%) fall within the domain of games. The most popular theme on Bilibili is game assets (41.41%), followed by the players (27.27%) and teams (16.16%) themes. This observation suggests that the predominant focus of audiences on Bilibili centers around game-related themes. Interestingly, we also find that 14.14% of topics discuss themes beyond the game on Bilibili, such as genders and entertainment stars. For instance, a sexist bullet chat that denigrated women by characterizing them negatively is observed. This demonstrates that the content carried by bullet chats of esports livestreaming has gone far beyond the game itself. Besides, Cai Xukun, a Chinese idol who sent a lawsuit notice to Bilibili about videos mocking him in 2019,[16] and his fans (named ikun) are often mentioned in the bullet chats.

Topics on Huya. For Huya, we observe 87.16% of topics are related to games while 22.94% of them are related to themes beyond games, respectively. Compared with Bilibili users, Huya users discuss more themes beyond games in bullet chats. Considering game-related themes, we find Huya users allocate the equally largest attention to two themes (players and teams), both comprising the highest proportion at 28.44%. At the same time, game assets-related topics also accounted for almost the largest proportion (27.52%). Regarding themes beyond games, we observe more topics related to genders and entertainment stars in Huya than in Bilibili. Furthermore, we find one theme related to non-esports athletes, which does not exist in Bilibili. For instance, a user in Huya posts: "James is the all-time No. 1 player,"[17] where James refers to LeBron James, an American professional basketball player.

[16]https://technode.com/2019/04/15/bilibili-threatened-with-lawsuit-about-videos-mocking-chinese-idol.

[17]To ensure readability for readers with different language backgrounds, the bullet chat examples have been translated from Chinese into English in this paper.

Scope	Theme/Code	Translation of Example	Description	Bilibili	Huya
Games	Game assets	What is the monkey[a] doing	Comments that contain in-game champions, equipment, operations, etc.	41.41%	27.52%
	Players	TheShy[b] tried his best	Comments that discuss the players' operations, records, etc.	27.27%	28.44%
	Teams	Come on EDG	Comments that discuss teams' rankings, performance, etc.	16.16%	28.44%
	Commentators	Say less, Mille[c]	Comments that discuss commentators' technique, appearance, etc.	6.06%	5.50%
	Audiences	Hater, speak up	Comments that discuss or provoke (online and offline) audiences.	3.03%	4.59%
	Others	The channel two of livestreaming is not stuck	Game-related comments that do not belong to any themes above.	11.11%	8.26%
Beyond Games	Genders	Bad woman	Comments that discuss different gender groups, such as female professionals.	3.03%	3.67%
	Entertainment stars	Is there any ikun[d] in the live room	Comments that discuss entertainment starts' acting skills, private lives, etc.	1.01%	7.34%
	Non-esports athletes	James[e] is the all-time No. 1 player	Comments that discuss non-esports athletes' performance, historical status, etc.	0.00%	7.34%
	Others	I don't eat anything that isn't Baixiang[f]	Beyond game comments that do not belong to any themes above.	10.10%	5.50%

[a] A champion named Monkey King Wukong in LoL.
[b] A game player in WBG.
[c] A regular commentator for LPL.
[d] Fans of Cai Xukun, a Chinese singer-songwriter, dancer, and rapper.
[e] This refers to LeBron James, an American professional basketball player.
[f] A food brand in China.

Table 5: The codebook on sampled bullet chats.

6.2 Toxic Topics

To further understand the themes of toxic bullet chats, we utilize our fine-tuned COLD to filter out toxic bullet chats used in Section 6.1. We perform the same process to topic modeling these toxic bullet chats. In the end, we obtain 60 and 90 toxic topics on Bilibili and Huya, respectively. Similar to the coding steps on normal bullet chats, we perform open coding on these toxic topics with a focus on identifying the target groups. In the end, these toxic topics are coded in six themes (kappa = 0.82 and 0.83 on Bilibili and Huya), as shown in Table 6.

Topics on Bilibili. We observe many topics that contain toxic bullet chats towards multiple targets and groups. Specifically, 91.66% of topics contain toxic bullet chats related to games and their stakeholders. For instance, a user of Bilibili posts: "RNG is bullshit," which contains slander against the team RNG. Besides, another bullet chat conveys an insult to a certain AD (a specific player role), saying: "who could tell me why the weakest AD player is not Huanfeng."

It is intuitive to observe such game-related toxic bullet chats in the game livestreaming. However, there are 6.25% and 4.17% topics containing toxic content against gender identities and celebrities outside of the games. Examples are "what a fucking slutty pregnant woman" and "is James even fit to carry Kobe's shoes."

Topics on Huya. As shown in Table 6, we also find topics containing the same themes in Bilibili. Specifically, topics that are toxic to games and their stakeholders account for a similar proportion (92.41%) on Huya as on Bilibili. Nevertheless, on Huya, the proportion of toxic topics targeting celebrities outside of the games reaches 8.86%, which is 2.12× than that on Bilibili. At the same time, the proportion of toxic topics on Huya related to gender groups is 3.80%, which is lower than Bilibili. Furthermore, we observe two themes that do not exist in Bilibili's toxic topics, showing severe discrimination and insult to different professions and regions, respectively. For instance, a Huya user posts: "sports students are all gorillas," which implies the inappropriate stereotype that sports students lack intelligence. Besides, another user demonstrates strong mockery and discrimination to different regions posts: "how can barbarians from Guangdong and Guangxi survive without eating mice."

Theme / Code	Translation of Example	Description	Bilibili	Huya
Games and their stakeholders	RNG is bullshit	Toxic bullet chats that target game assets and games' stakeholders, e.g., players, teams, etc.	91.66%	92.41%
Celebrities outside of the games	Is James even fit to carry Kobe's shoes[a]	Toxic bullet chats that target celebrities outside of the games, e.g., entertainment starts, non-esports famous athletes, etc.	4.17%	8.86%
Genders	What a fucking slutty pregnant woman	Toxic bullet chats that target different gender groups, e.g., women.	6.25%	3.80%
Professions	Sports students are all gorillas[b]	Toxic bullet chats that target people with different professions, e.g., sports students, deliverymen, etc.	0.00%	5.06%
Regions	How can barbarians from Guangdong and Guangxi[c] live without eating mice	Toxic bullet chats that target regions or the people living in these regions	0.00%	3.80%
Others	Extremely low in both quality, taste, and cultivation	Toxic bullet chats that do not belong to any of the themes above.	4.17%	1.27%

[a] Here LeBron and Kobe are LeBron James and Kobe Bryant respectively, they are both professional basketball players.
[b] This implies that they have physical prowess but lack intelligence.
[c] Guangdong and Guangxi are two provinces of China.

Table 6: The codebook on sampled toxic bullet chats.

7 Related Work

Toxicity in Esports. In (Bartle 1996), the concept of a "killer" type of game player was introduced, representing an early categorization of toxic players who derive pleasure from distressing others. This finding is paralleled in (Lin and Sun 2005), where it was observed that most toxic behaviors involved cursing other players.

With the rising popularity of MOBA games, particularly LoL, several studies have explored toxicity in LoL through text-based communication (Blackburn and Kwak 2014; Cook et al. 2019; de Mesquita Neto and Becker 2018) and in-game features (Grandprey-Shores et al. 2014; Kwak, Blackburn, and Han 2015; Kou 2020). For instance, (Blackburn and Kwak 2014) employed the Random Forest classifier to analyze over 10 million user reports with crowdsourced labels, categorizing toxic behaviors. Similarly, (Cook et al. 2019) conducted an in-depth analysis of messages sent by trolls, their teammates, and opponents in LoL, examining the characteristics of trolling interactions and their distribution among different actors. Moreover, (Kou 2020) developed a new taxonomy for toxic behaviors in LoL and conducted a comprehensive measurement of these behaviors. The above work has made vital contributions to defining, understanding, and governing toxic behavior in games. However, existing research primarily focuses on the interaction and exchange of toxic messages and behaviors within esports games. Bullet chat, as a novel and timely comment system for esports livestreaming, has not been well studied in previous studies.

Bullet Chats. In traditional sports, spectator behavior is often analyzed through indicators such as future attendance, intention to purchase team merchandise, and participation in online forums (Stavros et al. 2014; Chang and Wann 2022). In contrast, esports introduces a novel interaction through the real-time nature of bullet chatting, offering audiences an immediate platform for information exchange.

A significant body of research has been dedicated to understanding bullet chats in the context of viewer engagement and communication. (Chen, Gao, and Rau 2017) highlighted that audiences engage with videos featuring bullet chats for information, entertainment, and social connection. Extending this, (Li and Guo 2021) discovered that social presentation and information sharing are key motivations for users to participate in bullet chatting. To show how bullet chatting is (dis)similar to traditional comments, (Wu et al. 2018) conducted an analysis of 38K bullet chats and 16K forum comments on Bilibili, revealing a higher propensity for viewers to use bullet chats over forum comments while watching videos. This study also noted a tendency for bullet chats to contain more negative comments compared to forum discussions across various video genres. However, a contrasting observation was made by (He et al. 2021) in their study of COVID-19-related videos on Bilibili, where bullet chats were found to be less negative than forum comments. These findings indicate that the nature and tone of bullet chats can vary significantly depending on the video content and the context, underscoring the complexity of analyzing bullet chat content.

8 Discussion and Conclusion

In this paper, we study esports from the perspective of livestreaming bullet chats. With 21M bullet chats collected from two official livestreaming platforms, we quantitatively explore how bullet chats vary with respect to the season, team, and match. We find that teams with higher rankings consistently attract higher interest across all platforms. Additionally, our work contributes to a model specified for identifying toxic bullet chats, which outperforms existing toxic detection methods. Interestingly, while bullet chats are regarded as a novel comment system providing more timely and focused interaction with the game, we observe a considerable portion of bullet chats about (toxic) topics beyond the game, such as genders, entertainment stars, etc. Below, we discuss the implications of our findings for researchers and stakeholders interested in esports and bullet chats.

Consistency and Changes in Bullet Chats Across Platforms. Our research shows that the team is a significant factor in impacting the quantity of bullet chats. This is possible because LOL is a game where teams compete against each other. Even though the user base varies in platforms, teams with higher rankings consistently trigger more discussion. This is enlightening for advertisers. Placing ads with higher-ranking teams is more likely to gain more exposure. Additionally, while toxic bullet chats are observed on both platforms, their toxic content is significantly different due to their diverse user base. This can lead to further research on the differences in bullet chats among online communities.

Toxic Bullet Chat Detection. Bullet chats, with their unique characteristics like buzzwords and shorter lengths, pose a challenge for current toxicity detection methods. The model we propose in this paper for toxic bullet chat detection could be a pioneering tool for this new challenge. Given the substantial volume of toxic bullet chats identified in this study (over 4,000 for each match on both platforms), we see the detection of toxic bullet chats as a promising yet largely unexplored task. Considering the recurring patterns of toxic bullet chats observed in our study, incorporating temporal features in detection methods can also be a possible improvement direction.

Discussion Scope of (Toxic) Bullet Chats Has Gone Far Beyond Livestreaming Itself. A considerable portion of topics (14.14% on Bilibili and 22.94% on Huya) discuss themes beyond esports, including genders, entertainment stars, non-esports athletes, and so on, indicating that esports may appeal to audiences outside of its typical fanbase. An example is an identified sexist in our study, who uses bullet chats to spread hate speech against women in LPL livestreaming. The extensive user base and the instantaneous nature of bullet chats mean that toxic messages can reach a large audience quickly. Besides, given that large language models (LLMs) require massive training data, which could be collected from the Internet (e.g., bullet chats), bullet chats containing toxic information may cause LLMs to output harmful content toward different social groups. The aforementioned cases would highlight the urgency to examine how these (toxic) topics emerge, propagate, and evolve. Our study lays the groundwork for understanding bullet chats in esports livestreaming.

Limitations. Our work has limitations. First, we investigate bullet chats from livestreaming platforms in China. We acknowledge that bullet chatting deployed in other regions, e.g., Korea, may not follow the same findings in this work. However, as our annotators primarily have expertise in LPL, we leave the analysis in other regions for future work. Second, our analysis focuses on a specific esports game, League of Legends (LoL). There also exists other esports such as DotA and Counter-Strike: Global Offensive (CS: GO). Since LoL is one esports obtaining the largest esports user base and we are the first to study the ecosystem of bullet chats for esports games, we believe that focusing on one esports game across an entire game season can provide relatively comprehensive insights. We continue the investigation of other esports games for future studies.

References

Aguerri, J. C.; Santistebanand, M.; and Miró-Llinares, F. 2023. The Enemy Hates Best? Toxicity in League of Legends and Its Content Moderation Implications. *European Journal on Criminal Policy and Research*.

Bartle, R. 1996. Hearts, Clubs, Diamonds, Spades: Players Who Suit MUDs. *Journal of MUD Research*.

Bholowalia, P.; and Kumar, A. 2014. EBK-Means: A Clustering Technique based on Elbow Method and K-Means in WSN. *International Journal of Computer Applications*.

Blackburn, J.; and Kwak, H. 2014. STFU NOOB! Predicting Crowdsourced Decisions on Toxic Behavior in Online Games. In *International Conference on World Wide Web (WWW)*, 877–888. ACM.

Chang, Y.; and Wann, D. L. 2022. Effects of Game Outcomes and Status Instability on Spectators' Status Consumption: The Moderating Role of Implicit Team Identification. *Frontiers in Psychology*.

Chen, Y.; Gao, Q.; and Rau, P.-L. P. 2017. Watching a Movie Alone Yet Together: Understanding Reasons for Watching Danmaku Videos. *International Journal of Human-Computer Interaction*.

Cook, C.; Conijn, R.; Schaafsma, J.; and Antheunis, M. 2019. For Whom the Gamer Trolls: A Study of Trolling Interactions in the Online Gaming Context. *Journal of Computer-Mediated Communication*.

de Mesquita Neto, J. A.; and Becker, K. 2018. Relating Conversational Topics and Toxic Behavior Effects in A MOBA Game. *Entertainment Computing*.

Deng, J.; Zhou, J.; Sun, H.; Zheng, C.; Mi, F.; Meng, H.; and Huang, M. 2022. COLD: A Benchmark for Chinese Offensive Language Detection. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 11580–11599. ACL.

Gough, C. 2022. Engagement with eSports in Selected Countries Worldwide in 2020, by Country. https://www.statista.com/statistics/1250078/esports-engagement-by-country. Accessed: 2022-06-18.

Gough, C. 2023a. Engagement with eSports in Selected Countries Worldwide in 2020, by Country. https://www.statista.com/statistics/443147/estimate-of-global-market-revenue-of-esports-by-region/. Accessed: 2022-08-29.

Gough, C. 2023b. eSports audience size worldwide from 2020 to 2025, by type of viewers. https://www.statista.com/statistics/490480/global-esports-audience-size-viewer-type. Accessed: 2023-08-29.

Grandprey-Shores, K.; He, Y.; Swanenburg, K. L.; Kraut, R.; and Riedl, J. 2014. The Identification of Deviance and Its Impact on Retention in a Multiplayer Game. In *ACM Conference on Computer Supported Cooperative Work (CSCW)*, 1356–1365. ACM.

Grootendorst, M. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *CoRR abs/2203.05794*.

Hamari, J.; and Sjöblom, M. 2017. What Is eSports and Why Do People Watch It? *Internet Research*.

He, C.; He, L.; Lu, T.; and Li, B. 2021. Beyond Entertainment: Unpacking Danmaku and Comments' Role of Information Sharing and Sentiment Expression in Online Crisis Videos. *Proceedings of the ACM on Human-Computer Interaction*.

Howarth, J. 2023. How Many Gamers Are There? (New 2023 Statistics). https://explodingtopics.com/blog/number-of-gamers. Accessed: 2023-08-10.

Huang, Z.; Xu, S.; Xue, F.; Zhao, L.; Tan, Z.; and Chen, Y. 2023. Bullet Chatting Use Cases. https://w3c.github.io/danmaku/usecase.html. Accessed: 2023-06-12.

Kamberovic, R. 2023. League of Legends Player Count: These Are The Stats. https://riftfeed.gg/more/player-count-league-of-legends. Accessed: 2023-07-06.

Kanungo, T.; Mount, D. M.; Netanyahu, N. S.; Piatko, C. D.; Silverman, R.; and Wu, A. Y. 2002. An Efficient k-Means Clustering Algorithm: Analysis and Implementation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Kou, Y. 2020. Toxic Behaviors in Team-Based Competitive Gaming: The Case of League of Legends. In *ACM SIGCHI Annual Symposium on Computer-Human Interaction in Play (CHIPLAY)*, 81–92. ACM.

Kou, Y.; and Gui, X. 2014. Playing with Strangers: Understanding Temporary Teams in League of Legends. In *ACM SIGCHI Annual Symposium on Computer-Human Interaction in Play (CHIPLAY)*, 151–169. ACM.

Kou, Y.; and Gui, X. 2020. Emotion Regulation in eSports Gaming: A Qualitative Study of League of Legends. *Proceedings of the ACM on Human-Computer Interaction*.

Kwak, H.; Blackburn, J.; and Han, S. 2015. Exploring Cyberbullying and Other Toxic Behavior in Team Competition Online Games. In *Annual ACM Conference on Human Factors in Computing Systems (CHI)*, 3739–3748. ACM.

Lambert, C.; Rajagopal, A.; and Chandrasekharan, E. 2022. Conversational Resilience: Quantifying and Predicting Conversational Outcomes Following Adverse Events. In *International Conference on Web and Social Media (ICWSM)*, 548–559. AAAI.

Li, X. 2017. The League of Legends Worlds final reached 60 million unique viewers. https://dotesports.com/league-of-legends/news/lol-worlds-final-viewership-18796. Accessed: 2017-11-17.

Li, Y.; and Guo, Y. 2021. Virtual Gifting and Danmaku: What Motivates People to Interact in Game Live Streaming? *Telematics and Informatics*.

Lin, H.; and Sun, C.-T. 2005. The" White-Eyed" Player Culture: Grief Play and Construction of Deviance in MMORPGs. In *DiGRA International Conference (DiGRAC)*.

Lu, Y.; Mei, Q.; and Zhai, C. 2011. Investigating Task Performance of Probabilistic Topic Models: An Empirical Study of PLSA and LDA. *Information Retrieval Journal*.

Mei, F. 2021. Bullet Chats in China: Bilibili, language, and interaction. *Transformative Works and Cultures*.

Mimno, D.; Wallach, H. M.; Leenders, E. T. M.; and McCallum, A. 2011. Optimizing Semantic Coherence in Topic Models. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 262–272. ACL.

Shen, X.; He, X.; Backes, M.; Blackburn, J.; Zannettou, S.; and Zhang, Y. 2022. On Xing Tian and the Perseverance of Anti-China Sentiment Online. In *International Conference on Web and Social Media (ICWSM)*, 944–955. AAAI.

Stavros, C.; Meng, M. D.; Westberg, K.; and Farrelly, F. 2014. Understanding Fan Motivation for Interacting on Social Media. *Sport Management Review*.

Telefonica. 2023. What are e-sports and why are they so popular? https://www.telefonica.com/en/communication-room/blog/what-are-e-sports-and-why-are-they-so-popular. Accessed: 2023-02-22.

Torres-Toukoumidis, A. 2022. *Esports and the Media: Challenges and Expectations in a Multi-screen Society*. Taylor & Francis.

Webb, K. 2019. More than 100 million people watched the 'League of Legends' World Championship, cementing its place as the most popular esport. https://www.businessinsider.com/league-of-legends-world-championship-100-million-viewers-2019-12. Accessed: 2019-12-18.

Wikipedia. 2023a. 2019 League of Legends World Championship Final. https://en.wikipedia.org/wiki/2019_League_of_Legends_World_Championship_final. Accessed: 2023-08-14.

Wikipedia. 2023b. 2019 League of Legends World Championship Final. https://en.wikipedia.org/wiki/Huya_Live. Accessed: 2023-08-04.

Wikipedia. 2023c. Danmaku Subtitling. https://en.wikipedia.org/wiki/Danmaku_subtitling. Accessed: 2023-08-10.

Wu, Q.; Sang, Y.; Zhang, S.; and Huang, Y. 2018. Danmaku vs. Forum Comments: Understanding User Participation and Knowledge Sharing in Online Videos. In *ACM International Conference on Supporting Group Work (GROUP)*, 209–218. ACM.

Zhang, L.-T.; and Cassany, D. 2020. Making Sense of Danmu: Coherence in Massive Anonymous Chats on Bilibili.com. *Discourse Studies*.

9 Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes.
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes.
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes.
   (e) Did you describe the limitations of your work? Yes.
   (f) Did you discuss any potential negative societal impacts of your work? Yes, see Section 8.
   (g) Did you discuss any potential misuse of your work? Yes, see Section 8.
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? No.
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes.
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? NA
   (b) Have you provided justifications for all theoretical results? NA
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA
   (e) Did you address potential biases or limitations in your theoretical framework? NA
   (f) Have you related your theoretical results to the existing literature in social science? NA
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? No.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes, see Section 5.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? No.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? No.
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes, we evaluate toxicity detectors on manually labeled data and report their performance in Table 3.
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? Yes, we recognize the potential for false positives in toxicity detection. Therefore, we only report statistically significant results in our work.
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? Yes.
   (b) Did you mention the license of the assets? No.
   (c) Did you include any new assets in the supplemental material or as a URL? Yes, for instance, the Google's Perspective API.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes, in the Ethical Considerations.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes, in the Disclaimer and Section 1.
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? NA
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? NA
6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...
   (a) Did you include the full text of instructions given to participants and screenshots? NA
   (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA
   (d) Did you discuss how data is stored, shared, and deidentified? NA