2024 27th International Conference on Computer and Information Technology (ICCIT)
20-22 December 2024, Cox's Bazar, Bangladesh

Analyzing Audience Engagement in Esports: Sentiment and LLM-Based Topic Insights from Live Chats in South Asia

Md. Tanbeer Jubaer*, Mayeesha Farjana†, Barisha Chowdhury‡, Md. Shahid Uz Zaman§, Azmain Yakin Srizon¶, Md. Minhazul Islam‖
*†‡§¶‖ Department of Computer Science & Engineering
Rajshahi University of Engineering & Technology (RUET), Rajshahi, Bangladesh
E-mails: *tanbeerjubaer@gmail.com, †farjana.mayeesha@gmail.com, ‡nitub81@gmail.com,
§szaman22.ruet@gmail.com, ¶azmainsrizon@gmail.com, ‖minannu.2001@gmail.com

***Abstract*—This study investigates the dynamics of audience engagement in esports through sentiment analysis of live chat data and topic discovery, focusing on the popular game PUBG Mobile across Bangladesh, India, and Pakistan. A dataset encompassing nearly 15 million live chat messages and video metadata was utilized, employing a pre-trained RoBERTa model for sentiment classification, categorizing user sentiments into positive, negative, and neutral. The analysis revealed a significant increase in positive sentiment among Bangladeshi viewers, suggesting a shift towards a more favorable perception of esports. Additionally, the Gemma 7b-it large language model was applied to identify key discussion topics within the live chat, uncovering themes related to gameplay strategies and community interactions. The findings indicate a strong correlation between view counts and audience engagement, highlighting opportunities for advertisers to connect with dedicated esports fans. Despite limitations such as the focus on official YouTube channels and the resource constraints of sentiment analysis, this research offers valuable insights into real-time audience engagement in esports, paving the way for future studies to explore broader contexts and multimodal data integration.**

***Index Terms*—Esports, YouTube, RoBERTa, LLM, Sentiment, Topic Modeling**

I. Introduction

The esports industry has experienced exponential growth over the past decade, establishing itself as a mainstream form of entertainment with global appeal. By 2023, it was estimated that over 532 million people worldwide regularly engage with esports, either as competitors or as spectators on platforms such as Twitch, YouTube, and Facebook [1]. These platforms not only broadcast live gameplay but also enable real-time audience interaction through live chat features, providing an active space for viewers to express their thoughts, emotions, and opinions during events. This feature-rich environment offers a valuable opportunity to analyze audience sentiment and engagement, which has been relatively underexplored in the context of esports. Sentiment analysis, a branch of natural language processing (NLP), has emerged as a powerful tool to assess emotions, attitudes, and opinions expressed in text. In the context of esports, sentiment analysis of live chat can provide insights into the audience's reaction to in-game events, player performances, and match outcomes.

Building upon this foundation, the research further explores user engagement by integrating video metadata from multiple YouTube esports channels. This dual approach—analyzing both live chat and video metadata—enables the capture of real-time audience sentiment and an understanding of how video-specific factors such as views, likes, and duration correlate with audience interaction levels. By including data from different countries (Bangladesh, India, and Pakistan), a more comprehensive, cross-cultural view of audience behavior during esports tournaments is offered. Moreover, large language models (LLMs), such as the Gemma 7b-it model, are utilized to enhance the depth of topic discovery within the chat data. LLMs excel at uncovering thematic trends, allowing for the exploration of broader contextual insights related to gameplay, player performance, and fan reactions. The generative capabilities of LLMs facilitate nuanced topic modeling, further expanding the understanding of community discourse during esports events.

Previous studies have focused on analyzing social media sentiment related to sports events, entertainment broadcasts, and online communities [2] [3], but few have delved into the real-time sentiment dynamics unique to esports live chat, where interactions are driven by fast-paced, event-driven commentary. This paper presents a comprehensive framework for conducting sentiment analysis on live chat data from esports events. A large dataset of live chat messages collected during multiple esports tournaments is utilized, applying standard natural language processing (NLP) techniques alongside state-of-the-art machine learning models to classify and quantify audience sentiment. The study explores the correlation between audience sentiment and specific in-game events such as kills, team objectives, or turning points in matches. Additionally, temporal sentiment patterns are investigated to understand how audience reactions evolve throughout the course of an esports event.

The primary contributions of this work are threefold. First, a sentiment analysis model tailored to the linguistic character-

979-8-3315-1909-4/24/$31.00 ©2024 IEEE

istics of esports live chat is developed, addressing the prevalence of abbreviations, jargon, and rapid-fire text exchanges. Second, an exploratory data analysis (EDA) is performed to identify trends in sentiment during key in-game moments, contributing to a better understanding of fan engagement in esports. Third, the analysis provides actionable insights for event organizers, broadcasters, and game developers on how to improve audience experience by integrating real-time audience feedback. For example, broadcasters can adapt their commentary or presentation style based on audience sentiment trends, and game developers can use this feedback to refine the spectator experience in future updates. This research builds on previous sentiment analysis studies in similar domains, such as social media analysis during traditional sports events [4], and contributes to the emerging field of esports analytics. The methodology outlined in this paper demonstrates how live sentiment analysis can be systematically applied to esports, offering a real-time understanding of audience engagement, emotional responses, and viewer satisfaction. Moreover, by examining the emotional trajectory of viewers during key moments in a match, a deeper insight is provided into how audience sentiment can fluctuate in response to high-stakes plays and pivotal moments. Furthermore, this study not only enhances the current understanding of audience behavior in esports but also serves as a foundation for future research into real-time sentiment analysis in dynamic, high-engagement environments.

II. Literature Review

Esports has rapidly evolved from informal gaming competitions into an international phenomenon rivaling traditional sports in terms of fan engagement and media coverage, attracting millions of viewers worldwide. With the growth of streaming platforms such as YouTube and Twitch, esports fans can interact with live broadcasts in real-time, providing a rich data source. This data resource has broadened the scope of research conducted to navigate the emerging growth of this field across different regions of the world.

Assessing the trend and direction of the esports phenomenon is crucial due to its promising commercial potential. A comparatively recent study [5] focused on four representative esports games: TEKKEN 7 (Fighting), Dota2 (Multiplayer Online Battle Arena - MOBA), PUBG (Battle Royale - BR), and CS:GO (First-Person Shooter - FPS). The study analyzed their Steam reviews utilizing a BERT-based model and performed aspect-based sentiment analysis (E2E-ABSA). The analysis revealed that players and spectators had more critical views regarding specific gameplay elements and issues than their overall enjoyment of the game's graphics and design.

Studies have demonstrated the effectiveness of machine learning (ML) techniques in sentiment analysis and trend detection within esports communities. In a study by [6], sentiment analysis was performed on Danmaku live comments employing the Naïve Bayes method. The performance of multiple sentiment classification algorithms, including the proposed Sentiment Dictionary-Naïve Bayes (SD-NB) model, N-gram-Naïve Bayes (N-gram-NB), N-gram-Support Vector Machine (N-gram-SVM), and TextCNN were compared. The SD-NB model outperformed other methods, achieving an accuracy of 88.2%. In a study [7], machine learning models such as LR, SVC were used to detect sentiment from Reddit comments on the Contemporary Israel-Palestine Conflict.

Certain ML algorithms have proven to be more adept at handling nuanced text data. A study [8] explored several machine learning algorithms, such as Naive Bayes, Support Vector Machines (SVM), and Random Forests, to classify the sentiments of live chat messages as positive, negative, or neutral. The findings indicated that some models, particularly Random Forest and SVM, performed better in detecting sentiment in the Twitch chat environment due to their ability to handle the noisy and informal language often found in live chats. However, it is noted that ML techniques often fail to understand the contextual nuances of live comments and texts.

A plethora of research has been conducted on Twitch.tv as a primary online streaming platform. A study [9] utilized three feature selection methods on Twitch live comments: Least Absolute Shrinkage and Selection Operator (LASSO), Support Vector Machine-Recursive Feature Elimination (SVM-RFE), and Chi-square test to identify important keywords that could predict the number of views in live streaming. Support Vector Machine (SVM) was then used to evaluate the performance of the candidate feature subsets identified through the feature selection methods.

Deep learning has also been applied in the context of viewer engagement analysis through live chat. In [10], the L-Char-LSTM model was employed to process audience chat comments at the character level for video highlight prediction, allowing it to handle the complexities of internet-style slang, multilingual text, misspellings, emojis, and abbreviations common in online chat. For example, words like "happy" might be misspelled as "hapy" or exaggerated as "happpppy", and character-level processing allows the model to recognize these variations. [11] deployed three word-embedding schemes (word2vec, fastText, and GloVe) and five deep learning architectures: Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Bidirectional RNN with attention mechanism, Gated Recurrent Unit (GRU), and Long Short-Term Memory (LSTM) networks alongside conventional supervised ML approaches. Long Short-Term Memory (LSTM) networks achieved the highest predictive performance, demonstrating the efficiency of deep learning methods.

Different state-of-the-art models have proven to be more sophisticated in the task of live chat analysis. A study [12] deployed various BERT models, including BERT, XLM-RoBERTa, DistilBERT, WangchanBERTa, and TwHIN-BERT to detect hate speech in football news live streaming chat. The results showed that XLM-RoBERTa performed the best, yielding an F1 score of 0.9669. Similarly, [13] applied BERT, RoBERTa, and XLNet, with RoBERTa outperforming the others.

In some cases, topic modeling has been more appropriate

TABLE I: Video Information Data Example

ID	Title	Views	Likes	Duration	Date
knch7Sh_6Jc	[EN] 2024 PMWC x EWC Main Tournament Day 3..	6914	133	PT6H3M30S	2024-07-28
SaDGubynxoU	[EN] 2024 PMWC x EWC Main Tournament Day 2..	5344	44	PT6H49S	2024-07-27

TABLE II: Live Comment Data Example

Datetime	Author	Message	Video ID
2021-12-02 16:53:11	ea420f90	overall kill leaderboard dikhao :folded_hands:	x5QxAF7Y5ZE
2024-06-30 06:30:54	3bf5a7c4	hello hello	szvvDZCxwXs

for discovering the factors that drive viewer engagement and shifts in player or viewer preferences. In [14], guided Latent Dirichlet Allocation (LDA) was applied to extract topics from real-time comments on Douyu.com. Another study [15] benchmarked LDA along with four different feature representations of TF-IDF. One of LDA's primary drawbacks is the requirement for the predicted number of topics as an input parameter. Another study [16] conducted topic modeling using various techniques, including word frequency analysis, bigram analysis, and structural topic modeling. Additionally, a study by [17] opted for TF-IDF representation and clustering techniques like Self Organizing Maps to identify significant events.

After sufficient evaluation of the relevant studies, it is evident that there is room for contributions in revealing the catalysts for esports development by assessing the underlying sentiment and topics in users' comments. Utilizing state-of-the-art models for both sentiment analysis and topic discovery can facilitate the drawing of accurate insights into fan sentiment and evolving trends. This can assist viewers in finding streaming videos according to their preferences and provide players with feedback on their team strategies and gameplay.

III. Materials and Methods

A. Dataset Description

Our dataset is uniquely tailored to esports analysis, consisting of live chat data specifically scraped from YouTube, unlike traditional esports datasets which often lack such granular, dynamic user interactions. The dataset utilized in this research comprises two primary types of data: video information and live chat information. Both data types were sourced from the official PUBG Mobile YouTube channels of three countries: Bangladesh, India, and Pakistan. The dataset encompasses a comprehensive collection of live-streamed videos, specifically focusing on user engagement and interactions during these live broadcasts.

The video information dataset includes metadata associated with live videos streamed on the official PUBG Mobile YouTube channels. This data was collected for each video and encompasses attributes such as video_id, title, views, likes, duration, etc. Example of video information dataset is presented in Table I.

The live chat information dataset comprises nearly all messages posted in the live chat during each video stream. This dataset captures the real-time interactions of the audience, offering a window into audience sentiment and engagement. The attributes collected for each chat message include time, author, message, and video_id. To protect user privacy, the author names have been encrypted. An overview of the live chat data is provided in Table II. The live chat data furnishes a rich set of user-generated content that reflects the audience's reactions, emotions during the live broadcasts.

Data collection was performed manually using the YouTube API and the pytchat module. A threshold was applied, and data was collected spanning from 2020 to 2024. Specifically, data was gathered from 475 videos for the Bangladesh channel, 475 videos for the Pakistan channel, and 567 videos for the India channel. Regarding live chat messages, approximately 2.8 million comments were collected from the Bangladesh, 3.38 million comments from the Pakistan, and 9.13M comments from India.

B. Data Preprocessing

To prepare the dataset for analysis, several preprocessing steps were applied to both the video information and live chat data. For the live chat messages, the focus was on cleaning and standardizing the text. This process involved the removal of special characters, URLs, and redundant whitespace. Additionally, a key transformation was performed by converting emoji text descriptions (e.g., ":smile:") into their corresponding emoji symbols to preserve the original sentiment and meaning of the chat messages.

For the video information, consistency in date and time formats was ensured, and any missing values were addressed by either imputing reasonable defaults or excluding incomplete records. These preprocessing steps were essential in cleaning, normalizing, and standardizing the dataset, thereby making it suitable for further analysis, including sentiment analysis and topic modeling.

C. Sentiment Analysis

For the sentiment analysis of live chat messages, a pre-trained RoBERTa (Robustly Optimized BERT Pretraining Approach) model [18] was utilized. This model was fine-tuned for sentiment classification, assigning each chat message one of three labels: positive, negative, or neutral. RoBERTa's capability to handle informal and context-dependent language, as often observed in live chat messages, made it an ideal choice for capturing the emotional tone of user interactions during the live streams.

D. Topic Discovery

To discover the key topics within the live chat data, Gemma 7b-it, a large language model (LLM) designed for topic modeling and text generation, was utilized. The model was applied to segments of the chat data to identify and extract recurring themes and prominent discussion points. This process uncovered major topics such as gameplay strategies, in-game

events, player performance, and overall community reactions. Gemma 7b-it's generative capabilities provided a broader understanding of user discourse throughout the streams.

For topic discovery, a chunk of text consisting of approximately 100 comments was provided as context, and a query was used to generate topics from them. After collecting the topics, they were used again as context, and a query was submitted to summarize these topics, ultimately refining the list to derive a more concise set of topics.

To analyze the relationships among various topics, hierarchical clustering was employed, and the results were visualized with a dendrogram. The process began by compiling a list of unique topics relevant to the research focus. The TfidfVectorizer from scikit-learn was then utilized to convert these topics into feature vectors, capturing the significance of each term in relation to the entire set of topics.

The Ward method was then applied to compute the linkage matrix, minimizing within-cluster variance. This approach revealed topic closeness, which was visualized in the dendrogram, showing clusters and their interrelationships.

IV. Experimental Outcomes

A. Experimental Settings

In the experimental setup, computational optimizations were applied to deploy the Gemma 7b-it model for topic discovery and RoBERTa for sentiment analysis, addressing the challenge of limited resources. For the Gemma 7b-it model, precision was reduced by using 4-bit precision for certain weights, along with 16-bit floating point (float16) for computation, optimizing memory usage and performance. BitsAndBytesConfig with NF4 quantization helped further reduce the model size without significantly impacting performance. Memory usage was minimized by disabling caching of past key values, and automatic device mapping allowed efficient use of available hardware. In terms of sentiment analysis with RoBERTa, Kaggle's T4×2 GPU was leveraged to handle large volumes of live chat data, ensuring accurate and efficient processing. These optimizations allowed both models to perform effectively within resource constraints.

B. Result Analysis

In Figure 1, a comparative relplot between Bangladesh and Pakistan video information is observed, showcasing changes in view counts per month across different years (2020 to 2024). The shaded regions represent the uncertainty or variability in the data for each country. Both countries experienced an increase in viewership in 2021, with Pakistan showing a sharper rise than Bangladesh. In 2022, Pakistan's view count peaks significantly, whereas Bangladesh sees a slight decline. From 2023 onwards, both countries demonstrate fluctuating patterns, but Bangladesh shows a steady rise in 2024, suggesting growing engagement. This trend indicates a resurgence of interest in online content, particularly in esports, after a period of decline following the pandemic. India is not included in this comparison since it receives significantly more views

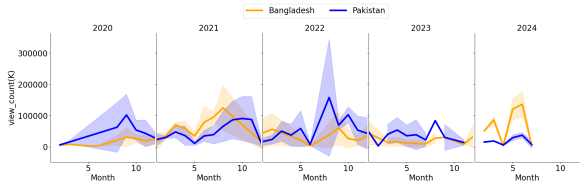

Fig. 1: Bangladesh and Pakistan Engagement(Views)

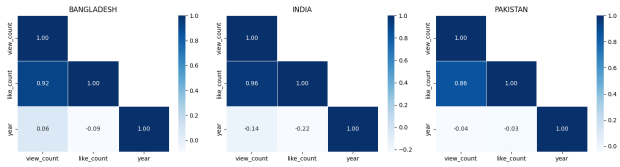

Fig. 2: Correlation analysis of Bangladesh, Pakistan & India

compared to Bangladesh and Pakistan. India will be addressed in the next section.

In Figure 2, for Bangladesh, the correlation between view count and like count is strong (0.92), showing a close relationship between the two, while their correlation with year is weak (0.06 and -0.09). In India, the view-like correlation is even stronger (0.96), though both metrics show a slight negative trend with time (-0.14 for view count and -0.22 for like count). For Pakistan, the correlation between view and like counts (0.86) is still strong, but the relationship with year is minimal (-0.04 and -0.03). In summary, the view count correlation is stronger for Bangladesh compared to India and Pakistan, suggesting that future engagement in Bangladesh is likely to increase.

In Figure 3, the live chat engagement for three countries—India, Pakistan, and Bangladesh—is observed. The data shows that India leads significantly, with over 9 million live chats, indicating a highly active audience in the esports scene. In comparison, Pakistan has recorded 3.38 million live chats since 2020, which is also a substantial number. Bangladesh, however, trails behind with only 2.2 million live chats. This suggests that, while Bangladesh is experiencing growth in esports engagement, its neighboring countries, particularly India and Pakistan, are advancing at a faster pace in terms of audience interaction and involvement in the esports domain.

Introducing the number of unique commentators is crucial, as it provides valuable insight into the number of individual users engaging through comments. In Figure 4, India leads with over 884K unique commentators, highlighting its large and active audience. Pakistan follows with 91K unique commentators, while Bangladesh, with 71K, ranks last. This data underscores the varying levels of interaction across these countries, with India exhibiting significantly higher engagement in comparison.

In one of the previous studies [19], the log frequency ratio was plotted to analyze U.S. political parties' attitudes toward immigrants. Similarly, eight frames—four positive and four negative—were taken and compared across Bangladesh, India, and Pakistan. The data from Figure 5 shows that in India, more

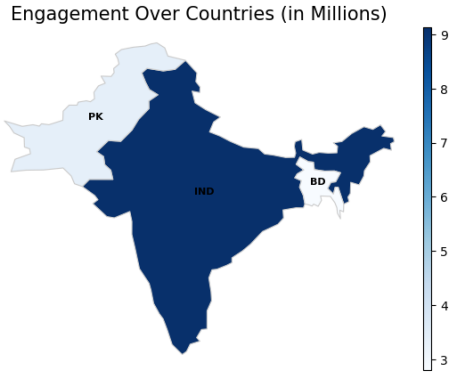

Fig. 3: Engagement over three countries shown on a worldmap

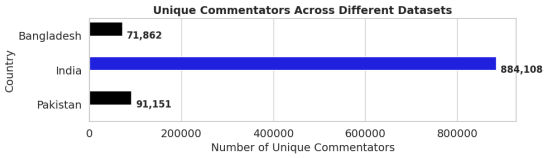

Fig. 4: Unique Audience

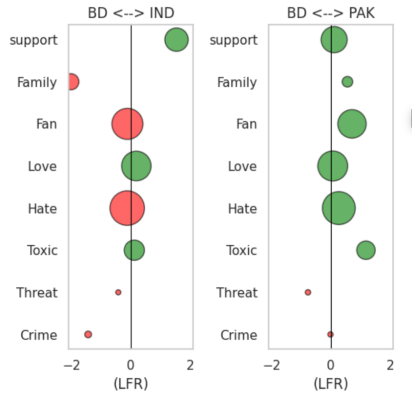

Fig. 5: Log Frequency Ratio

positive words are used, while in Bangladesh, negative words are more prevalent. The same pattern is observed in Pakistan, where Bangladeshis use negative frames such as “Threat” and “Crime” more frequently. This suggests that the fanbase in Bangladesh exhibits a more negative attitude compared to the other countries.

In Figure 6, a horizontal stacked bar chart displaying sentiment analysis trends from 2020 to 2024 is observed, with three categories: negative, neutral, and positive sentiments. The sentiment shift within the Bangladeshi community over the years is analyzed using the RoBERTa model, known for its effectiveness in understanding nuanced emotional expressions in text. The data reveals a positive trend, with positive sentiment showing consistent growth from 2020 to 2021 and into 2022. However, there was a noticeable decline in positive sentiment in 2023. It is important to note that since the complete data for 2024 is not yet available, the results for this year carry

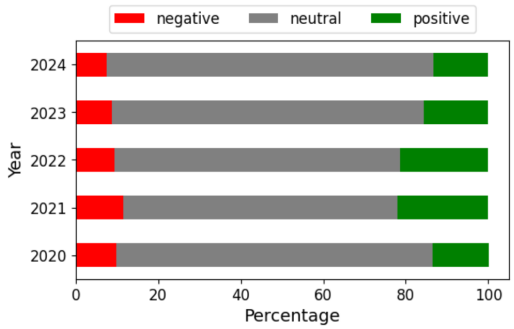

Fig. 6: Sentiment Over Year

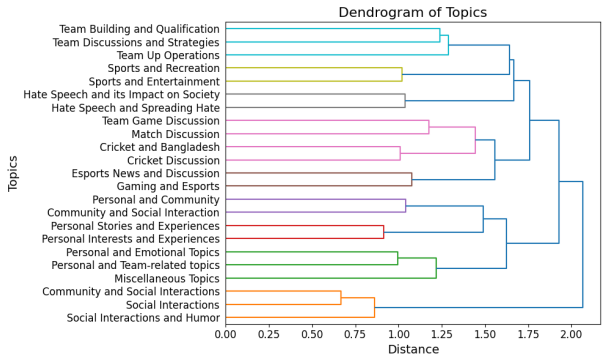

Fig. 7: Extracted Topics Using LLM (Gemma-7b-it)

less significance in the analysis.

On a more encouraging note, a decline in negative sentiment over the years is also observed, which is a promising development. This decrease suggests that the community’s overall perception is becoming more positive, indicating a shift towards a more optimistic outlook. Together, these trends highlight important dynamics within the Bangladeshi community, reflecting both challenges and improvements in sentiment over time.

In Figure 7, a dendrogram illustrating the hierarchical clustering of various topics is presented. The y-axis lists different topics, while the x-axis indicates the distance or dissimilarity between them. Topics positioned closely together on the dendrogram, such as “Team Discussions and Strategies” and “Team Up Operations”, indicate a greater similarity, whereas those that are farther apart represent more distinct themes. The hierarchical structure of the dendrogram allows the observation of how topics group into larger clusters. For instance, “Gaming and Esports” is closely associated with “Esports News and Discussion”, suggesting a strong relationship between these areas. In contrast, more distinct clusters include “Hate Speech” and “Social Interactions and Humor”.

This visualization is particularly valuable for understanding the natural grouping of topics based on their characteristics. Notably, topics such as “Cricket Discussion” and “Cricket and Bangladesh” indicate that within the esports section, discussions about cricket and other diverse subjects, such as personal stories, are present. This highlights the multifaceted

nature of conversations in the community, where interests extend beyond just esports.

C. Discussion

This analysis highlights key trends in esports engagement across Bangladesh, Pakistan, and India, offering valuable insights for stakeholders looking to invest in this growing market or target audiences with advertising. The comparative relplot indicates that interest in esports in Bangladesh is on the rise, particularly in 2024, suggesting a promising opportunity for brands to connect with a dedicated audience. The strong correlation between view counts and like counts signifies that viewers are engaged, making them more likely to respond positively to marketing efforts. While India leads in live chat activity and unique commentators, Pakistan's growth in viewership is noteworthy, signaling a potential goldmine for advertisers aiming to tap into emerging markets. The sentiment analysis reveals a shift toward more positive perceptions among Bangladeshi viewers, accompanied by a decrease in negative sentiment. This change creates an ideal backdrop for brands to engage with the community, allowing for the tailoring of messages to align with the evolving attitudes of the audience. Additionally, the hierarchical clustering of topics uncovers a wide range of interests, indicating that promotional strategies should extend beyond esports to include related themes such as cricket and social discussions. This broader approach enables more effective connections with the audience. For instance, cricket discussions were identified within the topics generated by the Gemma 7b-it model. Based on this insight, it is recommended that investors consider incorporating advertisements related to cricket, as the audience exhibits a strong interest in this sport as well. This strategy could enhance viewer engagement and brand relevance within the community.

V. Conclusion

This study presents a comprehensive sentiment analysis framework tailored to live chat data from esports events, specifically focusing on PUBG Mobile broadcasts across Bangladesh, India, and Pakistan. By leveraging natural language processing techniques and machine learning models, the research identifies sentiment trends that correlate with in-game events such as kills and match outcomes. The integration of the Gemma 7b-it large language model for topic discovery further uncovers thematic trends related to gameplay strategies, community reactions, and player performances, thereby enhancing the understanding of user discourse beyond mere sentiment analysis. These insights provide practical value for broadcasters and event organizers aiming to enhance the spectator experience and engage more effectively with their audience. However, the study is subject to several limitations. The analysis is confined to data from official PUBG Mobile YouTube channels within the Indian subcontinent, potentially overlooking significant engagement from unofficial channels and other geographic regions. Future research should address these limitations by incorporating a broader range of data sources, expanding geographical coverage, and utilizing more sophisticated language models to achieve a more holistic understanding of esports engagement.

References

[1] Newzoo, "Global esports market report," 2023. https://newzoo.com/resources/trend-reports/newzoo-global-games-market-report-2023-free-version.

[2] P. R. Cavalin, M. d. C. Gatti, C. N. dos Santos, and C. Pinhanez, "Real-time sentiment analysis in social media streams: The 2013 confederation cup case," *Proceedings of BRACIS/ENIAC*, vol. 2014, 2014.

[3] Y. Yu and X. Wang, "World cup 2014 in the twitter world: A big data analysis of sentiments in us sports fans' tweets," *Computers in Human Behavior*, vol. 48, pp. 392–400, 2015.

[4] Y. Xu, "Analyzing spectator emotions and behaviors at live sporting events using computer vision and sentiment analysis techniques," *Scalable Computing: Practice and Experience*, vol. 24, no. 3, pp. 475–486, 2023.

[5] Y. Yu, D.-T. Dinh, B.-H. Nguyen, F. Yu, and V.-N. Huynh, "Mining insights from esports game reviews with an aspect-based sentiment analysis framework," *IEEE Access*, vol. 11, pp. 61161–61172, 2023.

[6] Z. Li, R. Li, and G. Jin, "Sentiment analysis of danmaku videos based on naïve bayes and sentiment dictionary," *Ieee Access*, vol. 8, pp. 75073–75084, 2020.

[7] K. N. Nushin, M. S. Uz Zaman, and M. Ahmed, "Analyzing sentiment and unveiling geopolitical perspectives: A comprehensive study of reddit comments on the contemporary israel-palestine conflict," in *2024 6th International Conference on Electrical Engineering and Information Communication Technology (ICEEICT)*, pp. 788–793, 2024.

[8] A. Chouhan, A. Halgekar, A. Rao, D. Khankhoje, and M. Narvekar, "Sentiment analysis of twitch. tv livestream messages using machine learning methods," in *2021 fourth international conference on electrical, computer and communication technologies (ICECCT)*, pp. 1–5, IEEE, 2021.

[9] W.-K. Chen, L.-S. Chen, and Y.-T. Pan, "A text mining-based framework to discover the important factors in text reviews for predicting the views of live streaming," *Applied Soft Computing*, vol. 111, p. 107704, 2021.

[10] C.-Y. Fu, J. Lee, M. Bansal, and A. C. Berg, "Video highlight prediction using audience chat reactions," *arXiv preprint arXiv:1707.08559*, 2017.

[11] A. Onan, "Sentiment analysis on massive open online course evaluations: a text mining and deep learning approach," *Computer Applications in Engineering Education*, vol. 29, no. 3, pp. 572–589, 2021.

[12] P. Pookpanich and T. Siriborvornratanakul, "Offensive language and hate speech detection using deep learning in football news live streaming chat on youtube in thailand," *Social Network Analysis and Mining*, vol. 14, no. 1, p. 18, 2024.

[13] Z. Gao, S. Yada, S. Wakamiya, and E. Aramaki, "Offensive language detection on video live streaming chat," in *Proceedings of the 28th international conference on computational linguistics*, pp. 1936–1940, 2020.

[14] W. Wang and J. Fan, "Topic mining of real-time discussions: what catches the attention of live-streaming esports viewers?," *European Sport Management Quarterly*, vol. 24, no. 2, pp. 323–344, 2024.

[15] S. A. Curiskis, B. Drake, T. R. Osborn, and P. J. Kennedy, "An evaluation of document clustering and topic modelling in two online social networks: Twitter and reddit," *Information Processing & Management*, vol. 57, no. 2, p. 102034, 2020.

[16] P. A. Toussaint, M. Renner, S. Lins, S. Thiebes, and A. Sunyaev, "Direct-to-consumer genetic testing on social media: Topic modeling and sentiment analysis of youtube users' comments," *JMIR infodemiology*, vol. 2, no. 2, p. e38749, 2022.

[17] C. Comito, A. Forestiero, and C. Pizzuti, "Bursty event detection in twitter streams," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 13, no. 4, pp. 1–28, 2019.

[18] J. Camacho-collados, K. Rezaee, T. Riahi, A. Ushio, D. Loureiro, D. Antypas, J. Boisson, L. Espinosa Anke, F. Liu, and E. Martínez Cámara, "TweetNLP: Cutting-edge natural language processing for social media,"

[19] D. Card, S. Chang, C. Becker, J. Mendelsohn, R. Voigt, L. Boustan, R. Abramitzky, and D. Jurafsky, "Computational analysis of 140 years of us political speeches reveals more positive but increasingly polarized framing of immigration," *Proceedings of the National Academy of Sciences*, vol. 119, no. 31, p. e2120510119, 2022.