Participant profiling on Twitch based on chat activity and message content

Jari Lindroos[1], Ida Toivanen[1], Jaakko Peltonen[2], Tanja Välisalo[1,3], Raine Koskimaa[1] and Sami Äyrämö[1]

[1] *University of Jyväskylä, Seminaarinkatu 15, 40014, Jyväskylä, Finland*
[2] *Tampere University, Kalevantie 4, 33100 Tampere, Finland*
[3] *National Archives of Finland, Rauhankatu 17, 00170, Helsinki, Finland*

Abstract

Identifying participant profiles based on commenting activity in esports livestream platforms enhances our understanding of esports audiences and patterns of viewer engagement. Active chat participants represent a core audience of esports viewers due to their high level of engagement. In this study, we identify participant profiles based on chat data collected from Twitch livestreams of CS:GO Majors tournaments from 2022 and 2023. Profiling was conducted based on two types of features: chat activity and message content. We performed clustering to both sets of features to get insights about the communication patterns of chat participants and the contents of messages they sent during matches. The results show that livestreaming chat data even on a larger scale can be harnessed to understand participation in livestream chats from multiple viewpoints. Combining both of these approaches can give us a comprehensive way of analyzing and forming participant profiles based on chat participants' message behavior.

Keywords

esports, Twitch, clustering, chat, participant profiling

1. Introduction

Chat is an essential part of esports livestreams in multiple ways. The chat in esports livestreams in Twitch has been described as a significant part of the esports experience, "a proxy for noise, transmitting affects compelling continued viewing and consumption" [1]. Chat participation has been seen as part of 'audience work', which is essential for the esports economy via advertising, sponsorship, and various other revenue channels [2]. In-depth analysis of esports chat discussions is crucial for understanding audience's behavior, preferences, and opinions [3, 4]. Analyzing livestream chat messages helps us to gain a better understanding of what the viewers are interested in, what they dislike, and how they engage with the content, the stream provider and.

Recently, there has been a growing amount of machine learning research using massive chat data to discuss the chat cultures in livestream platforms like Twitch [5, 6, 7, 8, 9, 10, 11, 12, 13]. Manually going through massive amounts of chat data is a resource-heavy process, which has given us an incentive to focus on research that introduces large-scale datasets and makes primary use of machine learning solutions in the context of Twitch chats.

The chat data used in this study was collected during two Counter-Strike: Global Offensive Major Championships tournaments (CS:GO Majors) in 2022 and 2023, from two different broadcasters, Pelaajat.com and Yle on Twitch.tv. Using this data, we investigate audience engagement with esports livestreams by studying the behavior of esports spectators who participate in Twitch chat. To do this, we construct features in two ways: 1) describing the variation of chat participants' chat activity over time and matches, and 2) categorizing chat messages by content types. We refer to the first as "activity-based" and the second as "content-based". We use machine learning methods of exploratory

9th International GamiFIN 2025 (GamiFIN 2025), April 1-4, 2025, Ylläs, Finland.

jari.m.m.lindroos@jyu.fi (J. Lindroos); ida.m.toivanen@jyu.fi (I. Toivanen); jaakko.peltonen@tuni.fi (J. Peltonen); tanja.valisalo@jyu.fi (T. Välisalo); raine.koskimaa@jyu.fi (R. Koskimaa); sami.ayramo@jyu.fi (S. Äyrämö)
0000-0003-2275-3866 (J. Lindroos); 0000-0003-0440-7337 (I. Toivanen); 0000-0003-3485-8585 (J. Peltonen); 0000-0001-8678-4683 (T. Välisalo); 0000-0002-1492-4074 (R. Koskimaa); 0000-0002-7532-2771 (S. Äyrämö)

data analysis, particularly clustering and dimensionality reduction, to discover and analyze distinct participant profiles within the esports livestream audience based on their commenting behavior. We also examine whether the same structure of participant profiles is found from both tournaments.

The primary contributions of this work are:

1. We provide a method for forming participant profiles via clustering by identifying and categorizing features of chat participant activity and the types of content in those chat messages.
2. We propose that the participants' chat behavior can vary significantly – by forming six different activity-based participant profiles and 11 different chat content-based participant profiles. We then demonstrate the variety of these behaviors.

2. Previous research

2.1. Audience engagement in livestreaming

Research on audience engagement in livestreaming is largely focused on audience surveys or questionnaires [14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24], rather than the analysis of massive the chat data the audience produces. However, computational methods have also been employed to investigate this phenomenon, offering new ways for understanding the dynamics of audience engagement [25, 26].

Studies on chat activity itself have often focused on small datasets and qualitative methods exclusively. Having a larger span of adjacent viewers is important in discerning which communicative characteristics can be found from massive chat data. For example, in [27], Twitch chat datasets consisting of 50-message segments from livestreams with up to 10,000 concurrent viewers were analyzed by hand. The study suggested that the unique property of the type of communication found in massive chats could be described as "crowdspeak". Conversational norms usually found in small-scale interactions dissipate in a setting of massive chats, making repetitive and shorter speech more prominent to the degree of the chat appearing to be chaotic and serving no communicative purpose. However, the researchers found that in large-scale chat setting bricolage, shorthanding and voice-taking help in having more coherence in the communication. While bricolage refers to having a small set of elements that are re-used and re-arranged for further communicative use, shorthanding describes the deliberate choice to fit speech into a smaller frame. Sharing viewpoints and mannerisms in online communities is referred to as voice-taking. Conveying meaning has been found to be disrupted in massive online group settings in a similar study [28]. The researchers discovered that communication in massive chats tends to resemble a cacophony, which is represented by repetitive and information-poor messages, as well as lower per capita participation.

2.2. Machine learning in livestream chat analysis

Machine learning solutions have been employed in the analysis of Twitch livestream chat data to mitigate the limitations of time and resources of manual processing when going through massive amounts of data. This can be seen in studies regarding automatic chat bot detection [5] or identifying toxic language [9] and spam [10, 11], for instance.

To gain a deeper understanding of the livestream chat dynamics, machine learning has been used to predict viewer engagement, and also its impact on the popularity of a stream. For example, shallow artificial neural networks were used in [12] to predict low or high engagement based on gameplay events in PlayerUnknown's Battlegrounds livestreams. Their analysis encompasses both chat logs and game telemetry data, with specific gameplay features (e.g., player health status, in-game choices, and map location). The results demonstrate the possibility of accurately predicting continuous viewer engagement based exclusively on key gameplay events. Predicting the viewer count of a Twitch stream was investigated in [13]. The authors categorized these reactions into textual features, such as n-grams and sentiment, and non-textual features, including chat frequency and the active time of how long a viewer engages on the chat. They found that while textual features are important for predicting popularity when analyzing the entire chat log, non-textual features become more crucial for early prediction, within the first 15 minutes of a stream. More recently a novel machine learning based approach was proposed in [8] for analyzing emotional responses in Spanish video game streams on Twitch. A unique corpus of Spanish Twitch chat messages was created and manually annotated for polarity and emotions. The results show that a BERT-based model achieved the highest accuracy in detecting polarity (78%) and emotions (68%), outperforming other methods.

Clustering has also been used to analyze the differences between first-time visitors and regular participants in Twitch chats [29]. In the study, it was found that subscribers and regular users were the most common participants followed by moderators, leaving the new and other participants, like bots, to be the smallest group of participants. This may tell us that normative communication behavior in massive chats is maintained by those who have the most experience in interacting in said environments. Certain behaviors, such as messaging repetitively and outside typical streaming time, mentioning irrelevant topics and other channels, and having no response in chat interactions clearly stood out among the human-generated input and were linked to bot-generated content.

In [6], k-means clustering was applied for word vectors to reveal more about the innate qualities of chat data of 10,000 tokens collected from Twitch. They found that clustering word vector data is possible but unfolds odd shapes, depending on the chosen word vector method (e.g., skip-gram with negative sampling; see more in [30]). Features like streamer popularity were found to correlate with the two clusters found in the study. One cluster was interpreted as having more game-specific terminology and the other as likely consisting more of general speech and terms used throughout different Twitch channels.

One method frequently applied to chat data is topic modeling. For example, the Twitter-LDA topic modeling algorithm was employed in [31] to investigate the message content of chat participants from two Dota 2 tournament broadcasts from May 2016 and December 2016. The results suggested that various stages of the broadcast elicit distinct types of communication behavior among viewers depending on the content types. Active game sessions tended to include shorter and more emotional expressive messages, whereas during breaks and inactive moments more analytical discussions and social interactions occurred. Crowd behavior was also investigated in [3] to distinguish more meaningful and coherent communication among thousands of chat participants that contributed to upholding "crowdspeak" in chat. The authors utilized structural topic modeling and cross-correlation analysis to examine topical and temporal patterns of chat participants' messages during the Dota 2 International tournament in 2017, investigating their relationship to in-game events. They showed that in-game events significantly influence communication within large-scale chats, shaping the emergent topical structure. During inactive periods of game activity, boredom and frustration was expressed through emotes and recursively replicated "copypasta". During highly anticipated phases of the in-game events chat participants often trigger a high volume of short, emotionally charged messages and emotes (e.g., "ez", "PogChamp"). Chat participants also engage in ongoing discussion regarding what is happening on the screen – examples of this include cheering or supporting teams and players in chat messages. This behavior occurred with less correlation to in-game event triggers.

In summary, while prior research has applied machine learning to analyze Twitch chat from content moderation to predicting viewer engagement, the construction of comprehensive participant profiles integrating activity- and content-based features remains largely unexplored to our best knowledge. Our study aims to address these gaps, offering a more comprehensive view of engagement patterns and participant behavior on Twitch.

3. Data

This study used chat data from Twitch.tv to explore audience engagement dynamics. Twitch is a livestreaming platform owned by Amazon with a focus on games and related activities. Since its launch in 2011, Twitch has significantly shaped how esports content is consumed and understood. By using machine learning methods to perform exploratory data analysis, we aim to analyze chat participation patterns and model distinct participant profiles within the esports livestream audience.

Table 1
Statistics of the number of participants, messages and streamed matches for games streamed by both broadcasters, Yle and Pelaajat.com.

Majors	Yle Participants	Messages	Streamed matches	Pelaajat.com Participants	Messages	Streamed matches
PGL 2022	6319	87036	42	2734	20884	28
BLAST 2023	3522	32197	36	8366	85327	34
Total	8602 (unique users)	119233	78	10018 (unique users)	106211	62

Our dataset consists of Twitch chat messages collected during the 2022 and 2023 CS:GO Majors tournaments, titled PGL Major Antwerp 2022 and BLAST.tv Paris Major 2023. Both tournaments consisted of 70 individual matches played over 12 days. The data were collected from two Finnish Twitch channels: 'yleeurheilu' and 'pelaajatcom'. These channels served as the primary Finnish-language broadcasters for the events. Yleeurheilu is operated by the publicly funded national broadcaster Yleisradio (abbrev. Yle) and Pelaajat.com is run by the Finnish esports media company Pelaajat.com. Chat activity across both channels produced over 107,900 messages in 2022 and 117,500 in 2023 (see Table 1). Streaming activity was also distributed due to concurrent matches: Yle was the main broadcaster for the entirety of the 2022 Majors, including a majority of the games featuring ENCE and "Aleksib", while Pelaajat.com provided secondary coverage. This pattern was reversed in 2023, with streaming more evenly distributed and Yle remaining the primary broadcaster for the finals. Importantly, the presence of Finnish organizations, such as ENCE, or Finnish players, such as "Aleksib", was a significant factor influencing viewership patterns on both channels. This is noticed in the high viewership observed during matches featuring these players and teams, regardless of the broadcasting channel.

For clarity, we refer to our datasets with shortened terms YLE for data collected from 'yleeurheilu' and PCOM for data collected from 'pelaajatcom' so that, e.g., YLE22 refers to data collected from yleeurheilu in 2022, PCOM23 refers to pelaajatcom data collected in 2023 etc.

4. Methods

In this study we form participant profiles by conducting two types of clustering: activity- and content-based clustering. These two types of clustering are conducted to explore audience engagement based on two different types of distinct features: 1) features based on chat participant activity, and 2) features based on uniqueness of the message content produced by chat participants (see summary in Table 2). Given the differing nature of these features, we applied two different clustering evaluation metrics to ensure optimal cluster separation and interpretability for each feature type.

4.1. Activity-based clustering

To analyze the activity of chat participants within the tournament, we start by manually separating timestamps when a match occurs. This lets us distinguish between activity occurring during active gameplay and activity in the post and pre-match period. To further understand participants' behavior during the tournament, we define several features to construct participant profiles based on observable conversational participant traits. Such traits can be categorized into two distinct aspects by message- and time related information. These activity features are the following:

- **activity_count**: The total number of matches the participant has taken part of (chatting during the matches).
- **comment_count_mean**: The average number of messages sent by the participant across their participated matches.

Table 2
Summary of the methods used for determining cluster size (k), number of features, and type of feature scaling (if applied) for each clustering type.

Clustering	Method for determining k	Number of Features	Feature Scaling
Activity-based	Silhouette	7	QuantileTransformer
Content-based	Elbow	4	-

- **comment_count_std**: The standard deviation of the number of comments the participant made across matches.
- **comment_count_10th_quantile**: The value separating the lower 10% of a participant's matches in terms of number of comments made from the upper 90%.
- **comment_count_90th_quantile**: The value separating the lower 90% of a participant's matches in terms of number of comments made from the upper 10%.
- **timestamp_diff_to_match_start_mean**: The average time elapsed in seconds between the start of a match and the participant's messages.
- **timestamp_diff_to_match_start_std**: The standard deviation of time elapsed in seconds between the start of a match and the participant's messages.

Pre-examination of the preprocessed data of the activity features showed sensitivity to outliers. Due to this, we applied the QuantileTransformer from the scikit-learn library [32] to each feature of the activity data prior to clustering to transform the features into a uniform distribution. Previous studies have shown it to be effective in reducing the impact of outliers while still maintaining the distribution of the data. For instance in [33], the authors showed that QuantileTransformer spreads common values more evenly, minimizing outlier influence without distorting the overall feature scale. This allows for better scaling balance compared to standard scaling techniques, which can be disproportionately affected by outliers.

While the transformation in the method is non-linear and it could alter correlation between features, we still believe that its advantages in handling outliers outweigh this potential drawback in our specific case. Since our primary goal is to identify meaningful clusters in the presence of outliers. To further justify this choice, we conducted a comparative analysis using the scikit-learn library, evaluating the impact of various scaling methods (including no scaling, Normalizer, MinMaxScaler, RobustScaler, StandardScaler, MaxAbsScaler, PowerTransformer) on the clustering results. The evaluation of the clusters was focused on interpretability, such as cluster separation and compactness, and the QuantileTransformer yielded the most balanced and interpretable clustering results.

We apply k-means clustering using k-means++ initialization via scikit-learn to the processed dataset with the activity features to identify potential groups of participants exhibiting similar patterns of chat engagement behavior. To select the optimal number of clusters for activity-based clustering, we employ the silhouette method [34]. Finally to visualize the clustering results, we employ t-SNE to perform dimensionality reduction of the activity data, projecting the participants from the original high-dimensional feature space into two dimensions [35] where the participants can be visualized as a scatterplot and the discovered clusters can be depicted with different colors.

4.2. Content-based clustering

As another framework for creating participant profiles with clustering, we use deep learning based content detection [36] that gives us features based on the uniqueness of message content generated by chat participants. After describing the deep learning model used to create these features, we explain how we are forming participant profiles based on the uniqueness features.

Features for the second clustering approach, which is based on the participants' chat content, are constructed by quantifying each participant's total messages as a distribution in terms of uniqueness.

Before any data processing, we filter out known channel bots, such as Moobot and Nightbot, to focus the content on authentic participants. We utilize four distinct features:

- **unique**: messages which stand out as original, informative, and directly relevant to the contents of the stream;
- **non-unique**: repetitive content such as emotes, game slang and copypasta;
- **essentially non-unique but human-written**: Formulaic content based on variations of certain patterns such as reactions and cheers;
- **commands and replies**: commands written to a channel bot and the responses of the bot.

For each chat participant, we calculated the share of their messages for each feature (values are in between 0 and 1). No scaling is applied. We apply k-means clustering also using k-means++ initialization via scikit-learn to the uniqueness features to identify potential groups of participants exhibiting similar chat engagement patterns. To select the optimal number of clusters for content-based clustering, we employ the elbow method [37]. Finally to visualize the clustering results, we employ t-SNE to perform dimensionality reduction, projecting the participants from the high-dimensional feature space of content features into two dimensions [35], where we then visualize participants as a scatterplot with clusters shown as colors of the participants.

5. Results

In this section we compare the results of the two different types of clustering, activity-based (features based on chat participant activity) and content-based (features based on uniqueness of the message content produced by chat participants).

5.1. Activity-based clustering

Looking at the number of clusters determined by the silhouette score for the activity-based features, we get varying values across datasets: YLE22 with six clusters (silhouette score = 0.5529), YLE23 with four clusters (silhouette score = 0.5631, six clusters score = 0.5506), PCOM22 with four clusters (silhouette score = 0.6164, six clusters score = 0.5905), and PCOM23 with six clusters (silhouette score = 0.5910). According to [38], an average silhouette score of 0.5 or higher provides good evidence that the clusters are clearly distinguishable. For the sake of interpretability and consistency, we decided to define the number of clusters as six for all datasets (see Figure 1). While the silhouette scores indicated different optimal sizes for some datasets, all scores were above 0.5 when using six clusters, providing sufficient evidence of well-separated clusters across the datasets. Defining six clusters allowed us to maintain a coherent structure for comparing participant profiles. From these clusters, we observe a similar underlying structure for participant profiles across different datasets, reinforcing the comparability of the results.

Altogether, we obtain six different participant profiles across the datasets (Table 3). These findings indicate that participants' chat behavior can be categorized into six different types. Profiles 5 and 6 describe the "come and go" type of chat participants who only leaves a single footprint during the whole tournament either earlier (profile 5) or later (profile 6) during an ongoing match. Then there are more active groups of participants in terms of messaging activity (profiles 1-4). Profile 4 consists of chat participants who send exactly one message per match they participate in. Profile 3 on the other hand consists of chat participants who leave at least one message in a match. Profiles 1 and 2 consist of even more active chat participants who write up to 100 messages per single match. The main difference between them is that chat participants from profile 1 always participate in at least two matches and send at least three messages. Chat participants in profile 2 mostly participate in a single match in the tournament, but when they participate in more than one match (four matches at maximum), they write the same amount of comments in each match they participate in. Profiles also exhibit distinct differences in interaction rate which means how many of the participant's messages are replies to or

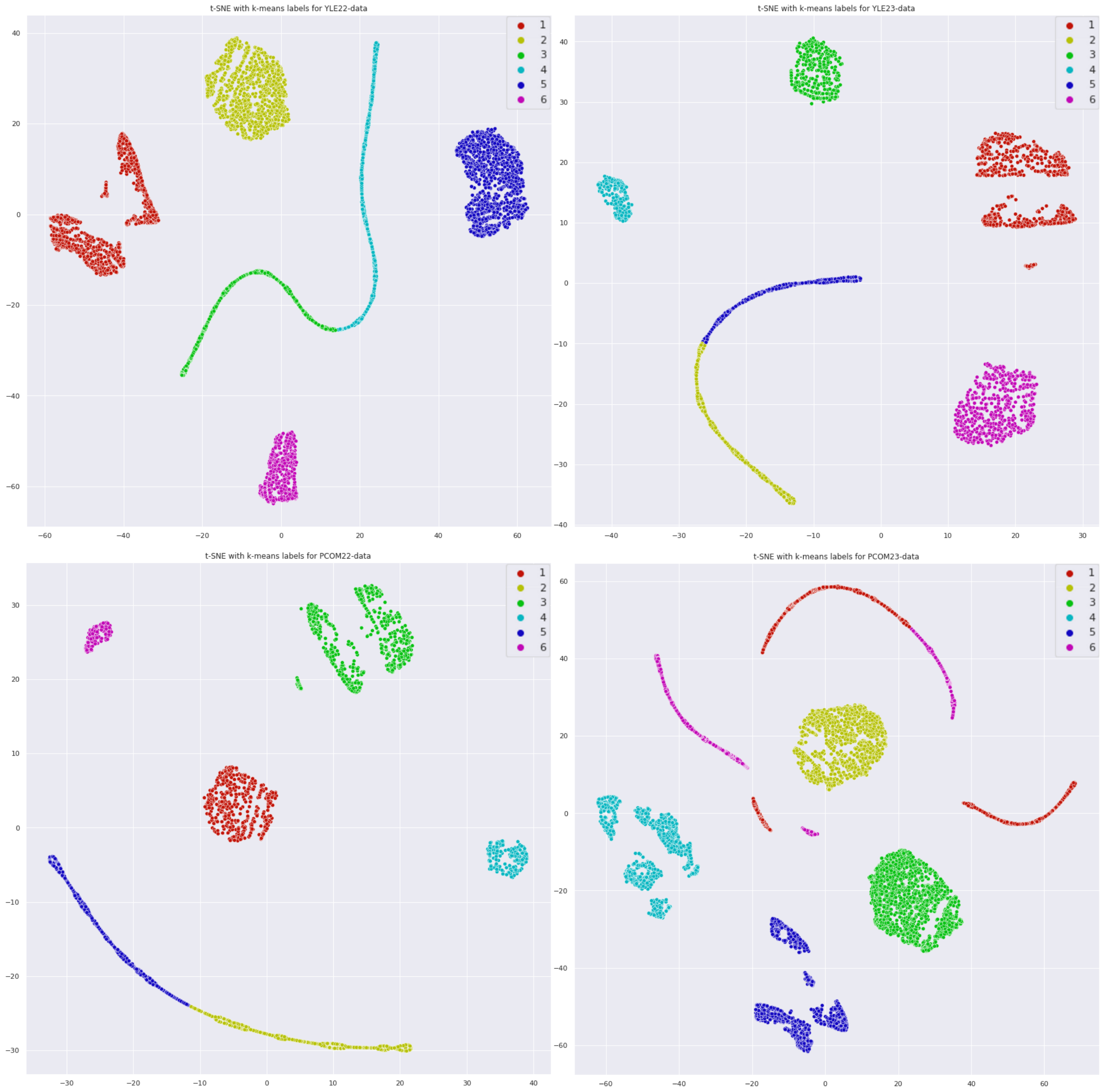

Figure 1: Activity-based clustering visualized with t-SNE and k-means labels for all of the four datasets: YLE22, YLE23, PCOM22 and PCOM23.

mention other participants. With activity-based clustering, we can differentiate between participants who have distinct attitudes to chat participation and use different amounts of effort to chat engagement.

5.2. Content-based clustering

We employed k-means clustering to identify distinct groups of participants based on four unique features within each dataset. In the cluster analysis, the number of clusters was determined using the elbow method by calculating the within-cluster-sum-of-square (WCSS) values for the uniqueness-based features, and resulted in values ranging from 8 to 10. The datasets YLE22 and PCOM22 both have 9 clusters, PCOM23 has 10 clusters and YLE23 has 8 clusters (see Figure 2).

In order to further analyze the clusters and form participant groups that span all datasets, we began by calculating the mean values across all participants for the four unique features for each identified cluster. We then used cosine similarity with a strong threshold of 0.95 to compare these mean feature values between clusters across different years and channels, identifying pairs of clusters that exhibit highly similar characteristics in terms of chat content (see Figure 3). Clusters with high similarity across

Table 3
Participant profiles by activity-based clustering

Profile	Description	Size	Key Characterics	Activity Level	Interaction Rate	Minimum Messages per Match	Maximum Messages per Match
1	Highly active participants	Large	Highest message volume. Second highest number of participants.	High	49%	3	105
2	Active participants, mostly in one match	Medium	Primarily active in one match. Some in multiple with consistent message counts per match. Distinct kidney-shaped t-SNE clusters.	Moderate – High	27%	2	132
3	Active participants, comment count 10th quantile is 1	Medium	Active participants commenting at least once in some match. Average of 3 messages per match.	Moderate	56%	1	26
4	Participants with matching activity and message counts	Small	Number of comments equals number of matches participated in. Minimum two matches/comments.	Low	6.9%	2	10
5	Single-message participants, messaging early on	Largest	Produce only one message, typically earlier in the match.	Low	6.4%	1	1
6	Single-message participants, messaging later on	Medium	Produce only one message, typically later in the match.	Low	6.9%	1	1

datasets are considered potential candidates for forming participant profiles.

Out of all the participant profiles (Table 4), there are six profiles (1-6) that are present in every dataset (i.e., there is one cluster fitting this profile in each dataset), and two participant profiles (7-8) that are present in three datasets (i.e., a cluster representing this profile is present in YLE22, YLE23 and PCOM22 datasets, but is missing a cluster from PCOM23 dataset). One participant profile (9) is also present in two datasets (YLE22 and PCOM22). This leaves out some clusters from the PCOM23 dataset that were not classified into any of the aforementioned participant profiles. The cluster pcom23_3 was classified into participant profile 10 and fairly similarly distributed clusters pcom23_0, pcom23_6, and pcom23_8 were grouped together into 11a, 11b and 11c.

Unlike other analyzed datasets, participant profiles involving PCOM23 showed a great deal of variety among participants by their message content in "commands and replies". Examining participant profiles 11a, 11b, 11c and participant profile 2, they all exhibit the majority of content focused on 'commands and replies'. Together they account for over 48% of all participants in PCOM23. Comparatively, this

Table 4
Participant profiles of content-based clustering

Profile	Description	Key Characteristics	Activity Level	Interaction Rate	Top 25% Activity (Msgs/-Tournament)	Matches Participated (Top 25%)
1	We are here to stay	Unique and essentially non-unique messages. (43% and 45% respectively) Very active in chat and mentions.	High	52%	18	4
2	Communicating with the system	Primarily engage with channel bots for giveaways. (98% of messages) Minimal social interaction.	Low	0.7%	3	1
3	We like to gg and LUL	Dominated by non-unique messages, primarily repetitive content and emotes. Limited direct social interaction.	Low	2.7%	2	2
4	Less unique members of the crowd	Primarily contain essentially non-unique messages.	Low	20%	1-2	1
5	Unique members of the crowd	Primarily contain unique messages. Largest group for PCOM22 and YLE datasets, but only third largest in PCOM23.	Low	19%	2	1
6	We keep the conversation flowing	Contain both unique and essentially non-unique messages. Most active group in terms of messages and interactions.	High	60%	20	5
7	Non original discussants	Contain both non-unique and essentially non-unique messages.	Moderate	34%	10	3
8	We like to LUL but also keep it original	Contain both unique and non-unique messages.	Moderate	35%	12	4
9	Commands are cool, but so are chat contents	Primarily contain commands and replies, with a significant portion of unique messages. Active only in the 2022 Majors.	Low to Moderate	20%	5	3
10	Jack of non giveaways	Contain a mix of non-unique, non-unique but human-written, and unique messages. Only PCOM23-specific.	High	33%	20	6
11(a,b,c)	Giveaways first, content second	Relies heavily on "commands and replies," with varying secondary content types (unique, non-unique, essentially non-unique). Only PCOM23-specific groups.	Moderate	25%, 8%, 26%	6-8	4-5

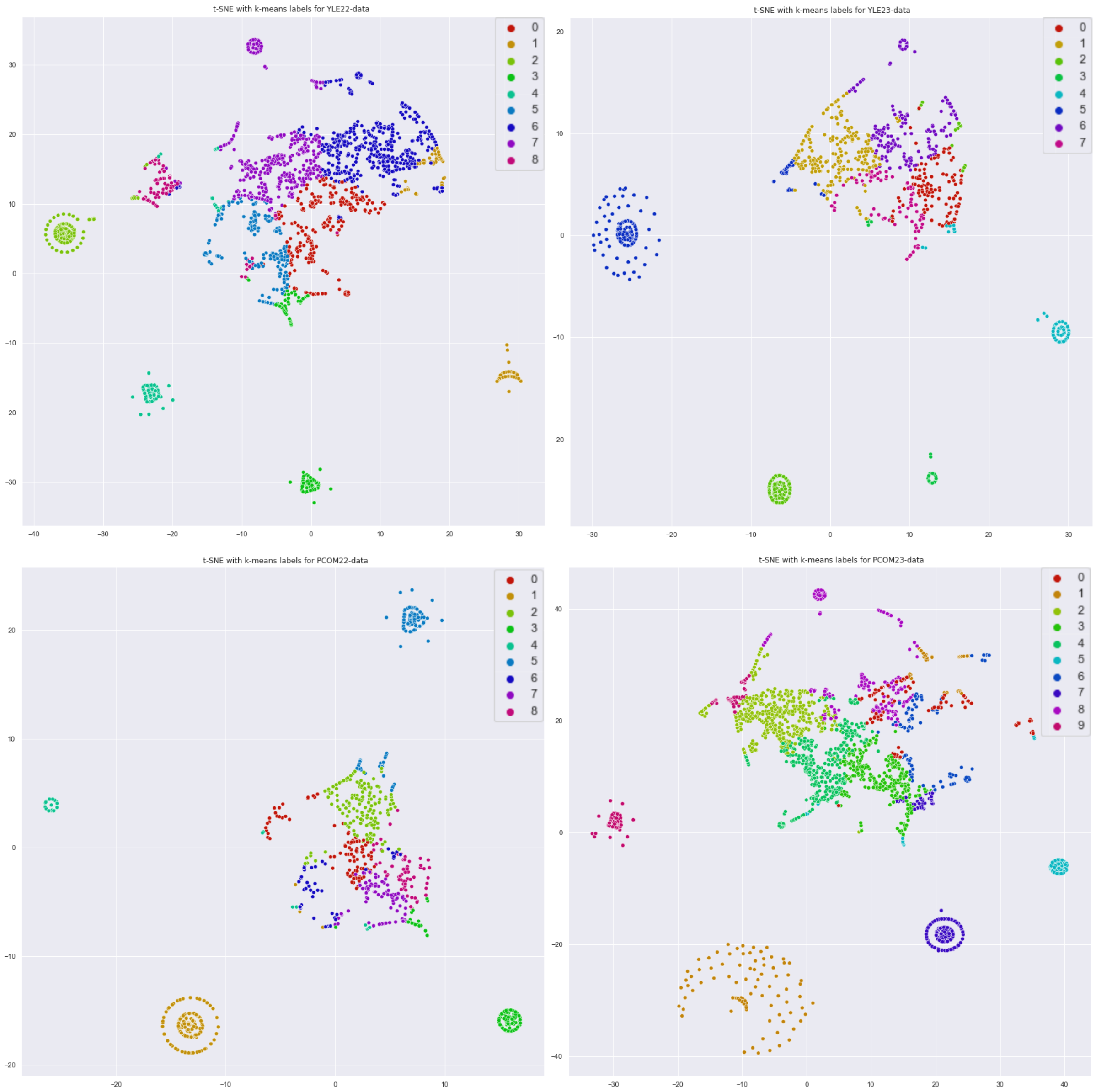

Figure 2: Content-based clustering visualized with t-SNE and k-means labels for all of the four datasets: YLE22, YLE23, PCOM22 and PCOM23.

pattern is less pronounced in other datasets: 22.7% of participants in PCOM22 (participant profiles 2 and 9), 10.5% of participants in YLE22 (same profiles), and only 5.7% of participants in YLE23 (profile 2). To understand why this behavior is so different between the channels we are next going to analyze this content type further.

While clustering may give us an idea about participant profiles appearing in chat data, chat behavior may also be described with tracking the most active moments in the timeline of the tournament (see Figure 4).

The distribution of the most active moments, overall and by content types, was visualized by percentages along a timeline of the tournament. This metric signifies the starting point of an 8-second window containing the highest concentration of activity. Most active moments overall across the channels generally concentrate towards the end of the timeline for both YLE22 and YLE23, but for PCOM22 and PCOM23 there is also some peak activity concentrated around the midle of the match. The most active moments vary between the chat content types. The non-unique messages gravitate greatly towards the end of the match, which makes sense in the form of the volume of “gg” (short for “good game”) messages

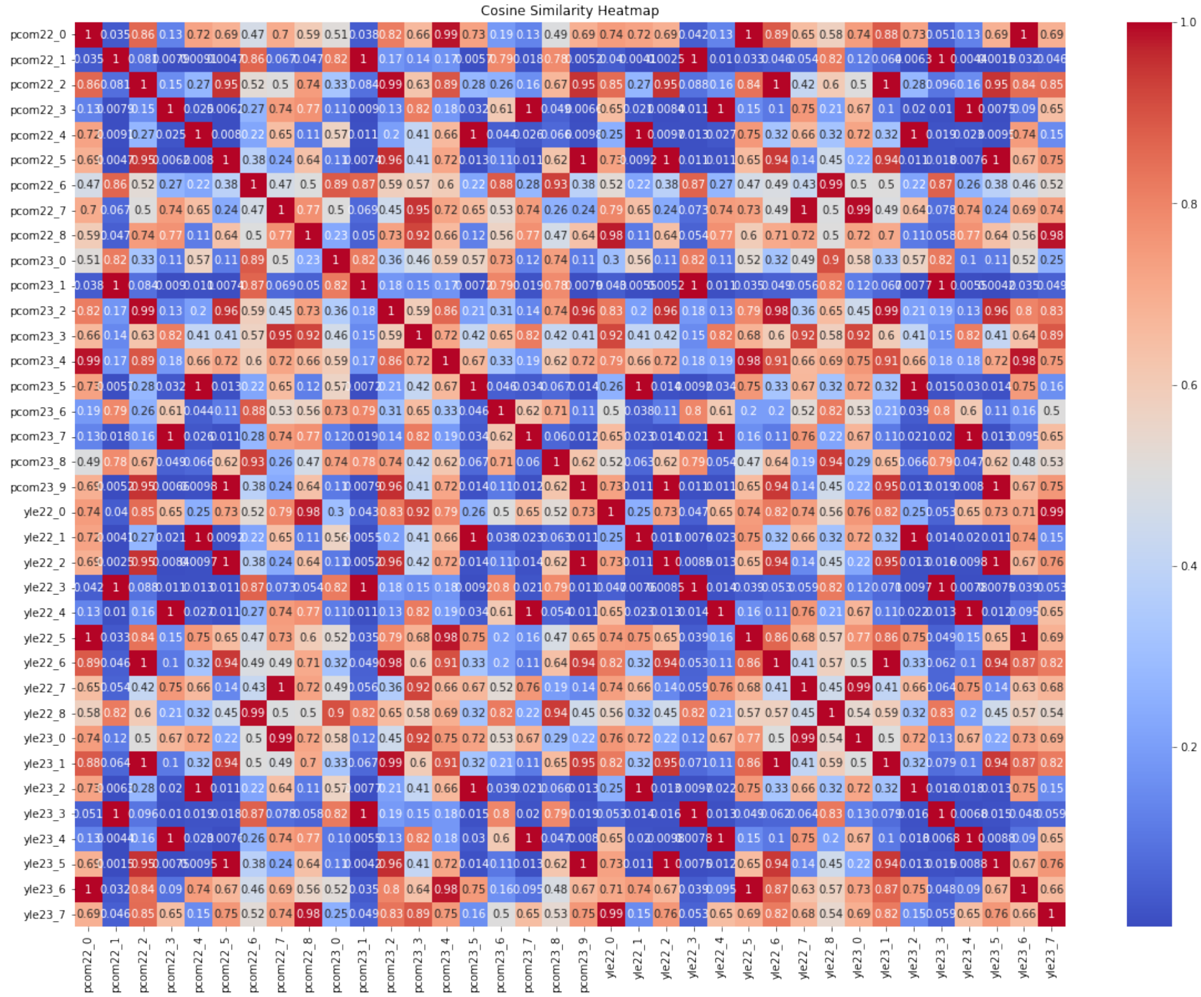

Figure 3: Correlation matrix for clusters produced with content-based clustering.

when a match ends. The highest activity for unique messages occurs slightly past the midpoint of the match for every dataset, which could be explained by discussion regarding in-game activity. For the 'human-written non-unique' messages no clear trend was found regarding the activity. This could indicate that these messages are triggered by various events on-screen rather than regular, repetitive events such as the beginning or ending of a match. Similar observations were noticed in [3]. The most active moment in 'commands and replies' reinforces the previous analysis on the differences of Pelaajat.com and Yle channels regarding interaction using these types of messages. For both of the YLE datasets, the chat activity greatly concentrated on the early timeline. This concentration could indicate that at the start of each match participants use informative commands, for example, in order to get further information about the ongoing match. For Pelaajat.com we see density throughout the whole match rather than a single spike at the start like for Yle. This indicates consistent behavior regarding the use of commands during a match.

The analysis of top commands used across channels provides insight into the distinct priorities of the participants in different livestreams (see Table 5). On the Pelaajat.com channel, a great number of commercially oriented commands dominate the stream. Commands like "!grandiosa" (popular brand of frozen pizza in Finland), "!beefmode" (partnership with a beef jerky brand), and "!gigantti" (Finnish electronics store) strongly suggest frequent giveaways and sponsor-driven promotions keeping participants engaged with the channel, especially during the 2023 tournament. In contrast, the chat participants on the Yle channel, which is lacking the commercial promotional campaigns, are primarily focused on core game details (as a public broadcast company, YLE does not run commercial advertisement

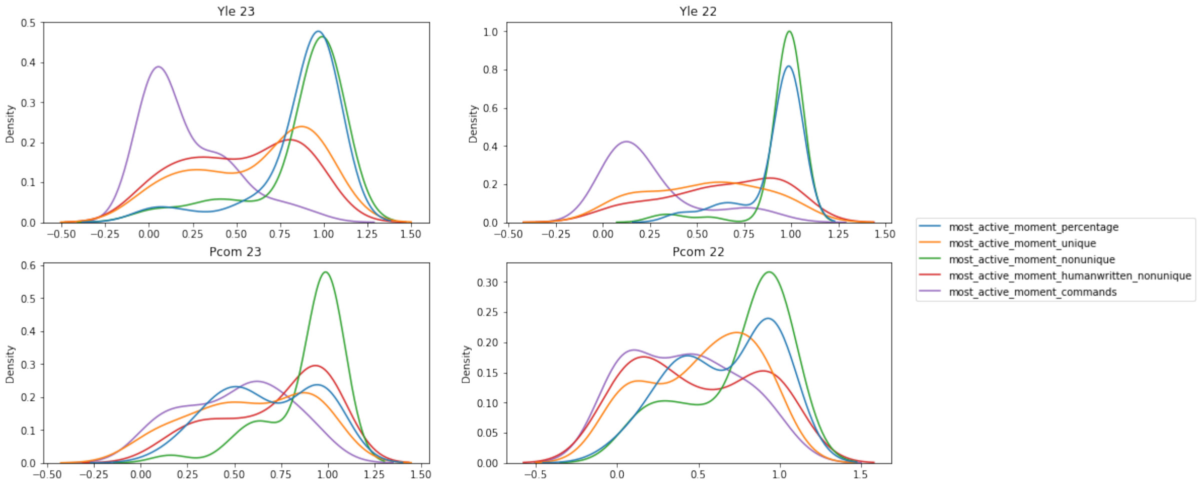

Figure 4: KDE plots visualizing the distribution of the most active moments, overall and by content types, during the tournament.

Table 5
Top 5 messages by commands, which are not bot replies, for each dataset.

	2022	#	2023	#
Yle	!maps	1050	!casters	325
	!casters	868	!maps	191
	!kartat	254	!results	31
	!ence	110	!bracket	12
	!ottelut	104	!caster	11
PCOM	!beefmode	414	!grandiosa	7849
	!arvonta	166	!beefmode	1635
	!maps	150	!gigantti	1089
	!BEEFMODE	130	!maps	418
	!kysely	81	!dick	394

campaigns). Commands such as "!maps", "!casters", and "!results" are the most common. Participants demonstrate a clear desire for additional context about the ongoing or upcoming match regarding the map and teams going against each other, and the commentators that are casting the match. The frequent use of the Finnish equivalents of these commands such as “!kartat” or “!selostajat” (Finnish for !maps or !casters) further points to an audience invested in understanding the specifics of the match in question.

Overall, these command usage patterns indicate a difference in audience’s interests and preferences on these channels.

6. Discussion and conclusions

In this paper, we aspired to use two types of clustering, activity and content-based, to classify the chat behavior of audiences on livestream platform Twitch for the two consecutive CS:GO Majors 2022 and 2023, focusing on Finnish language channels. The typical commenting behaviors were used to identify participant profiles. After carefully conducting a pair of clustering and analysis on the participant profiles they compose, we ended up with six participant profiles using activity-based clustering, and 11 participant profiles using content-based clustering. The disparity in the number of participant profiles between the two approaches suggests that when using the activity features the participant profiles are clean-cut (same profiles are visible in different channels and times), but using content features there is a lot more variety in the commenting behaviour. In content-based clustering, we first used a deep-learning based approach to classify chat messages into four distinct categories by "uniqueness" using a transformer-based chat content detection model to create content-based features for all participants. This approach made it possible for us to focus on the commenting style of participants, offering a chance to uncover deeper qualities of user-generated content, and provided insights into engagement patterns that are generalizable across different contexts.

Another observation in our analysis is that content-based clustering produced more clusters compared to activity-based clustering. This difference is likely due to the nature of the features used in each clustering approach. The features in activity-based clustering rely on message- and time related information, while content-based clustering incorporates features of the uniqueness of a message, which capture the differences in participants' communication styles in livestream chats. This leads to a wider range of participant profiles, reflected in the larger number of clusters.

Based on the t-SNE figures (see Figure 1 and 2), activity-based clustering shows more clearly defined clusters than content-based clustering, where for each dataset there are around four circle-shaped clusters lying outside of the other clusters that form a united group in the middle. These circle-shaped clusters situated mostly on the outside belong to participant profiles that are predominantly presenting only one content based feature. In activity-based clustering the participants that prefer sending a single message can be found in the string-like formations. Other clusters in activity-based clustering are distinctively far away from other clusters. Both types of clustering show us original properties of the data, complementing each other.

Moreover, we find our approach useful to understand participation in livestream chats from multiple viewpoints. Activity-based participant profiling gives information about the amount of effort the participants bring into the chatting, while content-based participant profiling describes the type of content participants create. Combining both approaches allows forming a comprehensive understanding of a particular participant's chat behavior, rather than preferring one type of clustering over another.

Besides differences between the two types of clustering, our analysis also reveals substantial disparities in the participant profiles between the Yle and Pelaajat.com channels, especially within the PCOM2023 data, based on chat content. Throughout the 2023 tournament Pelaajat.com exhibited four distinct participant profiles out of which three were only present on this channel, each characterized by a majority of 'commands and replies' within messages, consisting between 50% to 98% of their total message distribution and accounting for over 48% of all the active participants. These disparities can be explained by the frequent giveaways and sponsored promotions within the channel. This influence on participant behavior could also explain why, for instance, the cluster shapes for PCOM23 differ from other datasets in Figure 1. Therefore it seems that commercial promotions can substantially affect audience behavior in livestream chats. This information could provide insights for esports organizations aiming to enhance audience engagement. For example, knowing which clusters tend to respond to promotional content or giveaways could help businesses tailor their engagement strategies to fit specific participant profiles, leading to more effective interaction tactics and potentially improving viewer retention during these events.

Naturally, there are limitations to our work. These limitations include using data from only Finnish broadcasters. In this paper, we decided to focus mainly on a singular geo-political region, Finland, and its popular broadcasters, Yle and Pelaajat.com. Our study could be expanded by including livestream chat data from international broadcasters as well, or to compare the participant profiles created based on this data and data from international broadcasters.

The participant profiles in our analysis showed mostly 'unique' message content for the total messages of both of the tournaments combined across both channels. However, these patterns may not generalize well to the even more massive livestreams. There may very well be differences, e.g., in the type of uniqueness that permeates the chats of non-Finnish livestreams. Is there a distinct type of audience in these smaller national channels that differs from the larger "sports crowd" associated with the main international streams of the tournaments? Further research is needed to investigate potential differences in chat participants' behavior on an international scale based on their message content types.

Another limitation lies in the shortage of methods and features that could have been used. To keep the scope of our study reasonable, we only focused on the active part of the tournament and left out metadata that was originally collected from livestream chat data. In future research, the inactivity period between games as well as metadata, like information about banned participants' messages, could be included to examine chat behavior. The research we presented here could also be expanded by using several different clustering, or other unsupervised, methods and by making more technical analysis on the implications each method brings. There can also be some constraints related to the basis of the unique content detection model, because of the constantly evolving subculture of emotes and the emergence of new instances of text-based memes. For example, third-party browser based plugins like 7TV or BetterTTV allows integration of custom emotes into any word. This can add another layer of complexity to interpret the message, making it difficult to distinguish between the intentional use of a custom emote or a text-based meme. This poses an ongoing challenge for classification of these types of messages. Future research could focus on refining the chat content model to enable more enhanced classification of the proposed content types, as well. Potential examples include the direct categorization of 'non-unique' messages into distinct subcategories, such as 'emotes', 'gaming slang', and 'copypasta'. This fine-grained classification would provide deeper insights into the nature of communication patterns within chat participants' messages, especially as livestream subcultures continue to evolve.

The identification of distinct participant profiles in online livestream chat communities, as demonstrated in this study of Finnish-language Twitch users in esports streams, opens new areas for understanding these dynamic social spaces. By combining both activity-based and content-based clustering, this research offers a more broad understanding of user engagement. The resulting participant profiles not only describe audiences based on the volume of their chat activity (e.g., prolific vs. infrequent chatter) but also based on the nature of their content (e.g., unique contributors vs. repetitive posters). While we focused more on engagement patterns for broader applicability, future research might build upon these findings by examining thematic clustering within these chat messages, using approaches such as topic modeling, to explore specific conversation topics and community dynamics in more depth. While this methodological approach has only been applied here specifically into the Twitch context, it still holds promise for a more broad application in the analysis of online communities across different livestreaming platforms. It could give a different point of view to research made on chat data from other livestream platforms, like Youtube [39]. Additionally, instead of using an existing framework or theory for explaining participant behavior or communication styles and basing our profiling research on a set of assumptions [26] [17], we decided to utilize an unsupervised method (clustering) to bring out qualities that are not necessarily readily visible to the researcher but exist as a groupable substance.

Acknowledgments

This research was supported by the FIN-CLARIAH infrastructure project (The Research Council of Finland 358726).

References

[1] M. Carter, B. Egliston, The work of watching twitch: Audience labour in livestreaming and esports, Journal of Gaming & Virtual Worlds 13 (2021) 3–20.

[2] N. T. Taylor, Now you're playing with audience power: The work of watching games, Critical Studies in Media Communication 33 (2016) 293–307.

[3] D. Bulygin, I. Musabirov, A. Suvorova, K. Konstantinova, P. Okopnyi, Between an arena and a sports bar: Online chats of esports spectators, arXiv preprint arXiv:1801.02862 (2018).

[4] Y. Jiang, X. Shen, R. Wen, Z. Sha, J. Chu, Y. Liu, M. Backes, Y. Zhang, Games and beyond: Analyzing the bullet chats of esports livestreaming, in: Proceedings of the International AAAI Conference on Web and Social Media, volume 18, 2024, pp. 761–773.

[5] S. Jain, D. Niranjan, H. Lamba, N. Shah, P. Kumaraguru, Characterizing and detecting livestreaming chatbots, in: Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2019, pp. 683–690.

[6] C. Ringer, M. Nicolaou, J. Walker, Twitchchat: A dataset for exploring livestream chat, in: Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment, volume 16, 2020, pp. 259–265.

[7] C. Flores-Saviaga, J. Hammer, J. P. Flores, J. Seering, S. Reeves, S. Savage, Audience and streamer participation at scale on twitch, in: Proceedings of the 30th ACM Conference on Hypertext and Social Media, 2019, pp. 277–278.

[8] N. Merayo, R. Cotelo, R. Carratalá-Sáez, F. J. Andújar, Applying machine learning to assess emotional reactions to video game content streamed on spanish twitch channels, Computer Speech & Language 88 (2024) 101651.

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

[10] S. Yousukkee, N. Wisitpongphan, Analysis of spammers' behavior on a live streaming chat, IAES International Journal of Artificial Intelligence 10 (2021) 139.

[11] B. Janet, A. Nikam, et al., Real time malicious url detection on twitch using machine learning, in: 2022 International Conference on Electronics and Renewable Systems (ICEARS), IEEE, 2022, pp. 1185–1189.

[12] D. Melhart, D. Gravina, G. N. Yannakakis, Moment-to-moment engagement prediction through the eyes of the observer: Pubg streaming on twitch, in: Proceedings of the 15th International Conference on the Foundations of Digital Games, 2020, pp. 1–10.

[13] J. Kim, K. Park, H. Song, J. Y. Park, M. Cha, Learning how spectator reactions affect popularity on twitch, in: 2020 IEEE International Conference on Big Data and Smart Computing (BigComp), IEEE, 2020, pp. 147–154.

[14] D. Gros, B. Wanner, A. Hackenholt, P. Zawadzki, K. Knautz, World of streaming. motivation and gratification on twitch, in: Social Computing and Social Media. Human Behavior: 9th International Conference, SCSM 2017, Held as Part of HCI International 2017, Vancouver, BC, Canada, July 9-14, 2017, Proceedings, Part I 9, Springer, 2017, pp. 44–57.

[15] M. Hu, M. Zhang, Y. Wang, Why do audiences choose to keep watching on live video streaming platforms? an explanation of dual identification framework, Computers in human behavior 75 (2017) 594–606.

[16] Z. Hilvert-Bruce, J. T. Neill, M. Sjöblom, J. Hamari, Social motivations of live-streaming viewer engagement on twitch, Computers in Human Behavior 84 (2018) 58–67.

[17] W. B. Yu, J. Robb, P. Mirza-Babaei, Profiling livestream spectators, in: Extended Abstracts of the 2020 Annual Symposium on Computer-Human Interaction in Play, 2020, pp. 403–407.

[18] J. S. Lim, M.-J. Choe, J. Zhang, G.-Y. Noh, The role of wishful identification, emotional engagement, and parasocial relationships in repeated viewing of live-streaming games: A social cognitive theory perspective, Computers in Human Behavior 108 (2020) 106327.

[19] F. Neus, F. Nimmermann, K. Wagner, H. Schramm-Klein, Differences and similarities in motivation for offline and online esports event consumption, Proceedings of the 52nd Hawaii International Conference on System Sciences (2019).

[20] W. B. Yu, N. A. Beres, R. B. Robinson, M. Klarkowski, P. Mirza-Babaei, Exploring esports spectator motivations, in: CHI Conference on Human Factors in Computing Systems Extended Abstracts, 2022, pp. 1–6.

[21] P. Schuck, M. Altmeyer, A. Krüger, P. Lessel, Viewer types in game live streams: questionnaire development and validation, User Modeling and User-Adapted Interaction 32 (2022) 417–467.

[22] S. R. Kairam, M. C. Mercado, S. A. Sumner, A social-ecological approach to modeling sense of virtual community (sovc) in livestreaming communities, Proceedings of the ACM on human-computer interaction 6 (2022) 1–35.

[23] T. Tang, J. Kucek, S. Toepfer, Active within structures: Predictors of esports gameplay and spectatorship, Communication & Sport 10 (2022) 195–215.

[24] Y. Li, C. Wang, J. Liu, A systematic review of literature on user behavior in video game live streaming, International journal of environmental research and public health 17 (2020) 3328.

[25] F. Cauteruccio, Y. Kou, Investigating the emotional experiences in esports spectatorship: The case of league of legends, Information Processing & Management 60 (2023) 103516.

[26] V. Diwanji, A. Reed, A. Ferchaud, J. Seibert, V. Weinbrecht, N. Sellers, Don't just watch, join in: Exploring information behavior and copresence on twitch, Computers in Human Behavior 105 (2020) 106221.

[27] C. Ford, D. Gardner, L. E. Horgan, C. Liu, A. Tsaasan, B. Nardi, J. Rickman, Chat speed op pogchamp: Practices of coherence in massive twitch chat, in: Proceedings of the 2017 CHI conference extended abstracts on human factors in computing systems, 2017, pp. 858–871.

[28] A. Nematzadeh, G. L. Ciampaglia, Y.-Y. Ahn, A. Flammini, Information overload in group communication: From conversation to cacophony in the twitch chat, Royal Society open science 6 (2019) 191412.

[29] J. Seering, J. Hammer, G. Kaufman, D. Yang, Proximate social factors in first-time contribution to online communities, in: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 2020, pp. 1–14.

[30] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, Advances in neural information processing systems 26 (2013).

[31] K. Konstantinova, D. Bulygin, P. Okopny, I. Musabirov, Online communication of esports viewers: Topic modeling approach, in: Advances in Computer Entertainment Technology: 14th International Conference, ACE 2017, London, UK, December 14-16, 2017, Proceedings 14, Springer, 2018, pp. 608–613.

[32] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.

[33] T. M. Alenezi, T. H. Sulaiman, A. M. AbdelAziz, Applying machine learning models to electronic health records for chronic disease diagnosis in kuwait., International Journal of Advanced Computer Science & Applications 14 (2023).

[34] P. J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, Journal of computational and applied mathematics 20 (1987) 53–65.

[35] L. Van der Maaten, G. Hinton, Visualizing data using t-sne., Journal of machine learning research 9 (2008).

[36] J. Lindroos, J. Peltonen, T. Välisalo, R. Koskimaa, I. Toivanen, From PogChamps to insights: Detecting original content in twitch chat, Proceedings of the 58th Hawaii International Conference on System Sciences (2025).

[37] P. Bholowalia, A. Kumar, Ebk-means: A clustering technique based on elbow method and k-means in wsn, International Journal of Computer Applications 105 (2014).

[38] D. T. Larose, Data mining and predictive analytics, John Wiley & Sons, 2015.

[39] C. Liebeskind, S. Liebeskind, S. Yechezkely, An analysis of interaction and engagement in youtube live streaming chat, in: 2021 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/IOP/SCI), IEEE, 2021, pp. 272–279.