Formulating Video Watch Success Signals for Recommendations on Short Video Platforms

Srijan Saket[1], Sai Baba Reddy Velugoti[1] and Rishabh Mehrotra[1]

[1]*ShareChat, Bengaluru, India*

Abstract

With the rising prominence of short video platforms, the challenge of effective content recommendation has become more pressing, especially given the diverse range of video content and the sparsity of implicit user feedback signals. This research delves into the formulation of "successful video watches", particularly for short video platforms. We introduce various functional formulations to model video watch behavior and use these to define ranking objectives for training recommender models. Our findings reveal that, in contrast to the naive percentage-based thresholds, our proposed formulation — grounded on duration and watch percentile — aligns better with user retention and boosts engagement metrics. Moreover, while the standard approach tends to bias content recommendations towards extremes in video length, our methodology ensures a more balanced content recommendation, greatly impacting user experience on streaming platforms. This study underscores the potential nuances and implications of training recommender systems for video content.

1. Introduction

In recent years, video platforms have witnessed an unprecedented surge in popularity, transforming the way users consume and engage with digital content. These platforms offer a vast and diverse range of videos, encompassing varying lengths, genres, and categories, which pose unique challenges for effective content recommendation. To successfully train content recommender systems, platform designers have relied on implicit signals in the form of user feedback data. However, in the context of short video platforms, leveraging implicit signals, such as likes, shares, or downloads, remains a challenge due to the inherent sparsity of such signals.

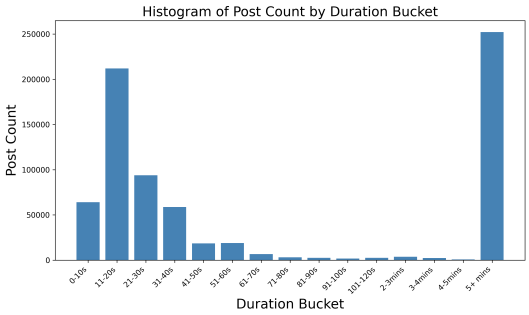
Video streaming time, on the other hand, is a widely available signal and is often leveraged to train and evaluate recommender systems. Given the heterogeneity of the video content, with a large number of short and long videos uploaded hourly on such platforms, naively choosing a label derived from video streaming time inadvertently causes bias towards certain type of video content. For example, a label based on successfully watching (say) 50% video will result in a larger proportion of shorter video watches being tagged as successful watches, as compared to longer videos. Indeed, shorter videos tend to have a larger watch perentage; e.g. users often would watch 10 seconds of a 20 second video, than watch 200 seconds of a 400 second video. Understanding the factors that define a successful video play, keeping in mind the video duration, is vital for developing effective recommendation strategies, improving user engagement, and optimizing content delivery.

In this paper, we present an in-depth investigation into different formulations of *successful video watches*, and propose various functional formulations that help us model video watch behavior on short video platforms. Subsequently, we use these formulations to define the ranking objectives and train candidate generation and ranker models. Specifically, we train Field Aware Factorization Machine model based on these objectives, and investigate how different formulations of successful video watches impact various user engagement and business metrics.

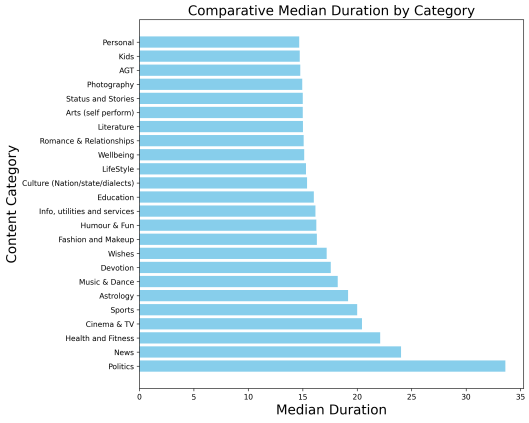
Compared with naive formulation of percentage based threshold, the proposed formulation based on duration and watch percentile is better correlated with user retention on the platform, and also results in better user engagement metrics when used as an objective for the FFM model. We also investigate how the platform level content distribution changes when these formulations are used to train the recommender system, and highlight that naive formulation of successful video watch biases the surfaced content towards the extremes, either on very short or very long videos; whereas the proposed formulation strives a better balance in terms of the video content surfaced, , significantly influencing the recommendations surfaced to users. We contend that our findings have implications on how recommender systems are trained and evaluated on video streaming platforms.

Workshop on Learning and Evaluating Recommendations with Impressions (LERI) @ RecSys 2023, September 18-22 2023, Singapore

srijanskt@gmail.com (S. Saket); saibabavelugoti@sharechat.co (S. B. R. Velugoti); erishabh@gmail.com (R. Mehrotra)
https://www.linkedin.com/in/srijansaket/ (S. Saket); https://rishabhmehrotra.com/ (R. Mehrotra)
0009-0006-9460-1203 (S. Saket); 0000-0002-0836-4605 (R. Mehrotra)



(a) Histogram plot of videos with duration bins on the x-axis



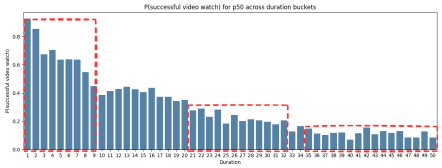
(b) Median duration of videos across categories

Figure 1: Figure illustrating the distribution of content on the platform across duration and categories. In (a), the chart displays the quantity of videos distributed among different duration bins. While there is a noticeable prevalence of shorter videos, a substantial portion comprises longer videos. In (b), the graph demonstrates the duration range of videos categorized by subject. Notably, categories such as Politics and News exhibit extended video durations compared to other categories.

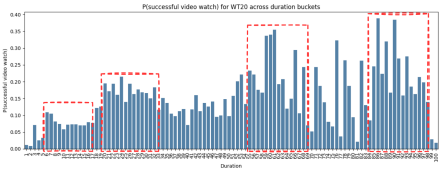
2. Related Work

Recommender systems modulate what billions of people are exposed to on a daily basis. Over the past decade a lot of research has gone into specifying how these systems are optimized for user engagement signals such as clicks, streams, likes, or a weighted combination of such sets [1]. There has been a growing interest in developing recommenders that optimize for objectives beyond accuracy such as diversity [2], novelty [3], sustainability [4], aiming at satisfying users' diverse needs.

Recent work has started to explore the impact the objective choice would have on the platform. Zahra *et al.* [5] use podcast recommendations example with two engagement signals: Subscription vs. Plays to show that the



(a) Percentage based definition (p50) of SVW



(b) Threshold based definition(WT20) of SVW

Figure 2: Figure depicting the impact of label selection on the determination of successful video watch. In (a) we display an instance of success defined based on a percentage, with a video watch surpassing 50% . Conversely, (b) shows instances of the conventional threshold selection is shown, where a watch time exceeding 20 seconds is considered.

choice of user engagement matter, and that optimizing for streams can bias the recommendations towards certain podcast types, undermine users' aspirational interests and put some show categories at disadvantage.

In the domain of short video recommendations, duration bias has been an under explored topic, especially for early stage content [6]. Wu *et al.* [7] investigated the bias of watch time and watch percentage from an aggregated level, i.e. the average of the watch time of all users towards each video. In other words, it merges all samples of the same video into one single data point, and compares with other videos to measure the video quality. Zheng *et al.* [8] propose an unbiased evaluation metric Watch Time Gain (WTG), which measures a user's relative engagement on a video against the average engagement of all users on videos with the same duration-level. Finally, Zhang *et al.* [9] propose a debiasied multiple semantics-extracting labeling framework constructs labels that encompass various semantics by utilizing quantiles derived from the distribution of watch time, prioritizing relative order rather than absolute label values.

3. Formulating Successful Video Watches as Objective

Modeling video watch behavior on video platforms presents several nuances that warrant careful consideration. Signals such as likes, shares, favorites, and clicks, which are commonly used in recommender systems, ex-

hibit sparsity in their occurrence, making them unreliable for capturing user preferences. In contrast, video watches is relatively less sparse compared to other signals, providing a more abundant source of information for modeling [Table:1]. However, the challenge lies in appropriately formulating the video watch signal. Indeed, a biased label choice might lead to the unintended promotion of shorter videos at the expense of longer-format videos, potentially cannibalizing the latter's visibility and engagement. Striking the right balance in label choice becomes essential for developing accurate and fair recommendation models in short video platforms.

We begin by describing the product context and production data used in Section 3.1 and present preliminary analysis that highlight the biases that might exist with naive formulations of video watch signal (Section 3.2). In Section 3.3 we propose a number of alternate formulations of the successful video watch signal; which we use to train the recommendation model described in Section 3.4.

3.1. Production Data Context

We consider production traffic from one the largest short video applications serving 200 million users, and randomly sampled user interaction data over the course of one week across 41,316,850 users, containing 14,549,333 video posts in the Hindi language, capturing both implicit signals such as Video Play - indicating a successful completion of a recommended video beyond a specified threshold, Skip, click, like, share, and favorite. We leverage Field-aware Factorization Machines (FFM) to extract 32-dimensional embeddings from user-item interactions for each signal. To ensure real-time learning, we adopted a dynamic approach that continuously updated the embeddings with every new interaction data point. The learning process utilized 7 days of logged data, while offline evaluation was conducted on 1 day of unseen data.

3.2. Prevalence of Duration bias

As baseline formulation of video watch signal, we propose two simple approaches:

Fixed Threshold (WT20) defines a binary label for successful video watch based on whether the user watched the video for more than 20 seconds.

Fixed Percentage (p50) defines a binary label for successful video watch based on whether the user watched the video for more than 50% of the video duration.

Considering both these definitions, we plot the proportion of videos that get a successful watch label of 1 using the above mentioned labels, across various video duration in Fig.2. It's evident that the definition of success is influenced by the video's duration: when using a percentage-based definition (p50)[Fig.2a], shorter videos are considerably favored over longer ones. Conversely, with a fixed threshold-based definition (WT20)[Fig.2b], longer videos are preferred over shorter ones. We anticipate that an ideal label should not exhibit such bias towards video duration, given that such a bias would significantly alter the overall content consumption pattern on the platform.

This also highlights the need for more nuanced formulations of video watch signal, which we consider in the next section.

Table 1
Signal sparsity in comparison to the video play signal as baseline. For instance, the *share* signal exhibits approximately 3% positive instances in comparison to the positive instances found in the *video play* signal, posing a greater challenge for modeling.

Signal Type	Relative % of positives w.r.t. video play
like	10.87%
share	3.10%
video play	100.00%
favourite	8.41%
comment	0.04%
skip	140.22%

3.3. Proposed Formulations of Video Watch Signal

As highlighted in Figure 2, naive formulation of video watch signals often causes duration bias in the video content surfaced as recommendations. To mitigate this, we introduce additional formulations of video watch signal that can serve as labels for training candidate generation or ranking models. These formulations can be categorized into two main categories: binary and continuous signals, each offering distinct approaches for defining them. For continuous signals, we further explore different definitions based on watch percentages, watch time, and percentile watches.

Table 2 provides details of the proposed formulations. The "*WT20*" signal employs a fixed time threshold of 20 seconds, serving as a quick engagement measure. In contrast, the "*L1PD*" signal is based on the logarithm of the video duration and dynamically adjusts the threshold based on the ratio of watch time to video duration, adapting to varying content lengths. The "p50" signal sets the threshold at 50% of video duration, wherein we assume if a user streams atleast half the video, it is a successful video watch.

The "*SVP*" signal introduces a nuanced binary metric of successful watch, stratifying videos into duration-based

Type	Label Name	Description
Binary	WT20	$\text{WT20} = \begin{cases} 1 & \text{if watch\_time} > 20\text{s} \\ 0 & \text{else} \end{cases}$
	L1PD	$\text{L1PD} = \begin{cases} 1 & \text{if watch\_time} > \log(1 + \text{duration}) \\ 0 & \text{else} \end{cases}$
	SVP	$\text{SVP} = \begin{cases} 1 & \text{if watch time} > \frac{(duration-min)\mu}{max} \text{\& duration} \in [5, 19) \\ 1 & \text{if watch time} > \frac{(duration-min)\mu}{max} \text{\& duration} \in [20, 34) \\ 1 & \text{if watch time} > \frac{(duration-min)\mu}{max} \text{\& duration} \in [35, 49) \\ 1 & \text{if watch time} > \frac{(duration-min)\mu}{max} \text{\& duration} \in [50, 63) \\ 1 & \text{if watch time} > \frac{(duration-min)\mu}{max} \text{\& duration} > 64 \\ 0 & \text{otherwise} \end{cases}$
	P50	$\text{p50} = \begin{cases} 1 & \text{if watch time} > 0.5 \times \text{video duration} \\ 0 & \text{else} \end{cases}$
Continuous	RootLogTimeWatch (RLTW)	$\sqrt{\log(1 + \text{watch\_time})}$
	RootLogPercentileWatch (RLPW)	$\sqrt{\log(1 + \text{watch\_percentile})}$
	RootLogTimePercentileWatch (RLTPW)	$\sqrt{\log(1 + \text{watch\_time} * \text{watch\_percentile})}$

Table 2
Proposed formulations of the video watch signal.

bins. Each bin's success definition is established from historical user watch behavior, allowing for targeted evaluation of engagement within specific temporal segments. Referring to Table.2, within a particular bin, a value of 1 is assigned if the video watch time surpasses the specified threshold; otherwise, a value of 0 is assigned. The threshold is decided based upon the min, max and mean of watch time within the duration bucket, from user historical watch time data.

For the group of continuous signals, we leverage logarithmic and square root functions, given their ability to gracefully handle a range of video duration and scale down the scores from a wide array of duration range. The "*RLTW*" metric is centered on raw watch time, and is quantified by taking the square root of the log of watch time.

Additionally, "*RLPW*" incorporates percentile of video watch, linking video plays to the percentile distribution of their respective bins, enabling context-aware assessment. We generate uniform 1-second intervals for videos of varying lengths. Then we calculate the percentile distribution of watch time within each interval. To get the value of the label, first, associate the video play event with its corresponding duration interval. Subsequently, within that interval, correlate the observed watch time with the respective percentile value. Finally, we propose "*RLTPW*", which leverages the interplay of watch time and watch time percentile, yielding a composite signal reflecting both engagement magnitude and relative positioning.

Together, these suggested indicators offer a wide range of choices, offering binary and continuous objectives that can be used to define labels to train recommender models. We present details of one such model in the next section.

3.4. Training recommenders based on Video Play Signal

To evaluate the efficacy of the proposed formulations of the video watch signal, we train a Field-Aware Factorization Machines (FFM) model takes userId, videoId and label as the training data input and learns a 32-dimensional vector based representation that captures interactions between user and video features. FFM extends the traditional Factorization Machines by introducing field information, which is crucial in recommendation systems where attributes can belong to distinct categories or domains. The FFM formula can be expressed as follows:

$$\hat{y}(x) = w_0 + \sum_{i=1}^{n} w_i x_i + \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \langle v_i, v_j \rangle \cdot x_i x_j$$

where x represents the input features, w_0 is the bias term, w_i are linear weights, v_i are the feature embeddings, and $\langle v_i, v_j \rangle$ denotes the inner product between the embeddings of features i and j.

Once trained, the FFM model provides us with user and video embeddings which we store in vector databases and leverage approximate neareast neighbor search approaches to fetch a recommendation of top-k closest videos to a given user embedding. These top-k fetched videos consitute the recommendations shown to the user.

Specifically around training, there are two modes of learning these embeddings via FFM models: batch and real-time. The batch setup updates the embeddings at a

specified frequency by collecting data over a period of few hours whereas the real-time method dynamically updates them based on every interaction data point. In this work, we designed a real-time embedding update system wherein FFM continually refines embeddings, enabling the model to adapt to evolving user preferences and item characteristics, enhancing the accuracy and relevance of generated recommendations. We omit the specific implementation details of the realtime embedding system since it is outside the scope of the current work, but we posit that the findings presented in this paper should generalize beyond FFM models.

4. Evaluation

We next present a detailed evaluation on how the proposed formulations of the video watch signal performs on a number of evaluation criterion. We begin by looking at the correlation of these signals with user retention on the platform (Section 4.1) and present an analysis on the overlap of information between these signals and traditional like/share/favorite signals (Section 4.2). Further, we use these labels to train a FFM based recommender models and evaluate its performance on recall based user engagement metrics (Section 4.3). Finally, we demonstrate the impact the design choice of this signal could have the overall platform level content distribution (Section 4.4).

4.1. Correlation of Labels with User Retention and Other Signals

4.1.1. How did we chose the labels:

An optimal label for training any machine learning model is one that demonstrates a correlation with user retention. Hence, gauging this correlation is of paramount significance. UserIDs are linked to video watch events, leading to the formation of labels, and subsequently, their correlation with retention is measured.

4.1.2. Key takeaway:

The outcomes of this analysis highlight that the proposed employment of quantile-based labels with square root transformations exhibits a stronger correlation with user retention when compared to the previous SVP signal and threshold-based approaches. This correlation is even higher than that achieved by the L1PW approach.

4.1.3. Remarks on functions:

It's observed that logarithmic transformations outperform threshold and percentage-based methods. The rationale behind this is yet to be determined. Additionally, combining square root with a logarithmic transformation results in further improvement, though the reason behind this improvement is still under investigation.

Table 3
Label/ROC_AUC_SCORE

Label	Like	Share	Fav	Vskip	Vclick
RLTW	0.5778	0.5897	0.5694	0.2618	0.5682
RLPW	0.5745	0.5909	0.5716	0.2701	0.5825
RLTPW	0.5803	0.597	0.5749	0.2549	0.5788
L1PD	0.5853	0.6063	0.5933	0.2504	0.5793
SVP	0.5487	0.5517	0.5286	0.2949	0.5451
WPER	0.5703	0.5866	0.5698	0.2605	0.5516

4.1.4. Remarks on signals:

Among percentile, watch time, and duration, percentile appears to be the most influential, followed by watch time and then duration. This hierarchy is attributed to the fact that duration-based measurements lack broad applicability across various data subsets due to their threshold nature.

4.1.5. Comparison between D1, D3, and D7 retention:

As we progress from D1 to D7, there is a slightly larger correlation value (0.12 vs. 0.09), indicating a higher correlation with long-term retention compared to short-term retention. Notably, no significant variance is observed across trends.

In conclusion, we anticipate that training a machine learning model with this objective should enhance user retention, though it's important to note that this is a view of correlation rather than causation.

4.2. Overlap between signals

Our objective is to assess whether the suggested labels convey similar information or exhibit varying degrees of overlap. Upon examination, it becomes evident that these labels differ and convey distinct information, as indicated by the heatmap[Fig.4]. Consequently, this dissimilarity implies that the recommendations derived from them should also differ. Once we establish the absence of label redundancy, we proceed to examine how these labels influence user satisfaction and the distribution of content in later sections.

4.3. User Engagement Evaluation

In order to determine the effectiveness of the proposed labels as potential training signals for ranker models aimed

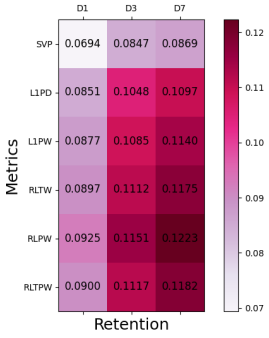


Figure 3: Retention correlation of proposed metrics with user retention

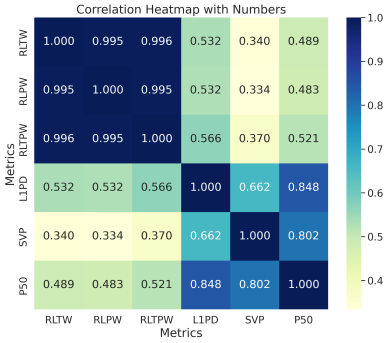


Figure 4: Heatmap showing the correlation between the proposed labels

Table 4
Recall Values for Different Labels

Label	Recall_Likes	Recall_Shares	Recall_Favs	Recall_Vskip	Recall_Vclick
SVP	0.3825	0.5057	0.4590	0.3874	0.5230
L1PD	0.4050	0.5487	0.5096	0.3674	0.5478
P50	0.3937	0.5293	0.4887	0.3764	0.5306
RLTW	0.4102	0.5506	0.5061	0.3726	0.5501
RLPW	0.4097	0.5535	0.5112	0.3767	0.5573
RLTPW	0.4125	0.5557	0.5112	0.3702	0.5536

at enhancing user satisfaction, a comprehensive evaluation framework was established. The core objective was to ascertain the suitability of the labels in capturing user engagement and subsequently training a Field-Aware Factorization Machine (FFM) model for improved ranking.

4.3.1. Methodology

To assess the viability of the labels, a seven-day dataset was utilized for training the FFM model. Subsequently, the trained model was employed to generate embeddings for the candidate items, which were then used to rank the items. The evalution was done on one day of unseen user data. The process was anchored in the analysis of interaction signals, which are pivotal indicators of user engagement.

4.3.2. Evaluation Metrics

The Receiver Operating Characteristic Area Under the Curve (ROC-AUC) was employed as the principal evaluation metric. In this context, the predicted scores were generated using the embeddings learned from the specific labels, and the actual labels were represented by the 'like' label in the corresponding column. The other assessment criterion selected is recall. Using the learnt embeddings, suggestions are formulated in accordance with suggested markers. Subsequently, the recall is calculated concerning the 'like' label as well as various other indicators of user engagement on unseen data.

4.3.3. Results and Insights

The evaluation results offer valuable insights into the effectiveness of the proposed labels in capturing user engagement and guiding the ranker model. Notably, the AUC scores varied across the different labels, shedding light on their relative performance.

The key observations from the evaluation is that the AUC score was highest for the "L1PD" label, indicating its effectiveness in capturing and predicting user engagement. The AUC scores followed a distinct order where "L1PD" was followed by "percentile based signals," "watch_time based" and "SVP" labels, respectively. The comparison highlights the varying degrees of effectiveness in capturing user engagement among the evaluated labels. From Table:4, it's evident that *RLTPW* demonstrates the highest recall concerning user engagements, along with exhibiting elevated values in relation to the newly proposed metrics. This alignment is consistent with the correlation pattern we previously identified.

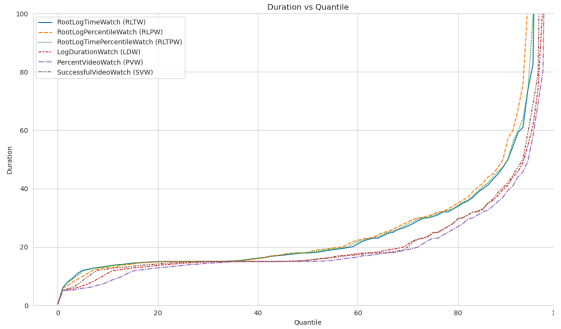


Figure 5: The figure illustrates the shift in video duration among suggested content, depicting changes across proposed labels using both percentiles and average values

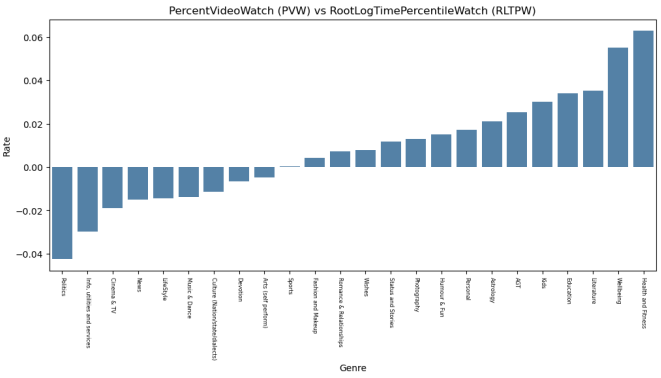


Figure 6: The figure illustrates the change in content distribution between the potentially optimal label (RLTPW) and a baseline label (p50)

4.4. Content Distribution

Our approach involved using labels to steer the model's behavior, subsequently influencing the content displayed to users. As a result, this choice of labels directly affects the content distribution across the platform. However, these interactions can lead to unintended and interesting outcomes. Furthermore, this interplay generates a feedback loop, amplifying potential negative effects. In the following sections, we demonstrate the alteration in content distribution across two aspects: duration of video and categories.

4.4.1. Across duration

From our analysis, we derive several insights shown in [Fig.5]. When employing a threshold-based label, shorter videos are favored, while the same preference holds for the percentage-based label. We observed a tendency towards longer videos when considering logarithmic watch times, although additional verification is necessary. In light of these findings, our proposed approach aims to strike a balanced combination, taking into account multiple factors for optimal content distribution.

4.4.2. Across categories

We validate our findings from the offline experiments, which demonstrate that various suggested labels present diverse content types to users. In [Fig.6], we depict the relative variation in content exposure across categories between the potentially optimal label (RLTPW) and a reference label (p50). Positive values indicate a higher recommendation frequency by the reference label (p50) compared to the optimal label (RLTPW), and vice versa for negative values. It's noteworthy that categories such as Politics and News exhibit increased recommendations through the proposed label, aligning with our initial assumptions. This observation reinforces our anticipation that the proposed label would exhibit less bias towards content duration.

5. Discussion & Conclusion

We posit that defining and understanding "successful video plays" in the context of video platforms that surface a diverse range of videos is a complex endeavor. Our investigations underscore the pitfalls of relying on naive label definitions, and higlights that such approaches suffer from duration bias and distort the content distribution on the platform in unintended ways, thereby promoting either excessive short or excessive long videos. We highlight that the proposed signals based on watch time, and watch percentile are more aligned with user retention, and when these labels were incorporated as recommender objectives, we observed a positive impact on various engagements metrics, attesting to their potential in enhancing the recommendation quality and user satisfaction.

As we move forward, several areas warrant further exploration. First, a pressing question remains about what the optimal metrics for gauging video success should be. While we have made headway in establishing some promising objectives, a comprehensive evaluation of how these formulations fare as evaluation metrics remains to be explored. Second, while our study introduced various functional formulations, we imagine future work to explore learnt formulations of video successful watch signals. Third, to truly gauge the applicability and effectiveness of our findings, they need to be validated in real-world, online scenarios wherein the recommender models trained on these signals are deployed online and evaluated. Lastly, an underexplored dimension of this

research pertains to content creators. The ways in which these formulations and definitions affect the creators, both in terms of their motivation and the content they produce, is a crucial aspect to understand.

References

[1] L. Hong, M. Lalmas, Tutorial on online user engagement: Metrics and optimization, in: Companion Proceedings of The 2019 World Wide Web Conference, 2019, pp. 1303–1305.

[2] C. Hansen, R. Mehrotra, C. Hansen, B. Brost, L. Maystre, M. Lalmas, Shifting consumption towards diverse content on music streaming platforms, in: Proceedings of the 14th ACM international conference on web search and data mining, 2021, pp. 238–246.

[3] C. H. Teo, H. Nassif, D. Hill, S. Srinivasan, M. Goodman, V. Mohan, S. Vishwanathan, Adaptive, personalized diversity for visual discovery, in: Proceedings of the 10th ACM conference on recommender systems, 2016, pp. 35–38.

[4] S. Tomkins, S. Isley, B. London, L. Getoor, Sustainability at scale: towards bridging the intention-behavior gap with sustainable recommendations, in: Proceedings of the 12th ACM conference on recommender systems, 2018, pp. 214–218.

[5] Z. Nazari, P. Chandar, G. Fazelnia, C. M. Edwards, B. Carterette, M. Lalmas, Choice of implicit signal matters: Accounting for user aspirations in podcast recommendations, in: Proceedings of the ACM Web Conference 2022, 2022, pp. 2433–2441.

[6] M. Agarwal, S. Saket, R. Mehrotra, Memer-multimodal encoder for multi-signal early-stage recommendations, in: Companion Proceedings of the ACM Web Conference 2023, 2023, pp. 773–777.

[7] S. Wu, M.-A. Rizoiu, L. Xie, Beyond views: Measuring and predicting engagement in online videos, in: Proceedings of the International AAAI Conference on Web and Social Media, volume 12, 2018.

[8] Y. Zheng, C. Gao, J. Ding, L. Yi, D. Jin, Y. Li, M. Wang, Dvr: micro-video recommendation optimizing watch-time-gain under duration bias, in: Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 334–345.

[9] Y. Zhang, Y. Bai, J. Chang, X. Zang, S. Lu, J. Lu, F. Feng, Y. Niu, Y. Song, Leveraging watch-time feedback for short-video recommendations: A causal labeling framework, arXiv preprint arXiv:2306.17426 (2023).