Proceedings of the 5th International Conference on Interdisciplinary Humanities and Communication Studies
DOI: 10.54254/2753-7064/2026.38825

Content Diversity and Hostile Language in Messi and Ronaldo Discussions on Bilibili and Reddit

Siyuan Li

The Affiliated High School of Peking University, Beijing, China
lsylsy080412@gmail.com

Abstract. This study examined online fan discussions about football on Bilibili and Reddit, two platforms that have recommender systems that select and rank online content to expose users to information. Specifically, the online discussions about Lionel Messi and Cristiano Ronaldo were analyzed to examine issues related to diversity and bias in online information exposure. In total, 200 online posts (50 posts in each of the four groups: Bilibili-Messi, Bilibili-Ronaldo, Reddit-Messi, and Reddit-Ronaldo) and 287 comments that were linked to 97 Reddit posts were analyzed. The findings showed that there were considerable differences in the topics of online discussions between the two players on Bilibili but not on Reddit. In terms of the way online content was presented, clickbait cues were found in 33% of the online posts in the Messi group and 36% of the posts in the Ronaldo group. Post-level attack flags were found in only 3% of the posts in the Messi group and 2% of the posts in the Ronaldo group. The findings of this study identified several dimensions that future studies on exposure audits can use to examine online information exposure. However, several limitations were found, including the use of different sampling frames, lack of body text in many of the online posts, and the need to validate the coding. In addition, the content was collected from the chronological search results and community listings on the two platforms. Therefore, the findings did not provide evidence of how personalized recommendations created uneven information exposure among users. Even though this study does not examine the information exposure by recommender systems in more detail, it provides a descriptive basis for such an audit on football-related topics, view points and presentation forms.

Keywords: Football Online Fan Discussions, Recommender Systems, Information Exposure, Diversity and Bias

1. Introduction

Recommender systems can shape users' information exposure by selecting and ranking content. Although various content dimensions can be measured when examining the exposure bias in football fans, the way in which popularity bias and feedback loops in recommender systems' output can develop into unequal visibility of online content for users is of special interest. Whether such bias affects exposure or even affects users' attitudes and behaviors remains to be seen and therefore needs to be analyzed in more detail.

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-7064/2026.38825

When looking at the discussions surrounding the two soccer players, it is possible to differentiate between the topics discussed, the points of view of the authors and their way of presenting information. Most posts deal with several topics, for example a player's recent matches, his upcoming transfers, his private life, or comparisons with his colleague. In terms of the point of view, most posts contain evaluations of the discussed player. Variety in terms of topics does not automatically also mean variety in terms of points of view and vice versa. Therefore, when speaking of diversity and engagement, it is necessary to differentiate between these two features [1, 2].

200 posts in four different groups of platform by player were selected for the study. The posts were accessed through Bilibili player-keyword searches as well as the new-post listings of the dedicated communities on Reddit, posted in chronological order. The study analyzed 200 posts and a supplementary sample of 287 comments linked to 97 Reddit posts.The selected content did not constitute part of the personalized recommendation feed of individual users, nor did it constitute the complete pool of items that were available to the recommender systems.

The corresponding research question is how to assess the impact of recommender systems on users' information exposure. Within the given framework of this study, content composition, presentation, stance and hostile language of the observed online posts are contrasted between the different sampling groups. The findings of this study thus may serve as a basis for future exposure audits to distinguish between the observed online content and the algorithms that recommend this content to users.

2. Literature review

2.1. Recommender bias and content diversity

From a general perspective, the information that a user can access on a platform is a subset of all the information that is published online. Recommender systems select and rank this information based on the user's past behavior (e.g. clicks, views, ratings). This form of information selection is affected by several biases that are described in detail by Chen et al. [3]. For the purpose of this study, however, it is sufficient to say that there are three main types of bias. The first type of bias is exposure bias. This type of bias refers to situations in which certain types of information are not exposed to a user at all. The second type of bias is selection bias. This type of bias occurs when a system selects a subset of available information for a user and presents this information to the user, rather than presenting the user with the entire pool of available information. The third type of bias is popularity bias. This type of bias occurs when the information that is most popular among all users is preferred by a recommender system over other information, even if this information is of poor quality.

The feedback loops of a recommender system can strengthen a popularity bias and reduce the aggregate diversity of user recommendations [4]. However, it is necessary to validate these outcomes for different types of recommendation systems [1]. For exposure to be a relevant dimension of diversity for a certain system, not only the outcomes of the system but also the design of the system itself needs to be validated for the respective dimension [1].

Measuring content diversity therefore requires more than just one index, because different dimensions can develop in different ways. For example, a sample of online messages about a topic can include many different topics, but most of them can be written by only a few authors. Or, on the contrary, a large number of authors can be represented by only a few evaluations that are all very similar. All these different aspects have to be measured and evaluated separately. The fact that online exposure of differing views does not in itself affect attitudes towards a political issue was found by

DOI: 10.54254/2753-7064/2026.38825

Nyhan et al. [5] in a Facebook experiment conducted during the 2020 US presidential election, in which users were exposed to more or fewer messages by other users of like-minded or opposite views. In that case, there was no significant change in the participants' before-study attitudes.

Research on football fandom, in particular on Messi and Ronaldo, has distinguished between star attraction and team identification [6]. Thus, mentions of players and mentions of teams are of different nature and therefore should not be used for the measurement of loyalty in samples of content for exposure audits in excess of mere access routes of sampled content. The routes through which this study's samples of content were accessed do not register personalized recommendation feeds of individual users nor the total set of items that have been made available by the recommender system of a platform for potential exposure by such individual users.

RQ1: How do player mentions, team mentions, topic diversity, and author diversity vary across the sampled platform and player groups?

2.2. Engagement optimization and content presentation

Engagement (in the sense of how users interact with online content) is relevant to information exposure because systems that recommend content online typically are optimized on user engagement on that content (e.g. clicks, time spent reading, etc.). Research on the diffusion of political messages has found that moral-emotional language increases diffusion within ideological groups [7, 2]. The latter study found that references to political out-groups in particular were strong predictors of sharing on Facebook and Twitter. But this does not necessarily translate to all online athlete fan engagement, let alone algorithmic promotion of hostile attitudes.

For example, clickbait is typically identified through attention-seeking headlines [8]. However, such a headline does not necessarily contain an attack, and therefore is distinct from other lexical indicators. Furthermore, the use of technical vocabulary in a title does not necessarily mean that the author is an expert on the topic being discussed, and thus the title-body correspondence indicator is only valid if the corresponding body text is observable and correctly linked to the title. Thus, although useful as operational proxies, lexical indicators must be used with caution and their validity assessed within the specific language and context of a sample.

The extent to which a platform's structure and governance supports communities and facilitates participation by its members is examined by Massanari [9]. Thus, while the algorithm of a platform could explain visible content, other aspects of a platform, such as its design and the communities and their members on that platform, also need to be considered. Therefore, the Bilibili keyword searches and the dedicated communities on Reddit for the two players provide different pools of content. Moreover, views on Bilibili and scores on Reddit are different measures of engagement and thus should be analyzed separately from each other. The within-group associations between the presentation features and engagement can, however, be a useful basis for future exposure audits, and for the separate question of the effects of personalized ranking.

RQ2: For each of the platform/player group combinations studied, how do terminology-based expertise proxies, clickbait indicators, emotional intensity and title-body correspondence vary? For each platform/player group combination, what relationship exists between clickbait and emotional intensity and engagement.

2.3. Rivalry stance and hostility in fan discourse

For many fans of football it is not just enjoyment of games between teams with their favorite player in them that they derive from their fan identity. Here Cialdini et al. [10] provide a useful starting

DOI: 10.54254/2753-7064/2026.38825

point to consider how identity-related expressions by fans of football are communicated online. However, as noted above simply supporting a player on a team that the fan of that player follows does not mean that fan is hostile to rivals of said team, as opposed to for example being critical of said team and/or its player(s), expressing mixed emotions concerning said team and/or its player(s), or even Expressing aggression or attacking a rival team and/or its player(s). Thus, supportive, critical, mixed and even hostile are to be treated as distinct, rather than as on a continuum of polarized fan behavior.

Different contexts can facilitate online incivility. The amount of incivility in newspaper website comments varies depending on a number of factors including the topic of the article and the quoted sources of the article [11]. Automated text analysis also must be validated for extensive use in specific applications [12]. Therefore, wh negative wording such as that contained in adjective grabs may express criticism of athletes and teams, it could be quoted in order to repudiate an insult, or it could be written in a sarcastic manner. Therefore, an indicator of hostile online communications must establish who is speaking, whether that speaker is being endorsed by the author of the post, and who or what is being attacked by the speaker. Simply disagreeing with an assessment of athletic performance is not equivalent to an online attack of a player or of a group of fans.

All the above variables of interest for research into information exposure can occur repeatedly within one post, and even within one comment. Also, within a sample of rivalry fan postings, topic variety can be prevalent while most of the instances of that topic variety will express repeated hostile evaluations of the opponents. In addition, rivalry fan content need not be hostile. Therefore, RQ3 is to investigate the coded instances of expressions of stance and of attacks in the sample of postings and in the supplementary comments, as separate from an analysis of the various measures of topic diversity and of author diversity. With regard to comments, as opposed to postings, the variably nested set of comments within a posting (i.e. the comments plus the post itself) is of interest. The available Reddit comments constitute a supplementary sample of visible comments on the same rival players as the analyzed fan postings.

RQ3: How do coded stance (positive, negative, neutral) and explicit attacks (yes/no) in fan postings of both platforms compare with those found in the available comments of Reddit postings?

3. Methods

3.1. Design and sampling

This study employs a cross-sectional purposive quota design.We collected 200 posts, distributing them evenly across the four groups of Bilibili-Messi, Bilibili-Ronaldo, Reddit-Messi, and Reddit-Ronaldo posts. One post had to be excluded from the analyses as there was a name-collision with another user and, thus, was not correctly attributed to the appropriate user in the recorded collection protocol.

The procedure for sampling Bilibili publications and Reddit posts was recorded. In order to collect Bilibili publications, keyword searches for the names of the two players were performed in chronological order of publication date. Posts on Reddit were collected from the new posts page for the communities r/messi and r/cristianoronaldo, excluding posts by the AutoModerator. Cross queries that returned duplicates were replaced by the duplicate in the collection procedure. The search for public keyword searches on Bilibili and a community for one of the players on Reddit created two different sampling frames. Hence, no causal effect of platform membership on content selection can be determined with this study.

DOI: 10.54254/2753-7064/2026.38825

Publication dates for the Bilibili samples considered in the analysis range from 1 December 2025 to 8 August 2026 for the 'Messi' stratum and 7 September 2025 to 8 August 2026 for the 'Ronaldo' stratum. For Reddit, the analysis uses a range of dates from 30 July to 8 August 2026 for 'Messi' and 17 July to 8 August 2026 for 'Ronaldo'. Two absolute publication dates were unavailable at the time the source material for this report was created. One of these was recorded only as a relative time stamp.Publication dates were recorded where available; two records lacked an absolute publication date. The source material for this report has not been independently verified with respect to the complete and accurate online replication of historical publication dates and/or online time stamp conversion to a precise local online time zone.

The supplementary comment data, that were collected within the posts, consist of 287 comments of 97 Reddit posts. For each of the Reddit posts with comments, up to three of the most visible comments were kept using the default ordering. Comments on Bilibili were not collected. Since there were no comments collected from Bilibili in this study, therefore we are unable to compare and contrast the comment hostility between the Chinese platform and Reddit.

3.2. Measurement and coding

The initial dataset includes sampling stratum and also for the posts the indicator for whether the post explicitly mentions the other player. The post-level measures for the posts are primary topic of the post, team mentions (for all teams mentioned in the post), terminology-based expertise (as a lexical proxy), clickbait indicators, sentiment, stance and lexical attack indicators. The team mentions for a post are multivalued, whereas the other measures are single-valued. The measure for terminology-based expertise, for example, counts how often certain terms for expertise are used in a post. This indicator can only be scored when sufficient body text is available. If no body text is observable, then the case is reported separately.Cases with unavailable text for this measure are reported separately.

Two human coders were given the same codebook to code the same set of posts. Intercoder reliability was measured by means of percentage agreement and by means of Cohen's kappa for the single coded variables.As mentioned above, the full original lexicons and the extraction implementation are not available in the working materials for this report. Thus, the reported annotation was not reproduced from source text for this report.

3.3. Data integrity and evidence availability

The supplied workbook versions have been reconciled by record id and source URL. In the two 12345 versions, 801 raw-post cells contain the string '533' whereas the corresponding cells in the Fully_Coded workbook for source 12345 (as well as for the other source 12345) are all blank. Similarly for the 287 comment-score cells. The 801 raw-post cells as well as the 287 comment-score cells with string '533' are treated as missing evidence. Thirteen Bilibili included posts are missing both view counts and comment counts in the Fully_Coded workbook for source 12345. Engagement summaries are calculated based on the available data only. All original post labels have been retained.

One of the included Bilibili posts has tags relating to an organist but the title and body of the post relate to Messi. There are 30 records where title-body correspondence codes have been marked as 'observable' but this does not necessarily mean that all the bodies have been correctly sourced. As before, missing body text has been reported as separate cases. The results from the correspondence

DOI: 10.54254/2753-7064/2026.38825

coding are reported as they were from the recorded sampling procedure and therefore have not been independently verified as having achieved perfect semantic matching.

3.4. Analytical strategy

We will present a range of descriptive statistics for the number of posts and corresponding categories for each of the sampling frames as well as Wilson intervals for binomial distributions. For assessing topic diversity, we calculated the Shannon entropy $H = -\Sigma\ p_i\ \ln(p_i)$ for each of the sampling groups. The effective number of topics was then computed as exp(H), where zero-proportion categories contributed zero to the sum.The author diversity for a set of posts is computed as one minus the sum over all authors of the squared fraction of posts by that author. "Other" was treated as a substantive residual topic category, rather than as missing data. A supplementary analysis excluded this category in order to assess the sensitivity of the findings on topic-diversity.

For the larger sparse tables, the number of Monte Carlo samples is set to 99,999 and the seed to 20260929. All tests within the variable families, for both the pooled as well as the within-platform analyses, are then Holm-adjusted for multiple testing. These tests are all assumed to be of independent posts. Pooled player differences are not interpreted as platform interactions, and thus are not tested for such. The within-platform analyses, however, do test for differences between the two players.

For each platform-player group, engagement analyses were conducted and displayed separately from the main analysis because Bilibili video view counts and Reddit scores are different measures. The amount of available engagement data for the 47 Bilibili videos featuring Messi, the 40 Bilibili videos featuring Ronaldo, and the 50 videos from each of the two Reddit groups was recorded. In cases where engagement data was missing for a video, it was not imputed. To examine the association between a video's emotional intensity and that video's engagement, Spearman correlations were conducted for each group of videos. The comments section for supplementary comments for each video was examined and summarized descriptively; these comments were not treated as independent observation(s).

3.5. Ethics and reporting

In the analysis protocol that was recorded for the collection of the data the public sources were used and the authors were identified by hashed author identifiers. Thus, all records with identifiable quotations as well as hyperlinks to the public sources were treated with due care. However, the institutional ethics approval or exemption was not documented in the working materials, and thus re-identification was possible for the cases where it had not been avoided in the first place.

4. Results

4.1. Sample composition and evidence availability

The figures for the analysis of the 200 included cases are for the coding pass A for the 50 cases in each of the four groups of platform-by-player. Note that there is one additional record in the workbook for the record that was excluded from the collection protocol for a name collision with the name of another well-known person, not Lionel Messi. This dataset consists of 287 comments corresponding to 97 Reddit posts. In terms of the data within this dataset, there are 148 comments

DOI: 10.54254/2753-7064/2026.38825

from 50 posts in the Messi stratum and 139 comments from 47 posts in the Ronaldo stratum. We did not collect any comments from Bilibili.

The analysis uses results from coding pass A, a purposive sample of 200 records, for which the analyses are exploratory and the tests of the hypotheses are conducted using tests of ten variable(s) (separately for the analysis of all records and for each platform) and the unobservable title-body cases for the corresponding tests have been excluded from the analyses, and the distribution figures include all records, with the missing evidence noted. The analyses assume that the posts are independent, and do not take into account the fact that there are repeated authors and events.

4.2. RQ1 content composition and diversity

In the sampling stratum of the Messi sample 73% of the posts about Lionel Messi mentioned him alone, 6% mentioned both players and 21% did not mention either of them. In the sampling stratum of the Ronaldo sample 60% of the posts about Cristiano Ronaldo mentioned him alone, 11% mentioned both players, 1% mentioned only Lionel Messi and 28% did not mention either of them. These distributions are different from each other (Pearson chi-square = 132.52, df = 3, Cramer's V = .814, Holm-adjusted $p < .001$). Note, however, that these distributions are also a result of the sampling design and thus do not in themselves prove selective exposure or fan loyalty in online fan communities. See Figure 2.

The distribution of primary topics among the pooled sample of posts also differs significantly between the two strata of posts (Monte Carlo Fisher test, 99,999 simulations, p = .00057; Holm-adjusted p = .00342; Cramer's V = .354). Of the posts about Messi in the Messi-stratum of posts, 32% were about his personal life, compared to 11% of the posts about Ronaldo in the Ronaldo-stratum of posts. Conversely, 44% of the posts about Ronaldo in the Ronaldo-stratum of posts were about Other topics, compared to 27% of the posts about Messi in the Messi-stratum of posts. None of the posts in either stratum of posts were about commercial promotion of either player. A figure (Figure 1) illustrates the distribution of the ten topics among the two groups of posts.

Within Bilibili, the percentage of personal life-related content for Messi is 56% and for Ronaldo is 18%, while for the category of Other topics related to Messi it is 14%, and for Ronaldo it is 46%. Thus, the within-Bilibili test for difference in topic distributions between the two players is highly significant (Monte Carlo Fisher test, p-value = .00001; Holm-adjusted p-value = .00007; Cramer's V = .848). Within Reddit, however, the distributions for Messi and Ronaldo do not differ significantly (p-value = .98261; Holm-adjusted p-value = 1.000; Cramer's V = .026). As noted above, the absence of a significant difference between two groups in a test of association does not constitute evidence of an interaction effect between the groups. Figure 1 therefore reports all topic distributions within each platform for both players.

This figure plots the distribution of primary topics for the four groups of posts, Bilibili-Messi, Bilibili-Ronaldo, Reddit-Messi, and Reddit-Ronaldo. Each platform is plotted separately. All groups were made up of 50 posts and each post was assigned a single primary topic. The bars represent the proportion of posts assigned to each of the 10 primary topics and the error bars represent the 95% Wilson confidence intervals for the relevant proportion. The intervals were not calculated to account for the purposive sampling or for possible author clustering. The 'Other' category is a residual category for other topics and does not denote lack of evidence. None of the groups had any posts with the primary topic 'Commercial Promotion'.

DOI: 10.54254/2753-7064/2026.38825

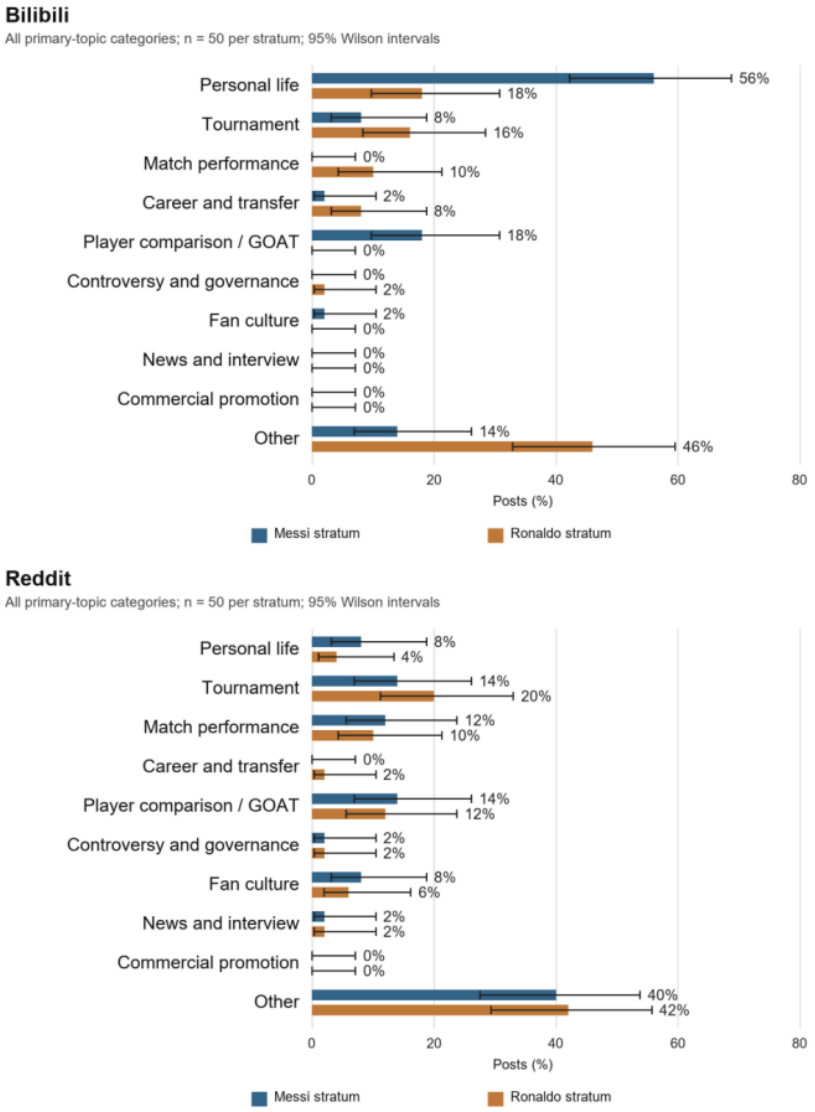


Figure 1. Primary topic categories by player and platform (Bilibili vs. Reddit)

Shannon entropy measures the diversity of primary topics within each of the sampling groups. The maximum possible Shannon entropy for 10 topics using natural logarithms is ln(10) ≈ 2.303 when all 10 topics are equally frequent. The highest Shannon entropy found for the primary topics was 1.732 for the Reddit-Messi sampling group, corresponding to an effective number of topics of 5.65. Note that this value expresses the number of topics that would be perfectly frequent at the found entropy, not the number of topics in a single post. The effective number of topics is a count, not a percentage.

Within each of the sufficently documented platforms (Bilibili and Reddit), the number of authors who wrote the 200 sampled posts is: 36 (Bilibili/Messi), 45 (Bilibili/Ronaldo), 34 (Reddit/Messi), 27 (Reddit/Ronaldo). Author diversity was measured as one minus the sum of the squared proportions of posts contributed by each author, using posts with available author identifiers.

The variable for team mentions was a multivalued variable, thus there were no restrictions on the number of teams mentioned in a single post. Therefore, the percentages of mentions for individual

DOI: 10.54254/2753-7064/2026.38825

teams do not have to sum up to 100%. We are counting explicit mentions of teams, not the level of fans' attachment to the teams.

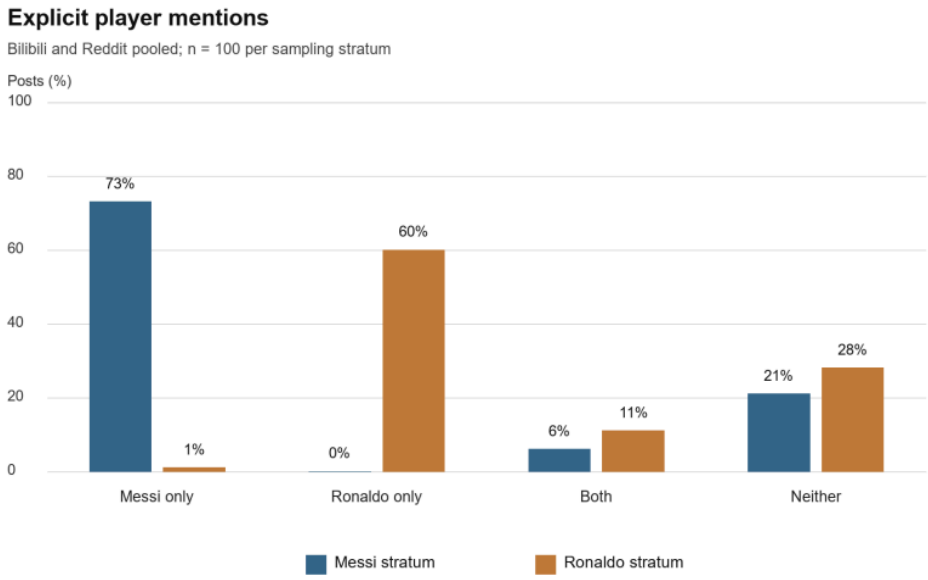


Figure 2. Explicit player mentions by sampling stratum

Each pooled player stratum contains 100 posts. Sampling strata identify collection groups, not the main subject or the direction of a post's evaluation. Categories are mutually exclusive.

4.3. RQ2 textual presentation and engagement

The frequency of the terminology-based expertise proxy (Low: 76% of the Messi-stratum postings, 72% of the Ronaldo-stratum postings, Medium: 21% and 22%, respectively, and High: 3% and 6%, respectively) did not reveal significant differences between the two player groups (Holm-adjusted p = 1.000). Note that this proxy measures the frequency of lexical matches rather than the author's qualifications or even the factual correctness of the postings.

The percent of cases coded for clickbait in the Messi sampling stratum was 33%, and 36% for the Ronaldo sampling stratum. For Bilibili, the percentages for clickbait in the Messi and Ronaldo strata were 30% and 42% respectively. The corresponding percentages for the Reddit sampling stratum were 36% and 30% respectively. In all cases, there was no significant difference for clickbait between the two players' sampling strata at the .05 level, after adjustment as necessary.

Title/body correspondence was unobservable for 85% of the 200 total samples (170 samples). The 30 samples of Bilibili posts where title/body correspondence was observable consisted of 23 posts from the Messi sampling stratum and 7 posts from the Ronaldo sampling stratum. The corresponding distribution of the within-category topic test for title/body correspondence was as follows: for the 23 Messi sampling stratum posts of Bilibili where title/body correspondence was observable, 7 samples were matches, 4 samples were partial matches and 12 samples were mismatches. For the 7 samples from the Ronaldo sampling stratum of Bilibili where title/body correspondence was observable, the corresponding figures were: 5 matches, 2 partial matches and 0 mismatches. The mismatch percentage for the 23 samples from the Messi sampling stratum was therefore 52.2%, while for the 7 samples from the Ronaldo sampling stratum it was 0%. The small and unequal number of samples precludes conclusions. The corresponding adjusted test statistic was

DOI: 10.54254/2753-7064/2026.38825

not significant (p = .2765). For the remaining platform/speaker, Reddit, as mentioned earlier, title/body correspondence was unobservable for all samples. In addition, for one of the records with a title/tag conflict, it was also unobservable whether the corresponding body text was correctly associated with the title/tag. The corresponding figures for the number of mismatches therefore should not be construed as having been validated for semantic matching in these cases. See Figure 3 above for corresponding figures for all samples.

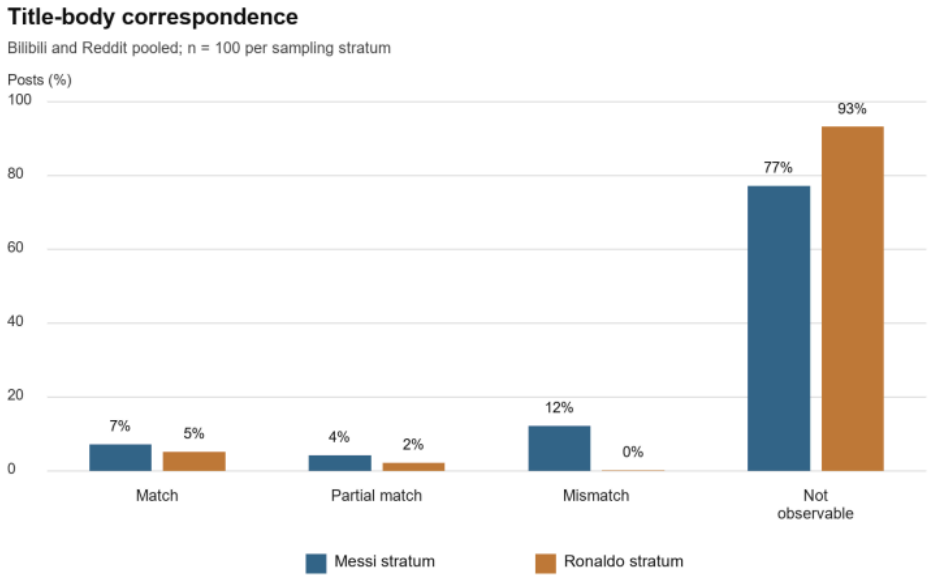


Figure 3. Title-body correspondence by sampling stratum

Not observable indicates unavailable evidence, not a mismatch. Of 200 posts, only 30 had observable correspondence; all were on Bilibili.

These 13 Bilibili post records contained Bilibili "view" and "comment" numbers that were not found in the corresponding locations in the source workbook. The subsequent versions of these same records contained a single number, 533, which had been repeated in the corresponding locations of the prior versions. No view counts or comment counts existed for thirteen of the Bilibili posts in the Fully_Coded workbook. The repeated value of 533 for other versions of the workbook was therefore treated as missing rather than actual engagement. All relevant data for engagement analyses consisted of the available observations (47 Bilibili posts in the Bilibili-Messi group and 40 in the Bilibili-Ronaldo group) as well as the full 50 posts in each of the Reddit groups, with all posts having existing engagement scores.

The engagement analyses investigate differences in view counts on Bilibili between videos that include clickbait and videos that do not. Furthermore, within each platform-by-player combination, the analyses investigate the association between emotional intensity and engagement.The associated correspondence correlation was of medium strength and significant after the Holm adjustment for all eight tests conducted (rho = 0.429, n = 47, unadjusted p-value = .00263; Holm-adjusted p-value = .0211). The same tests for the other seven correlations did not reach significance after the Holm adjustment. As with all findings based on the terminological-based expertise proxy, the post-view figures have not been adjusted for the age of the posts, authors' reach, or for event clustering. Thus, they do not provide evidence of the recommendation ranking of content with higher emotional intensity.

DOI: 10.54254/2753-7064/2026.38825

4.4. RQ3 stance and hostile wording

Neutral was the most common stance label for both strata (55% of the posts in the Messi stratum and 63% in the Ronaldo stratum). While there was some variation in the share of posts classified as supportive of the players in both strata (15% in the Messi stratum, vs. 20% in the Ronaldo stratum), this variation is not itself a finding about the two fan populations, since the stance labels are composed of both a target and a valuation, and the initial coding pass A included some inference about the community of experts from the text itself. Thus, corresponding findings about the valence of the content in a given post (i.e. sentiment) are not independent of these findings.

The count of explicit attack flags was low across all samples. Specifically, there were three posts in the Messi-stratum with attack flags (3% of all posts in this stratum) and two posts in the Ronaldo-stratum with attack flags (2% of all posts in this stratum). This does not provide strong evidence that there are equal levels of hostility in the two populations. In the Reddit comment data, for example, there were attack flags in one of 148 comments in the Messi-stratum of analyzable comments (0.7%), and in 9 of 139 comments in the Ronaldo-stratum of analyzable comments (6.5%). Note that the labels for the flags for incivility (Mild/Severe) have not been validated by human coders for these classifications. Hence these data should not be used as if they had been validated to measure hostility to players in online communities. No significance test was performed for these observations because comments were sampled as nested units within the posts, and thus the set of comments does not constitute a complete set of all online discussion about these events.

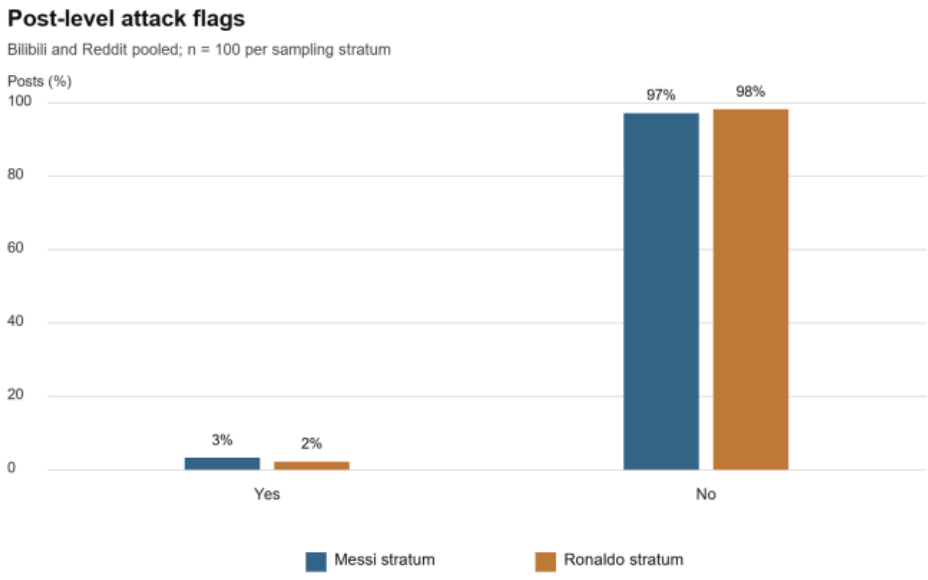


Figure 4. Post-level attack flags by sampling stratum

Yes and No are mutually exclusive and sum to 100% within each stratum. The flags are lexical coding outcomes, not a validated measure of psychological polarization.

4.5. Intercoder reliability

We calculated the intercoder reliability of two independent coders for the 200 or so postings that were used for the human coding for all variables. For the primary topic, the agreement was 99.5% ($\kappa = .994$); for clickbait the agreement was 77.5% ($\kappa = .411$); for emotional intensity the agreement was

DOI: 10.54254/2753-7064/2026.38825

60.0% (κ = .332); and for the explicit attacks there was 98.0% agreement (κ = .328). The lowest percentage agreement was for the emotional intensity variable but the agreement for the explicit attacks was high and therefore included in the analyses. Yet the Cohen's kappa for the latter variable was low because of its high category prevalence and hence, the findings for this variable have to be treated with caution as for the other variables with low intercoder agreement.

5. Discussion

5.1. Distinguishing content from fandom

The Messi-Ronaldo topic contrast is concentrated in the Bilibili sample whereas the Reddit topic distributions are similar to each other. This finding is purely descriptive and does not necessarily transfer to the two fan populations. In other words, a significant contrast in one platform and no significant contrast in the other platform does not constitute an interaction effect. Also, we cannot attribute our findings to the recommendation algorithms of the two platforms and therefore cannot generalize to the two fan populations based on this single design.

On the other hand, a post that mentions a football player can contain a discussion of his relatives, teammates, or even characters from a football game. Thus, a keyword search defines a bigger population than posts written about specific football players. This must be taken into account when designing an exposure audit, i.e. when determining the population of contents that will be analyzed and the relevance criteria.

5.2. Contribution and limits

We provide a measurement framework for a descriptive analysis of football-related content of sampled platforms. We distinguish between the analysis of the content's topic structure, the author structure, the structure of the way content is presented, the taker's stance, and the extent of hostile language usage. We note several limitations of the available data. Most importantly, the observations of individual recommendations as well as of shifts in users' attitudes towards sports-related news are not available. Consequently, we cannot establish whether observed exposure differences are caused by algorithms or by users' polarization. For a detailed analysis of these two separate empirical questions, see Nyhan et al. [5].

There are several limitations that we have to keep in mind when interpreting our results. For example, the sampling frames and the time span of the sampled items differ between the platforms. Some of the records need to be source matched and most of the media and body content is not available. We are missing engagement information for 13 of the Bilibili posts. Inter-rater agreement for the two human coders, however, differed between the variables. For the content features, especially for the emotion intensity of the posts, agreement was lower and the values for clickbait as well as for explicit attacks even achieved rather low kappa values.Some of the posts were written by the same authors or even refer to the same events. The comment sample only contains a few very visible comments and there is no Bilibili sample for comparison. Most importantly, we do not have information on how personalized recommendations would compare to the same content in a matched reference feed. Therefore, the content distributions that we present do not establish a causal effect of the algorithmic selection.

DOI: 10.54254/2753-7064/2026.38825

6. Conclusion

The discussions about football contained mostly different topics in the sample of football-related posts, with the biggest contrast between the two ends being the topic of comparison between Messi and Ronaldo in the Bilibili sample and similar distributions of topics in the Reddit sample. Few of the sampled posts contained explicit post-level attack flags under the coding rules used in this study. Therefore, in studying bias in user information exposure, it is important to study topic diversity, author diversity, presentation, stance, and hostility separately. The current findings describe the sampled content and therefore do not establish causal effects of the recommender systems studied here. However, future research could compare these various dimensions of bias in the content in a personalization system's output and in matched reference content, with consistent sampling and independent human validation of the coding.

References

[1] Möller, J., Trilling, D., Helberger, N., & van Es, B. (2018). Do not blame it on the algorithm: An empirical assessment of multiple recommender systems and their impact on content diversity. *Information, Communication & Society, 21*(7), 959-977.

[2] Rathje, S., Van Bavel, J. J., & van der Linden, S. (2021). Out-group animosity drives engagement on social media. *Proceedings of the National Academy of Sciences, 118*(26), e2024292118.

[3] Chen, J., Dong, H., Wang, X., Feng, F., Wang, M., & He, X. (2023). Bias and debias in recommender system: A survey and future directions. *ACM Transactions on Information Systems, 41*(3), Article 67.

[4] Mansoury, M., Abdollahpouri, H., Pechenizkiy, M., Mobasher, B., & Burke, R. (2020). Feedback loop and bias amplification in recommender systems. *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*, 2145-2148. https://doi.org/10.1145/3340531.3412152

[5] Nyhan, B., Settle, J., Thorson, E., Wojcieszak, M., Barberá, P., Chen, A. Y., Allcott, H., Brown, T., Crespo-Tenorio, A., Dimmery, D., Freelon, D., Gentzkow, M., González-Bailón, S., Guess, A. M., Kennedy, E., Kim, Y. M., Lazer, D., Malhotra, N., Moehler, D., . . . Tucker, J. A. (2023). Like-minded sources on Facebook are prevalent but not polarizing. *Nature, 620*, 137–144.

[6] Hayat, T. (Z.), Samuel-Azran, T., Laor, T., & Galily, Y. (2024). Unveiling the intersection of individual stardom and team loyalty in social network reflections: The case of soccer-stars Ronaldo and Messi. *American Behavioral Scientist*.

[7] Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences, 114*(28), 7313-7318.

[8] Chakraborty, A., Paranjape, B., Kakarla, S., & Ganguly, N. (2016). Stop clickbait: Detecting and preventing clickbaits in online news media. *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 9-16.

[9] Massanari, A. (2017). #Gamergate and the Fappening: How Reddit's algorithm, governance, and culture support toxic technocultures. *New Media & Society, 19*(3), 329-346.

[10] Cialdini, R. B., Borden, R. J., Thorne, A., Walker, M. R., Freeman, S., & Sloan, L. R. (1976). Basking in reflected glory: Three (football) field studies. *Journal of Personality and Social Psychology, 34*(3), 366-375.

[11] Coe, K., Kenski, K., & Rains, S. A. (2014). Online and uncivil? Patterns and determinants of incivility in newspaper website comments. *Journal of Communication, 64*(4), 658-679.

[12] Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis, 21*(3), 267-297.