Ahmed Shbair*, Lutf Ul Rahman Haqmal, Tuba Livberber and Melih Günay

Mapping offensive discourse in sports on X platform: a weakly supervised NLP and journalism-based analysis

https://doi.org/10.1515/omgc-2026-0027
Received April 23, 2026; accepted August 30, 2026

Abstract

Purpose: Social media platforms have become a central place for sport-related public discourse that often involves emotionally charged and offensive language. This study examines offensive sports discourse on X (formerly Twitter) through an interdisciplinary approach that combines natural language processing (NLP) and journalism research.

Design/methodology/approach: A dataset of 7,279 sport-related tweets that were collected between January and December 2025 was analyzed. An initial set of 350 tweets were manually annotated first and then used to train a machine learning classifier based on TF-IDF features and logistic regression. The classifier got an accuracy of 0.71. A confidence-based weakly supervised labelling strategy was then applied, resulting in 2,042 high-confidence labelled tweets.

Findings: Quantitative analysis shows that offensive tweets demonstrated higher average engagement levels than regular tweets in terms of likes, retweets, views, and replies, although these differences were not statistically significant. Unverified users predominantly produce offensive content and this contents appear more frequently in original tweets than in replies. Lexical analysis reveals that offensive discourse combines explicit profanity with sports-related terminology. The findings demonstrate how NLP-based methods can support journalism research by revealing large

Article Note: This article underwent double-blind peer review.

***Corresponding author: Ahmed Shbair**, Faculty of Communication, Akdeniz University, Antalya, Türkiye, E-mail: ahmedshbairpal@gmail.com. https://orcid.org/0000-0002-5319-211X
Lutf Ul Rahman Haqmal and Melih Günay, Department of Computer Engineering, Akdeniz University, Antalya, Türkiye, E-mail: 202351075502@ogr.akdeniz.edu.tr (L.U.R. Haqmal), mgunay@akdeniz.edu.tr (M. Günay). https://orcid.org/0009-0000-1740-6339 (L.U.R. Haqmal). https://orcid.org/0000-0001-5409-6720 (M. Günay)
Tuba Livberber, Department of Journalism, Faculty of Communication, Akdeniz University, Antalya, Türkiye, E-mail: tubalivberber@akdeniz.edu.tr. https://orcid.org/0000-0002-9879-2135

Open Access. © 2026 the author(s), published by De Gruyter. This work is licensed under the Creative Commons Attribution 4.0 International License.

scale patterns in digital sports communication, and how these methods can be used in interdisciplinary settings to provide unique and novel insights through the analysis of the available data.
Practical implications: This research assists the growing field of computational journalism by leveraging NLP techniques with journalism-based research in an interdisciplinary settings to gain unique and novel insights about the domain-specific online sports discourse.
Social implications: This study helps us to achieve empirical insights and intuitions into fan behavior, interaction and the pattern of their engagement. At the same time, it is very important to perform journalistic interpretation for the purpose of contextualizing computational findings within ethical, cultural, and professional frameworks.
Originality/value: This study adopts established NLP based methods to explore large scale communicative patterns in online sports discourse.

Keywords: sports journalism; offensive language; X platform; natural language processing; logistic regression; artificial intelligence

1 Introduction: sports journalism and social media discourse

1.1 Journalism in sport: background

Sports journalism differs from other forms of journalism in that it combines factual reporting, emotive storytelling, and the construction of a sense of community. Sports journalism plays an important role because it can influence how people think and feel, often shaping their attitudes in ways that can be just as powerful as the sporting events themselves (Mutz and Gerke 2021). Sports journalism was called a “toy department” but it is actually one of the most significant and attractive sorts of journalism (Perreault and Bell 2022). While watching sports, people are more emotionally involved, but the media also affects how people react (Mutz and Gerke 2018). Therefore, sports journalism has an impact on the emotions of players, causing public reaction due to its role in shaping global journalism (Srivastava and Singh 2025). Media has the ability to discuss multiple issues and shape public perception (Kurniawan 2024; Shbair and Hasfi 2022). It comprises various elements, including practice, profession, commerce, and ideological perspectives (Broussard et al. 2019). According to Deuze (2005), the ideological foundation of journalism shapes how journalists view their role, especially in terms of serving the public,

being fair and independent, delivering news quickly, and upholding professional ethical standards.

With the growing shift of sports discussions to social media platforms, sport has gradually become part of people's daily routines and social lives (Popek 2024). At the same time, journalists are now working in environments where audience participation happens quickly, directly, and on a global scale (Uth 2025). This development has also made social media an important strategic tool for sport clubs, as it helps them build trust, communicate with passionate supporters, expand their global presence, and manage both the benefits and challenges of online interaction (McCarthy et al. 2022). At the same time, it has also led to a noticeable increase in hostile and offensive comments (Kearns et al. 2022). It has also provided new spaces for public hate directed at athletes (Kavanagh et al. 2016). As a result, it creates challenges for journalistic ethics, affects the emotional well-being of athletes, and raises questions about how to maintain responsible communication in the media environment and media practices.

1.2 X platform background

With the emergence of digital media, the importance of producing and consuming sports content has increased (Hutchins and Rowe 2012). X (formerly Twitter) has the potential to influence sports content due to its ability to engage with a large number of users (Gibbs and Haynes 2013). Asiri and Saleh (2024) highlight the importance of the X platform in discussing social issues and user interactions. It is also an important space that brings users together to interact and express opinions (Curran et al. 2011). It is characterized by its ability to extract huge amounts of data from millions of users (Ghosh et al. 2020; Hussiny et al. 2023). On X platform, journalists, athletes, clubs, and fans interact with one another during live matches, transfer announcements, and moments of controversy (Burgers et al. 2023). Therefore, the X platform now attracts millions of users every day due to the high demand. This widespread use allows people to easily express and share their thoughts and feelings (Ghosh et al. 2020; Hussiny et al. 2023; Muindi 2018). Rua Gámez (2023) explains that the way the platform is built makes it much easier for hate speech to spread, because hurtful and offensive posts can travel very quickly and reach a lot of people. Unlike traditional media, where content is usually reviewed before it goes public, moderation on X is mostly automated and often happens after the material is already posted. This situation raises concerns about harmful communication going unchecked (Mathew et al. 2019), especially in emotionally intense sports discussions, and contributes to inter-fandom conflicts (Karki et al. 2024), making this a unique domain for

undertaking multi-disciplinary research and gaining valuable insights from these online conversations. Social media can play a very important role, as it gives citizens space to express their views and determine their attitudes and opinions in ways that are not available through traditional channels (Jost et al. 2018) It allows sports fans to effectively evaluate and discuss athletes' performance. (Klutse et al. 2023). In this context, X platform (formerly Twitter) can be understood as an opportunity for the rapid exchange of public discourse among users, and is considered an effective means of transmitting messages quickly (Hasfi et al. 2024: 330). This makes it a suitable subject for studying how offensive discourse spreads and gains engagement.

This paper explores how sports fans interact on the X platform, commenting on and sharing content that spreads more easily. Furthermore, the use of NLP in quantitative data analysis contributes to a relatively underutilized area of research on the importance of verifying more data-driven content, specifically through sport-related comments over a defined time period. This paper aims to expand research on audience interaction with social media in sports journalism by exploring the use of offensive language by sports fans on X platform. a) This paper complements research on audience interaction with social media in sports journalism is complemented in this paper by studying the use of offensive language by sports fans on the X platform. a) This paper studies the X platform users in the context of sports journalism. b) It analyzes user tweets through NLP to understand the reasons behind the offensive language and the most used comments. This study addresses user comments, interactions, dialogues, and offensive language on the X platform. This analysis is conducted from qualitative analysis of tweets and quantitative NLP analysis, promoting its use in journalism research and exploring modern methodological findings and approaches that many previous studies lack. Accordingly, the following research questions guide the analysis. The 2025 dataset is designed to reflect the numerous sporting events taking place globally. The authors also aimed to gather the most up-to-date fan engagement data up to the study date to obtain truly valuable insights. This paper is highlighted through its contribution to the knowledge of the patterns of offensive discourse, the interaction dynamics between and among users and the cultural qualities of sports communication in digital spaces.

2 Literature review/theoretical background

2.1 AI and NLP in sport journalism

The term "artificial intelligence" was first introduced by John McCarthy during a conference in the 1950s (Stone et al. 2022). Artificial intelligence (AI) is perhaps the oldest and most broad field of computer science, dealing with all aspects of

mimicking cognitive functions for real-world problem-solving and building systems that learn and think like people (Holzinger et al. 2019). According to Walker et al. (2021), artificial intelligence has diverse applications which influence various areas and domains such as social media, the Internet of Things, and financial sectors. According to Shehab et al. (2020), Artificial intelligence (AI) performs tasks more innovatively than humans because it can create AI-driven scenarios for potential outcomes.

Natural Language Processing (NLP), which is one of the most important subbranches of Artificial Intelligence, has achieved a lot of progress in recent years in a diverse range of fields such as sentiment analysis, information extraction and machine translation (Khurana et al. 2023). The main reason behind the recent progress of NLP is the rapid rise of large language models (LLMs) and the advancements in deep learning and neural networks (Krasadakis et al. 2024).

The integration of AI and NLP in computational journalism has fundamentally transformed how news is produced and published (Moran and Shaikh 2022; Palanimurugan and Shanthi 2025). This integration has been particularly revolutionary in the domain of sports reporting (Tesse 2024). According to (Sarkar et al. 2024), Artificial Intelligence has the potential to drastically transform the landscape of storytelling and story consumption in the sport journalism domain. Thanks to the availability of a large volume of online data in the sport domain, lots of research has been conducted to process and transform this sports data with the help of Artificial Intelligence and Natural Language Processing techniques (Canavilhas 2022). In the current era of digitalization, Artificial Intelligence technologies can be utilized to assist and enhance the journalistic domain, primarily due to their ability to address fundamental issues and challenges in contemporary journalism such as detecting and flagging fake news, adherence to the editorial standards and content personalization (Ali and Hassoun 2019). However, there are growing disagreements among scholars over how to utilize artificial intelligence in journalism. Stray (2019) has performed a survey that covers the researches that have utilized AI techniques in investigative reporting. Kothari and Cruikshank (2022) assert that some journalists are optimistic about AI's potential to improve content quality and streamline journalistic workflows, while others express concerns about job security within the media sector. There are several reasons why journalism has begun to shift, including cultural influences, technological advancements, and other changes (Lewis and Usher 2016). Therefore, in recent years, journalism has undergone significant changes in the business, leading media companies and journalists to utilize technological advancements as a means of survival in the journalistic work environment (Usher 2017).

In contrast, personal relationships and social interactions appear to be essential elements for expressing and interpreting emotions (Hussiny et al. 2023). Understanding and expressing emotions is one of the most important challenges that scholars in the field of Natural Language Processing (NLP) focus on (Ghosh et al. 2020). Sentiment Analysis (SA), also known as opinion mining (Pang and Lee 2008), is a technique employed to identify and assess emotions or attitudes inside text. It works by automatically detecting and extracting out the subjective meaning from written information (Ghosh et al. 2020). Sentiment Analysis (SA) is a part of Natural Language Processing (NLP) that helps computers understand how people feel and what they think based on text or other types of input. It is an analytical technique that studies people's feelings and opinions regarding a given topic, having seen drastic growth over the last few years (Zhao et al. 2013). It allows machines to understand, sort, and judge how people feel or think about different things (Hussiny et al. 2023). In most cases, sentiment analysis classifies data as positive, negative, or neutral as part of emotion identification.

Therefore, many previous studies have discussed analyzing audience speech across social media platforms and analyzing emotions in multiple ways, but few have addressed the NLP methodology in language processing. In sports journalism, with the assistance of NLP, we can perform systematic analysis of audience discourse. It usually helps us to achieve empirical insights and intuitions into fan behavior, toxicity, and the pattern of their engagement. At the same time, it is very important to perform journalistic interpretation for the purpose of contextualizing computational findings within ethical, cultural, and professional frameworks.

Rather than proposing new AI architecture, the study adopts established NLP based methods to explore large scale communicative patterns in online sports discourse. Accordingly, the following research questions guide the analysis:

RQ1: To what extent does offensive language appear in sports-related discourse on X?

RQ2: How does user engagement differ between offensive and normal sports-related tweets on X?

RQ3: What role does user verification status play in the production of offensive sports-related content on X?

RQ4: What lexical patterns characterize offensive sports-related discourse on X?

Accordingly, the following research questions guide both the descriptive and interpretive dimensions of the analysis:

RQ5: How can the identified offensive discourse patterns be interpreted within the cultural and communicative dynamics of sports journalism on X?

These research questions are addressed through a combination of NLP based text classification, quantitative engagement analysis, and qualitative journalistic interpretation.

2.2 Theoretical framework: the online disinhibition effect

The present study adopts the Online Disinhibition Effect as a theoretical framework to understand how patterns of expression and behavior may vary in online contexts. People may be less inhibited online and may express themselves differently than they would in face-to-face encounters (Suler 2004). The study identifies distinct types of disinhibition including toxic disinhibition that can take the form of impolite language, harsh criticism, rage, hostility and threats. This makes it a valuable concept to the study of abusive conversation in online environments.

Additional support for the concept has come from later studies; however, they have underlined the necessity of contextualizing it. Joinson (2007), for example, argues that online de-inhibition may include reduced concern with self-presentation and the judgment of others, and examines research of various forms of uncontrolled verbal behavior, including swearing and insults, in computer-based communication. However, the study emphasizes that this conduct should not be explained merely by anonymity, but by the user's goals and the broader social context.

Similarly, the study by Lapidot-Lefler and Barak (2012) demonstrates that toxic online disinhibition can be shaped by several characteristics of computer communication, not just by anonymity. Their experimental study found that the absence of eye contact was the most influential factor among those they tested, while anonymity had a direct effect on threats. These findings suggest that hostile and uncontrolled online behavior may result from the interaction of several communication conditions rather than a single factor.

Accordingly, the Online Disinhibition Effect provides a theoretical lens through which to interpret the patterns of abusive discourse observed in the current study in sports discussions on X platform. The study does not assume that the status of authentication is a direct measure of anonymity, but rather uses this framework to understand how the characteristics of the digital communication environment and the reduction of certain social constraints can be associated with patterns of hostile expression in online sports discourse.

3 Research methodology

A mixed-method approach was used in the article taking into consideration the inter-disciplinary nature of the research involving both NLP and journalism domains.

3.1 Data collection and dataset description

A third-party API was used to gather sports-related tweets from X over a one-year span from January 2025 to December 2025. A structured query was used to collect sports-related tweets that were likely to contain offensive or abusive discourse, rather than to directly classify tweets as offensive. The query combined sports-related keywords (e.g., football, team, referee, major tournaments, and prominent players) with keywords associated with offensive, abusive, or aggressive language (e.g., profanity, corruption charges, racial remarks, and incitements to violence). Only English-language tweets were gathered. The complete search query is provided in Appendix A for clarity and reproducibility. The initial dataset consisted of 7,412 tweets. After data cleaning, which included text normalization and deduplication, the final dataset had 7,279 distinct tweets.

A random subset of 350 tweets was manually annotated into two categories: *offensive* and *normal*. Of these, 208 tweets were labeled as normal and 142 as offensive. Tweets were classified as abusive when they included direct insults, offensive or derogatory language, discriminatory or racist comments, profanity, or hostile attacks targeting players, referees, teams, journalists, or fan groups. User information was anonymized to ensure ethical compliance, and only publicly available data were collected and analyzed. The classification of tweets as offensive or non-offensive is designed to boost the accuracy and credibility of comment selection, because it is aimed at retrieving sports discourse that is likely to contain offensive terms, rather than as a labeling mechanism. However, this does not mean that all the tweets retrieved are offensive; in certain circumstances, racial phrases could occur or be used in an abusive way, but they might be described in a neutral or critical way. In addition, the inclusion of racist terms does not necessarily mean that the tweets we retrieved are offensive language or hate speech, and such terms are sometimes utilized in a neutral or critical way. Additionally, all the potentially offending tweets are context-based, not just keyword-based. Therefore, manual classification was required to separate offensive tweets from the others.

3.2 Dataset analysis and classification approach

3.2.1 Logistic regression classifier

We split the manually labeled subset into training and test sets using a 70/30 split. We extracted text features using the TF-IDF vectorization technique, capturing unigrams and bigrams. We trained a logistic regression classifier as a baseline model because of its interpretability and suitability for journalism-oriented research.

The study applied a traditional machine learning approach. This method was suitable given the relatively small size of the classified dataset and enabled the use of an interpretable, computationally efficient, and reproducible binary classification method for the investigation. Moreover, the major purpose was to categorize the tweets into two classes, namely, offensive and normal, and not to investigate the intrinsic thematic patterns using theme modeling.

The model achieved an overall accuracy of 0.71. Class-wise performance showed an F1-score of 0.76 for normal tweets and 0.64 for offensive tweets. Given the limited size of the annotated dataset and the subtlety of offensive sports discourse, these results are to be expected.

To expand labeling coverage, a weakly supervised approach was employed. The trained model was applied to the remaining unlabeled tweets, and only predictions with a confidence score of 0.6 or higher were retained. This process resulted in 2,042 confidently labeled tweets, consisting of 1,381 normal and 661 offensive tweets as shown in the Figure 1.

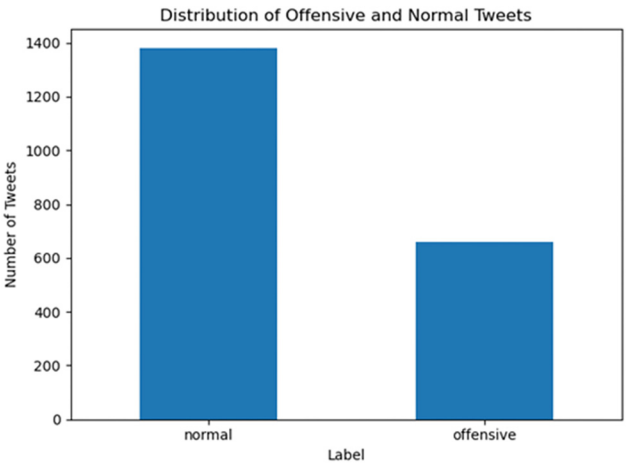


Figure 1: Distribution of offensive and normal tweets.

To improve label reliability, a manual spot-check was conducted on a subset of 200 high-confidence predictions. The review showed a strong alignment between model predictions and human judgment, supporting the credibility of the weakly supervised labeling process.

While the quantitative data allowed us to answer the first part of the research question regarding the difference in user interaction between abusive sports tweets and regular tweets on the X platform. Thereafter, we identified the main types of tweets that were most interactive and important, emerging from Twitter discourse, through text analysis of abusive and ordinary tweets, retweets, and interactions, which we illustrate in the text classification system.

4 Results and discussion: mapping offensive discourse in sports communication

4.1 Quantitative analysis of offensive discourse patterns

4.1.1 Classifier analysis

This section presents the quantitative results obtained from the implementation of an NLP-based weakly supervised classification model applied to sports-related X data. After training the model and expanding the labels based on confidence, a group of 2,042 tweets with high-confidence predictions was used for statistical analysis. The results show the user's verification status, how engaged they are, what kind of interactions a tweet gets, and the lexical patterns that go along with offensive language.

4.1.1.1 Engagement metrics: normal versus offensive tweets

We calculated the average engagement metrics for tweets that were either *normal* or *offensive* to see how people reacted to different kinds of content. The results are shown in Table 1.

Table 1: Engagement metrics.

Tweet type	Avg. likes	Avg. retweets	Avg. views	Avg. replies
Normal	40.26	7.58	1,680.38	1.53
Offensive	78.29	12.80	4,467.10	3.25

The results indicate that offensive tweets showed higher average engagement levels across all measured metrics. Offensive tweets had higher average likes, retweets, views, and replies compared to normal tweets. The table also shows that offensive content received almost twice as many likes and retweets as other content. However, an independent samples *t*-test conducted to compare engagement levels between offensive and normal tweets did not reveal a statistically significant difference between the two groups ($p > 0.05$). Therefore, these findings should be interpreted cautiously.

Similarly, the table shows that offensive tweets get more than 2.5 times as many views as regular tweets. Also, offensive tweets get more than twice as many replies, which may reflect higher interaction tendencies in digital sports communication. Figure 2 shows the bar plot with a logarithmic scale to show the highly skewed engagement distribution and to make small differences clear.

From the point of view of NLP and computational analysis, these results show that using aggressive or emotionally charged language makes people more visible and involved on social media.

4.1.1.2 Verification status of users posting offensive content

We further looked at the distribution of offensive and normal tweets based on user verification status.

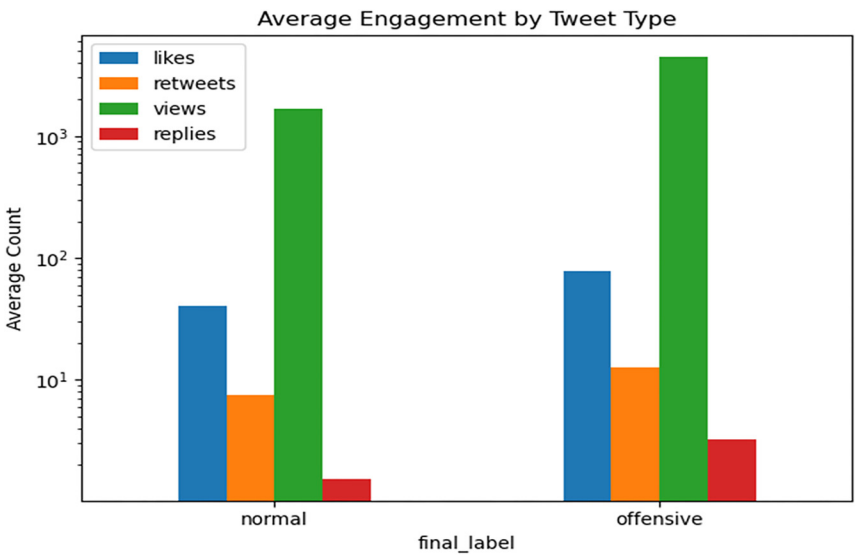


Figure 2: Average engagement by tweet types.

The analysis shows that most offensive tweets come from accounts that aren't verified. Verified users only posted a small number of offensive tweets (63), while unverified users posted 598 offensive tweets. Table 2 and Figure 3 show that verified users, on the other hand, post a higher percentage of normal tweets.

Table 2: Verification status of users.

Tweet type	Unverified	Verified
Normal	984	397
Offensive	598	63

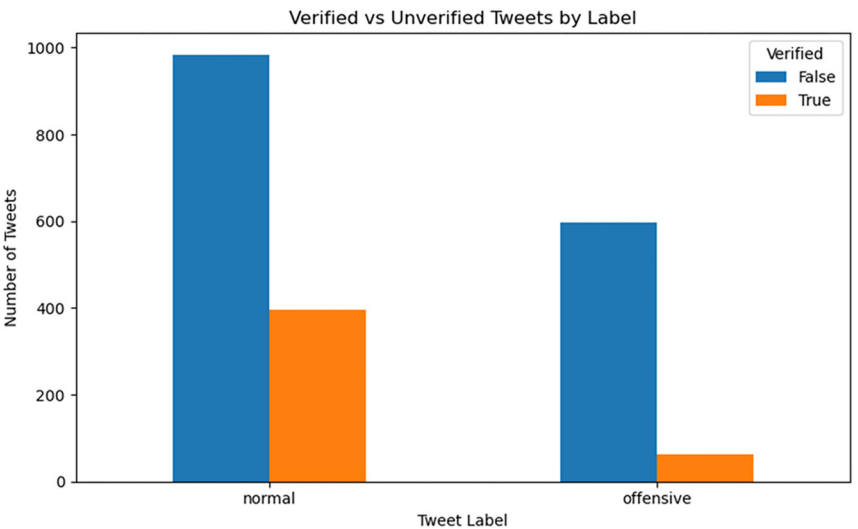


Figure 3: Verified versus unverified tweets by labels.

This quantitative pattern suggests that verified users are less likely to engage in offensive sports-related discourse, which may reflect reputational concerns or platform accountability mechanisms.

4.1.1.3 Offensive content in tweets versus replies

The dataset was also examined to distinguish between original tweets and replies. Table 3 shows that original tweets have more offensive content than replies, while

reply chains have more normal content. This means that people often use offensive language to start conversations instead of responding to them.

Figure 4 shows the results of this label-wise comparison between original tweets and replies.

From the point of view of AI and NLP, this difference is important because it shows how offensive language shows up differently in different types of conversations, which could help with future modeling.

4.1.1.4 Frequently used words in offensive tweets

We conducted a frequency-based lexical analysis on tweets that were classified as offensive. Table 4 lists the words that occur most commonly.

The vocabulary contains lots of explicit profanity (e.g., f**k, f**king, sh*t), combined with many sports-related terms (team, game, club, football, fans).This

Table 3: Offensive content in tweets versus replies.

Tweet type	Original tweets	Replies
Normal	518	863
Offensive	379	282

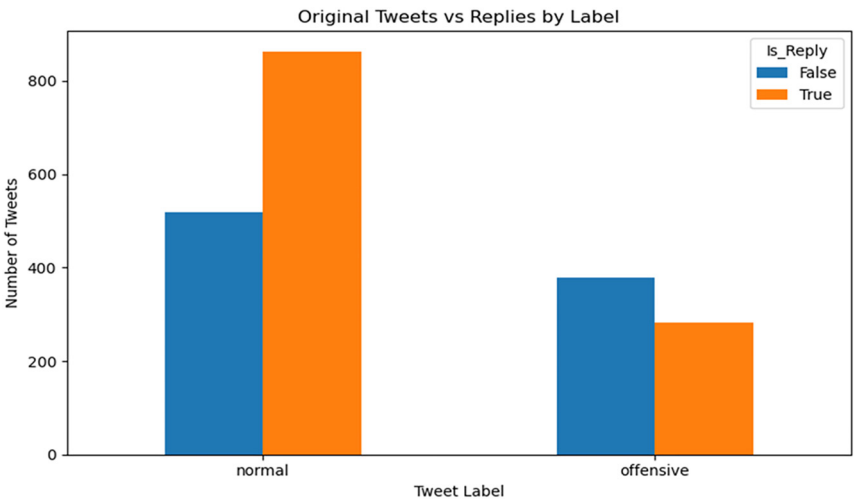


Figure 4: Comparison of original tweets versus replies by label.

Table 4: Words in offensive tweets.

Word	Count
F**k	568
team	283
game	239
F**king	122
hate	118
club	105
fans	98
football	76
shameful	72
Sh*t	71

lexical pattern suggests that offensive language in sports discussions often pertains to teams, players, and match outcomes, rather than being merely general abuse.

Also, emotionally charged words like “hate” and “shameful” show that moral judgment and emotional reactions are important parts of offensive sports communication.

4.2 Qualitative interpretation of offensive discourse patterns

4.2.1 Offensive discourse as high-engagement content

One of the most striking observations is the consistently higher engagements associated with offensive tweets. Offensive tweets attracted higher average levels of likes, retweets, views, and replies compared to normal sports discourse; however, the statistical analysis did not reveal a statistically significant difference between the two groups. From a journalistic standpoint, this suggests that emotionally charged and aggressive language functions as an engagement catalyst in online sports conversations. Following is one of the examples of tweets that has received higher engagement:

> f**k psg f**k champions league manchester city f**k pep f**k premier league f**k football f**k arteta f**k england f**k the referees f**k interpol f**k uganda police f**k uedcl f**k broke ladies f**k man united fans and f**k the ministry of transport. (User tweeted | 08/05/2025 12:01 AM)

Confirming the results we have indicated, the tweet above shows some offensive phrases that the user used to express his opinion, and it received thousands of likes

and views, while it received hundreds of retweets and replies. Fans' passion for sport, identity, and competition has always been a strong driving force, often creating intense emotions and encouraging the use of provocative language when reacting to events. The features of the X platform may also help explain these findings. Fans can react immediately during matches, and tweets can spread quickly. This may increase the visibility of offensive tweets and encourage more interaction, especially during emotional moments in sports.

Sports discussions are often more emotional and intense than many other news topics, making sports journalism a unique environment for studying fan interaction on the X platform. The X platform allows fans to communicate with each other during matches and share their opinions with a large audience in real time. This type of communication may contribute to more emotional discussions and help explain the higher engagement observed in offensive sports-related tweets.

4.2.2 The role of anonymity and accountability in offensive discourse

The dominance of unverified users among offensive content producers highlights the importance of perceived anonymity and reduced accountability in digital sports communication. Verified users, who are often journalists, athletes, public figures, or official organizations, contribute a small share of offensive content. A small proportion of the offending content is posted by verified users such journalists, athletes, public figures or official organisations. The study did not systematically follow individual users or account identities. Qualitative observations indicated that some of the most interactive offensive tweets were related to emotional fan feedback and rivalry-driven debates. Following is one of the examples of tweets from users whose accounts is unverified and have received a lot of interaction:

> if**king hate that f**king team makes the game so lame and boring f**k you (User tweeted | 11/12/2025 04:57)

> dear f**k you you had positioned yourself to be a mentor to a team of young men playing under the first black head coach at um so much opportunity for you and them but here we are selfish insecure liar i am so sad for your wife and your children f**k you (User tweeted | 11/12/2025 01:35 AM)

The results indicate that unverified users make the majority of offensive tweets and verified accounts are less likely to use offensive language in their tweets. Also from a professional ethics perspective, the results also suggest that reputation-conscious users are better able to control their emotions and exercise caution in their tweets. This difference supports the prevailing view in journalism that reduced

accountability and perceived anonymity may contribute to more hostile and lower-quality public discourse in digital environments.

The Online Disinhibition Effect contributes to further explaining these results. Suler (2004) claims that some social limitations seen in online environments may result in aggressive forms of communication, but that does not imply that accounts being used with offensive language are necessarily anonymous. Joinson (2007) argues that online disinhibition should not just be restricted to anonymity but also to the users' goals and the whole social milieu. Similarly, Lapidot-Lefler and Barak (2012) show that toxic online disinhibition can be influenced by numerous communicative contexts, not just anonymity. Thus, the higher proportion of offensive tweets by unverified users in the present study may be interpreted as a pattern that may be related to lower social constraints and a sense of accountability in online offensive sports discourse, without assuming unverified status as a proxy for anonymity.

4.2.3 Offensive language as a tool for initiating discourse

The results showed a high proportion of offensive language used in original tweets compared to replies, and accordingly, offensive content appears to be more engaging at the beginning of a conversation. Journalistically, this implies that offensive tweets function as attention-seeking entry points into the public discourse, designed to provoke reactions, debate, or confrontation. Following is an example of the use of offensive language in the original tweets:

> The Asian Elite Championship is a fantastic tournament, but it is shameful to choose such a disgraceful refereeing team. Uphold your reputation and the tournament's reputation by selecting those who deserve to officiate its matches with honor. (User tweeted | 24/11/2025 10:25 PM)

This indicates that the offensive speeches in the original tweets begin with words that may be opening words to provoke interaction through the use of inappropriate language, and that tweet gained thousands of views and a number of comments. In contrast, users' responses after the initial interaction tend to be more neutral and less hateful. Therefore, these results suggest that the beginning of the discussion is more engaging than its end.

4.3 Implications of the findings for sports journalism

Research that uses a mix method quantitative and qualitative approach in sports is necessary because it helps distinguish between abusive and non-abusive sports discourse measurably. It also helps explain the psychological, cultural, and

journalistic impacts of this discourse. This study focuses on sports journalism on the X platform, and this section explains the broader context behind the observed patterns in sports journalism and public discussion on X platform through the amalgamation of NLP-based machine learning and journalistic analysis. The research demonstrates the effectiveness of a weakly supervised classification framework on a medium-scale dataset. It demonstrates how computational methods can augment journalism research and reveal patterns that may be difficult to identify through manual analysis alone.

These results show that offensive content about sports consistently gets more likes, retweets, views, and replies than other types of content. These findings may initially imply a preference among the audience for aggressive or provocative content; however, a journalistic interpretation necessitates a more prudent analysis. More engagement doesn't always mean that someone agrees or supports something (Park et al. 2020). Instead, it might be a sign of controversy, anger, or conflict-driven interaction, which are all known factors that draw attention in digital media environments (Brady et al. 2017). Conversely, a robust correlation between unverified accounts and offensive language intensifies ongoing discussions regarding accountability in online communication. Verified users, who are often institutions, media organizations, or public figures, seem to follow professional and ethical communication rules more closely. This distinction demonstrates the importance of identity signaling in shaping discourse quality (Marwick and Boyd 2011).

Beyond the computational findings, these results also point to important implications for journalistic ethics and professional media practice. The spread of abusive sports-related discourse raises concerns about responsible coverage and content moderation. Therefore, media organizations and sports journalists need to adopt more robust approaches to prevent the perpetuation of abusive communication or the unintentional amplification of toxic narratives.

Compared with many other news topics, sports journalism is closely connected to fan identity, team loyalty, and rivalry. The X platform allows fans to react immediately during sporting events, and tweets can quickly reach a large audience. This may increase emotional discussions and highlight the importance of responsible reporting and content moderation. These findings suggest that sports journalists and media organizations should balance audience engagement with ethical communication practices, particularly during highly emotional sporting events.

The results indicate that verification status functions as both a technical indicator and a symbolic representation of accountability, influencing interpersonal communication in public sports environments. The results align with prior research concerning online hostility, emotional polarization, and algorithmic amplification, while providing a domain-specific focus on sports discourse. This research integrates

NLP techniques with journalistic analysis, bridging computational and interpretive methodologies and enhancing the domain of computational journalism.

However, from an NLP point of view, the quantitative analysis shows that:

a) Offensive sports related tweets receive higher average engagement than normal tweets.
b) Offensive content is predominantly produced by unverified users.
c) Offensive language is more common in original tweets than in replies.
d) Lexical patterns combine sports terminology with profanity and emotional language.

These results show that NLP-based classification works well in uncovering large-scale behavioral and linguistic patterns in social media sports conversation.

In this context, AI-based methods do not supplant traditional journalistic inquiry; rather, they function as analytical frameworks, enabling researchers to scrutinize extensive communicative dynamics while maintaining critical interpretive profundity.

In summary, the interpretation of the NLP driven findings mentioned above reveals that offensive sports discourse is usually increased through engagement mechanisms. Similarly, anonymity and lack of verification is strongly related with offensive expression. Additionally, offensive language often serves as an initiating strategy in online discussions.

This study illustrates how the integration of computational analysis and journalistic interpretation can enhance the comprehension of digital sports communication beyond mere technical metrics through the use of NLP-assisted methods.

4.3.1 Linking the findings to the research framework

This paper's findings highlight the importance of examining fan-generated tweets and sports-related interactions through the lens of Natural Language Processing (NLP) and may provide new methodological pathways for research in sports journalism. Social media platforms have become significant digital spaces for direct interaction between athletes and supporters, but they also expose athletes to various forms of cyber-aggression (Vujičić Stanković and Mladenović 2023). In this sense, this study contributes to identifying patterns of offensive language and dynamics of user engagement, thereby supporting more effective ways to understand and handle abusive communication in digital sports contexts. In addition, X platform offers a fast and large-scale communication environment that promotes the rapid dissemination of public discourse across geographical borders (Hasfi et al. 2021). In this context, Bandyopadhyay and Karmakar (2024) indicated that natural language processing-based models are effective in evaluating large-scale sports-related tweets and

achieved high accuracy in sentiment categorization. NLP-based techniques offer effective means for detecting, recognizing and categorizing objectionable discourse at large scale (Alqahtani and Ilyas 2022). Sports communication between fans is shaped by emotion, competition, loyalty, culture, and identity; therefore, the emergence and amplification of abusive discourse is critical in digital environments. Therefore, the findings show that the model applied in this paper has achieved good performance in recognizing offensive discourse patterns and identifying the user interaction dynamics associated with the propagation of abusive speech among sports fans on the X platform.

5 Conclusions

In this paper, we have deployed a mix-method approach by leveraging NLP based techniques and journalism research to examined offensive sports discourse by analyzing X platform tweets published in 2025. The tweets were analyzed firstly using manual annotation and then by training a machine learning classifier using TF-IDF features and logistic regression.

Our quantitative analysis showed that offensive tweets showed higher average engagement levels compared to unoffensive tweets in terms of likes, views and retweets and replies. Our findings also show that unverified users usually drive the offensive discussions. On the other hand, verified accounts usually refrain from initiating offensive conversations, meaning the verified status appears to function as an indicator of responsibility and professional reputation.

Additionally, the results show that offensive discussions are predominantly initiated in the original tweets compared to replies and retweets. This means that offensive discourse may function used as a strategy to initiate the conversation rather than merely responding to it.

Overall, this research assists the growing field of computational journalism by leveraging NLP techniques with journalism-based research in an interdisciplinary setting to gain broader insights about the domain-specific online sports discourse.

The findings of this study further highlight the unique characteristics of sports journalism in digital communication environments. Fans usually react to matches, players, and referees while events are still taking place, and these reactions are often influenced by team loyalty and rivalry. On the X platform, opinions can spread quickly, and emotional discussions may attract more attention and interaction. These characteristics help explain why offensive sports discourse develops differently from many other forms of online communication. Overall, this study demonstrates that combining natural language processing techniques with journalistic interpretation can improve our understanding of offensive sports discourse and may

support future research as well as more responsible media practices in digital sports environments.

The results of this study indicate that the Online Disinhibition Effect offers a valuable theoretical perspective to explain offensive behavior in online sports discourse. However, the study does not assume that the unverified status directly reflects anonymity. The higher proportion of offensive content originating from unverified users represents a pattern that can be understood in light of reduced social constraints and reduced perceived accountability in digital communication environments. Previous studies have demonstrated that anonymity alone cannot account for online offensive behavior. Instead, it should be understood in relation to a greater range of communicative contexts, user motivations, and the broader social environment of online engagement.

6 Limitations and future scope

Even though this study provides meaningful insights into offensive discourse relevant to sport on the X platform by applying NLP-based methods and journalistic analysis, we have to acknowledge several limitations. It is essential to recognize these limitations because it will help us to properly interpret the findings and guide the future research.

The dataset contains data from a limited time period i.e. from January 2025 until December 2025. This may limit the generalizability of the findings. Sports discourse is highly event-driven, and engagement patterns can differ significantly depending on tournaments, rivalries, or controversial incidents. It means the observed trends may not fully represent long-term or seasonal changes in sports-related online conversations.

Since the data collection strategy explicitly targeted tweets that were related to sport and that contained keywords that were offensive, abusive, or antagonistic, we should note that the resulting dataset is not representative of the full spectrum of sports discourse on X. Instead, it reflects a sample that is offense-leaning and which is designed to analyze patterns and characteristics of offensive communication. Accordingly, we should not interpret the findings in such as way as to reflect the overall prevalence of offensive language in sports-related tweets, but rather as identifying engagement dynamics, linguistic features, and user behaviors within a targeted subset of offensive-prone discourse.

We have employed a traditional machine learning model (logistic regression with TF-IDF features) in this study. The reason for choosing it was because of its

interpretability and suitability for journalism research. We must insist that such models may struggle to capture linguistic nuances such as sarcasm, irony, coded language, or contextual humor. These are features that are common in sports-related communication. Moreover, the manually annotated part was still a bit modest since manual annotation was labor-intensive and the study was of an exploratory nature. A larger annotated dataset may improve the classification performance and generalizability. Nevertheless, the annotation size was considered sufficient to construct a preliminary supervised framework for analyzing offensive sports speech.

In our future work, we might use models that are based on a transformer, or we may use hybrid approaches while maintaining interpretability through explainability techniques.

The binary classification of tweets into offensive and normal categories simplifies the complex picture of harmful speech. This research might not have fully captured subtle forms of harassment, implicit bias, or contextual insults.

In our future work, we might adopt annotation schemes that are multi class or hierarchical. These schemes are good in distinguishing between different types of offensive speech, such as insults, threats, hate speech, and abusive humor. Moreover, future research may benefit from larger manually annotated datasets and additional validation procedures to improve model robustness and strengthen the reliability of weakly supervised classification.

Research ethics: Not applicable.
Informed consent: Not applicable.
Author contributions: All authors have accepted responsibility for the entire content of this manuscript and approved its submission. Conceptualization, A.S., L.U.R.H., T.L., and M.G.; Methodology, L.U.R.H. and A.S; Data collection, A.S. and L.U.R.H.; Formal analysis, L.U.R.H. and A.S; Writing – original draft preparation, A.S. and L.U.R.H.; Writing – review and editing, A.S., L.U.R.H., T.L., and M.G.; Supervision, T.L. and M.G.
Use of Large Language Models, AI and Machine Learning Tools: Machine learning tools were used in this study for text classification and weakly supervised labeling. Specifically, TF-IDF was used for feature extraction, and logistic regression was applied to classify offensive and non-offensive tweets.
Conflict of interest: The authors state no conflict of interest.
Research funding: None declared.
Data availability: Not applicable.

Appendix A

Data Collection Query

```
(football OR soccer OR match OR game OR team OR player OR coach OR fans
   OR “World Cup” OR FIFA OR “FIFA World Cup” OR “FIFA Arab Cup”
   OR championship OR loss OR “red card”
   OR Messi OR Ronaldo OR Mbappe
   OR “women sport”
   OR handball OR foul OR “VAR check”
   OR #football OR #soccer OR #WorldCup OR #FIFA OR #FIFAWorldCup
   OR #AFCON2025
)
AND
(
   “hate this club” OR “fake supporters”
   OR “referee bias” OR “worst referee”
   OR “referee error” OR “referee mistake”
   OR conspiracy OR corruption OR “fixed match”
   OR “cheating club” OR “dirty tactics” OR “unfair decision”
   OR shameful OR “overrated player”
   OR thug OR scum OR monkey OR “racist fans” OR “racist referee”
   OR “kill yourself” OR “f**k you” OR “f**k yourself”
   OR “piece of shit” OR shit
   OR “stolen win” OR injustice OR disgrace OR rude
   OR bribe OR “referee bribe”
   OR “boycott FIFA” OR genocide OR freedom
)
lang:en since:2025-01-01 until:2025-12-13
```

References

Ali, Waleed & Mohammed Hassoun. 2019. Artificial intelligence and automated journalism: Contemporary challenges and new opportunities. *International Journal of Media, Journalism and Mass Communications* 5(1). 40–49.

Alqahtani, Abdulkarim Faraj & Mohammad Ilyas. 2022. Using natural language processing to detect offensive text and cyberbullying in social media: A review. *Malaysian Journal of Information and Communication Technology* 7(2). 107–118.

Asiri, Areej & Mohammed Saleh. 2024. SOD: A corpus for saudi offensive language detection classification. *Computers* 13(8). 211.

Bandyopadhyay, Indranil & Rahul Karmakar. 2024. Deep learning-based sentiment analysis of olympics tweets. *arXiv:2407.12376*. https://doi.org/10.48550/arXiv.2407.12376.

Brady, William J., Julian A. Wills, John T. Jost, Joshua A. Tucker & Jay J. Van Bavel. 2017. Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences* 114(28). 7313–7318.

Broussard, Meredith, Nicholas Diakopoulos, Andrea L. Guzman, Rediet Abebe, Michel Dupagne & Ching-Hua Chuan. 2019. Artificial intelligence and journalism. *Journalism & Mass Communication Quarterly* 96(3). 673–695.

Burgers, Christian, Camiel J. Beukeboom, Pamela A. L. Smith & Tammie van Biemen. 2023. How live Twitter commentaries by professional sports clubs can reveal intergroup dynamics. *Computers in Human Behavior* 139. 107528.

Canavilhas, João. 2022. Artificial intelligence and journalism: Current situation and expectations in the Portuguese sports media. *Journalism and Media* 3(3). 510–520.

Curran, Kevin, Kieron O'Hara & Sean O'Brien. 2011. The role of Twitter in the world of business. *International Journal of Business Data Communications and Networking* 7(3). 1–15.

Deuze, Mark. 2005. What is journalism? Professional identity and ideology of journalists reconsidered. *Journalism* 6(4). 442–464.

Ghosh, Soujanya, Asif Ekbal, Pushpak Bhattacharyya, Sriparna Saha, Vaibhav Tyagi, Animesh Kumar & Naman Kumar. 2020. Annotated corpus of tweets in English from various domains for emotion detection. In *Proceedings of the 17th international conference on natural language processing (ICON)*, 460–469. Available at: https://aclanthology.org/2020.icon-main.62/.

Gibbs, Chris & Richard Haynes. 2013. A phenomenological investigation into how Twitter has changed the nature of sport media relations. *International Journal of Sport Communication* 6(4). 394–408.

Hasfi, Nurul, Micah R. Fisher & Muhammad A. K. Sahide. 2021. Overlooking the victims: Civic engagement on Twitter during Indonesia's 2019 fire and haze disaster. *International Journal of Disaster Risk Reduction* 60. 102271.

Hasfi, N., V. Pednekar-Magal, L. Rimban, S. Sunarto & A. Yusriana. 2024. Digital dissent: How Indonesian and Filipino journalists converge with virtual communities to face state repression. *Global Media and Communication* 20(3). 329–356.

Holzinger, Andreas, Georg Langs, Helmut Denk, Kurt Zatloukal & Heimo Müller. 2019. Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 9(4). e1312.

Hussiny, Mohammad Ali, Mohammad Arif Payenda & Abdul Razaq Vahidi. 2023. SentiDariPers: Sentiment analysis of Dari-Persian tweets based on people's views and opinion. In *International conference on technologies and innovation*, 138–156. Cham: Springer Nature Switzerland.

Hutchins, Brett & David Rowe. 2012. Sport Beyond Television: The Internet, Digital Media and the Rise of Networked Media Sport.

Joinson, Adam N. 2007. Disinhibition and the internet. In Jayne Gackenbach (ed.), *Psychology and the internet: Intrapersonal, interpersonal, and transpersonal implications*, 2nd edn. 63–80. Burlington, MA: Academic Press.

Jost, John T., Pablo Barberá, Richard Bonneau, Melanie Langer, Megan Metzger, Jonathan Nagler, Joanna Sterling & Joshua A. Tucker. 2018. How social media facilitates political protest: Information, motivation, and social networks. *Political Psychology* 39(Suppl 1). 85–118.

Karki, Sejal, Siddharth Parasher, Pallav Divedi, Amit Aggarwal & Ankush Mittal. 2024. Comprehensive analysis of sports discourse: Sentiment dynamics, derogatory speech patterns, and topic modeling during FIFA 2018 world cup. *Research Square*. https://doi.org/10.21203/rs.3.rs-4755275/v1.

Kavanagh, Emma, Ian Jones & Lucy Sheppard-Marks. 2016. Towards typologies of virtual maltreatment: Sport, digital cultures & dark leisure. *Leisure Studies* 35(6). 783–796.

Kearns, Colm, Gary Sinclair, Jack Black, Mark Doidge, Thomas Fletcher, Daniel Kilvington, Katie Liston, Theo Lynn & Pierangelo Rosati. 2022. A scoping review of research on online hate and sport. *Communication & Sport* 11(2). 402–430.

Khurana, Diksha, Aditya Koli, Kiran Khatter & Sukhdev Singh. 2023. Natural language processing: State of the art, current trends and challenges. *Multimedia Tools and Applications* 82(3). 3713–3744.

Klutse, E. K., S. Nuamah-Amoabeng, H. Lyu & J. Luo. 2023. Dismantling hate: Understanding hate speech trends against NBA athletes. In R. Thomson, et al. (eds.), *Social, cultural, and behavioral modeling, lecture notes in computer science*, vol. 14161. Cham: Springer.

Kothari, Ammina & Sally Ann Cruikshank. 2022. Artificial intelligence and journalism: An agenda for journalism research in Africa. *African Journalism Studies* 43(1). 17–33.

Krasadakis, Panteleimon, Evangelos Sakkopoulos & Vassilios S. Verykios. 2024. A survey on challenges and advances in natural language processing with a focus on legal informatics and low-resource languages. *Electronics* 13(3). 648.

Kurniawan, S. T. 2024. Amplification of artificial intelligence in 2024 election news in online media in Indonesia. *Journal of Politica Governo* 1(4). 1–19.

Lapidot-Lefler, Noam & Azy Barak. 2012. Effects of anonymity, invisibility, and lack of eye-contact on toxic online disinhibition. *Computers in Human Behavior* 28(2). 434–443.

Lewis, Seth C. & Nikki Usher. 2016. Trading zones, boundary objects, and the pursuit of news innovation: A case study of journalists and programmers. *Convergence: The International Journal of Research into New Media Technologies* 22(5). 543–560.

Marwick, Alice E. & danah boyd. 2011. I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media & Society* 13(1). 114–133.

Mathew, Binny, Ritam Dutt, Pawan Goyal & Animesh Mukherjee. 2019. Spread of hate speech in online social media. *Proceedings of the 10th ACM conference on web science*, 173–182. New York, NY: ACM.

McCarthy, J., J. Rowley & B. J. Keegan. 2022. Social media marketing strategy in English football clubs. *Soccer and Society* 23(4–5). 513–528.

Moran, Rachel E. & Sonia Jawaid Shaikh. 2022. Robots in the news and newsrooms: Unpacking meta-journalistic discourse on the use of artificial intelligence in journalism. *Digital Journalism* 10(10). 1756–1774.

Muindi, Benjamin. 2018. Negotiating the balance between speed and credibility in deploying Twitter as journalistic tool at the daily nation newspaper in Kenya. *African Journalism Studies* 39(1). 111–128.

Mutz, Michael & Markus Gerke. 2018. Major sporting events and national identification: The moderating effect of emotional involvement and the role of the media. *Communication & Sport* 6(5). 605–626.

Mutz, Michael & Markus Gerke. 2021. Media presentations of olympic victories and nation-related identification among viewers: The influence of emotions induced by sportscasts. *International Journal of Sport Communication* 15(2). 117–126.

Palanimurugan, Priya & V. Shanthi. 2025. The Rise of AI in Journalism: Automation, Ethics, and Personalization in News Media.

Pang, Bo & Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval* 2(1–2). 1–135.

Park, Sora, Caroline Fisher, Terry Flew & Uwe Dulleck. 2020. Global mistrust in news: The impact of social media on trust. *The International Journal on Media Management* 22(2). 83–96.

Perreault, Gregory & Travis R. Bell. 2022. Towards a 'digital' sports journalism: Field theory, changing boundaries and evolving technologies. *Communication & Sport* 10(3). 398–416.

Popek, Adam. 2024. The role of social media in increasing interest in sport and physical activity. *International Journal of Arts, Humanities & Social Sciences* 2(6). 97–106.

Rua Gámez, Stefany. 2023. « The spread of hate speech on Twitter ». *Revista Kribí* 1(diciembre). 44–61. Available at: https://revistas.unicartagena.edu.co/index.php/Kribi/article/view/5281.

Sarkar, Soham, Tadisetty Sai Yashwanth & Animesh Giri. 2024. Advancing cricket narratives: AI-enhanced advanced journaling in the IPL using language models. *2024 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, 1–6. Piscataway, NJ: IEEE.

Shbair, Ahmed A. O. & Nurul Hasfi. 2022. The role of media television to emerge patriotism during conflict: Evidence from Palestine TV. *al-Irsyad: Journal of Islamic and Contemporary Issues* 7(1). 819–830.

Shehab, Mohammad, Laith Abualigah, Muath Ibrahim Jarrah, Osama Ahmad Alomari & Mohammad Sh. Daoud. 2020. (AIAM2019) Artificial intelligence in software engineering and inverse: Review. *International Journal of Computer Integrated Manufacturing* 33(10–11). 1129–1144.

Srivastava, Anurag & Kumar Mangalam Singh. 2025. Sports journalism and its psychological effects shaping a new trend of sports culture: A systematic review. *International Journal of Physical Education, Sports and Health* 12(4). 364–369.

Stone, Peter, Rodney Brooks, Erik Brynjolfsson, Ryan Calo, Oren Etzioni, Greg Hager, Julia Hirschberg, Shivaram Kalyanakrishnan, Ece Kamar, Sarit Kraus, Kevin Leyton-Brown, David Parkes, William Press, AnnaLee Saxenian, Julie Shah, Milind Tambe & Astro Teller. 2022. Artificial intelligence and life in 2030: The one hundred year study on artificial intelligence. *arXiv:2211.06318*. https://doi.org/10.48550/arXiv.2211.06318.

Stray, Jonathan. 2019. Making artificial intelligence work for investigative journalism. *Digital Journalism* 7(8). 1076–1097.

Suler, John. 2004. The online disinhibition effect. *CyberPsychology and Behavior* 7(3). 321–326.

Tesse, Alessio. 2024. AI and Sport Journalism. Available at: https://hdl.handle.net/20.500.14247/23832.

Usher, Nikki. 2017. Venture-backed news startups and the field of journalism: Challenges, changes, and consistencies. *Digital Journalism* 5(9). 1116–1133.

Uth, Bernadette. 2025. What does the audience expect when interacting with journalists? A Q-Sort study. *Journalism Practice* 1–19. https://doi.org/10.1080/17512786.2025.2551986.

Vujičić Stanković, Staša & Miljana Mladenović. 2023. An approach to automatic classification of hate speech in sports domain on social media. *Journal of Big Data* 10(1). 109.

Walker, Margath, Jamie Winders & Emmanuel Frimpong Boamah. 2021. Locating artificial intelligence: A research agenda. *Space and Polity* 25(2). 202–219.

Zhao, Tong, Chunping Li, Mengya Li, Qiang Ding & Li Li. 2013. Social recommendation incorporating topic mining and social trust analysis. *Proceedings of the 22nd ACM international conference on information & knowledge management*, 1643–1648. New York, NY: Association for Computing Machinery (ACM).