Gender Biases in Online Communication: A Case Study of Soccer

Mariana Macedo[1,2†] and Akrati Saxena[2*†]

[1]Data Science Department, Northeastern University London, United Kingdom.

[2]Khoury College of Computer Sciences, Northeastern University, Boston, United States.

[3]Leiden Institute of Advanced Computer Science, Leiden University, Niels Bohrweg 1, Leiden, 2333CA, South Holland, The Netherlands.

*Corresponding author(s). E-mail(s): a.saxena@liacs.leidenuniv.nl;

Contributing authors: mariana.macedo@nulondon.ac.uk;

†These authors contributed equally to this work.

Abstract

Social media and digital platforms allow us to freely and easily express our opinions to a wide audience. In this study, we investigate gender-based differences in online communication, specifically on Twitter (now X), in the context of soccer, by analyzing patterns of participation, sentiment, and engagement across male and female users. As one of the most popular sports, soccer engages a diverse audience on social media, regardless of expertise. We collected 9.5 million tweets related to soccer in English and Portuguese during three months (March - June 2022). Only 18.38% tweets were identified as written by women, highlighting a possible gender gap already in the number of people who participated actively in this topic. We analyze tweets in Portuguese and English, as these languages represent different perspectives on soccer. English serves as a proxy for global discussions, while Portuguese reflects more localized but deeply engaged communities. We observe that women and men communicate more between each other in Portuguese than in English, exhibiting lower homophily within their social networks. However, this difference in homophily does not appear to influence

how women and men express emotions and sentiments, suggesting that these aspects may be shaped by other factors such as societal expectations, gender role socialization, online community dynamics, visibility constraints, or other gender-related norms and characteristics. Women express their emotions more intensely in response to events than men, regardless of the differences in the number of tweets, and men tend to be more negative in their tweets than women. Our study reveals persistent gender gaps across both weeks and hours through qualitative and quantitative analyses, including detailed text-level and network-level examinations. These findings underscore the importance of identifying and reporting gender disparities in online communication to foster more inclusive spaces where individuals can freely express their opinions.

Keywords: Gender Biases, Social Media, Online Communication, Homophily

1 Introduction

Social media platforms, such as Twitter[1] (currently known as X), Instagram, and Reddit, have emerged as significant channels for self-expression. The versatility, convenience, and widespread accessibility of social media encourage people to express and garner reactions from their targeted audience. Fans, sports organizations, and players heavily rely on these platforms for interactive communication and real-time information about ongoing events [1, 2]. It is also likely that breaking news comes first on Twitter than traditional media channels, providing an excellent medium to access instantaneous information from official as well as unofficial sources [2].

However, on social media, it is often observed that some influential groups – commonly majorities – dominate the communication, and as a consequence, the minority or less influential groups do not feel as free to express [3]. Past research has explored gender-related differences in offline as well as online communication spaces, such as Twitter [4, 5], Stack Overflow [6], Wikipedia [7], and YouTube [8]. For instance, Messias et al. [9] showed that, in general, white men are more likely to be followed on social media and acquire higher positions in rankings and ratings on Twitter; only after the top 14% most followed users, there are higher fractions of women than of men. This can also be translated to specific topics (e.g. Sports, Politics, News, Technology, Industry, and Art) [10, 11] where women tend to be less influential as well as hold lower-ranking/percentage in top influential females. Thus, thoughts and narratives of a particular group (e.g. males) linger more on Twitter and are more likely to get traction probably due to this participation imbalance, gender biases and constructs [5]. Many social, contextual, and developmental factors play a role in perpetuating biases and stereotyping in society that is manifested on social media [12–14], and the only way to change this wave is to encourage better online environments and healthier conversations.

[1] In this work, we will refer to it as "Twitter" due to the following two reasons: during the data collection process, the platform was known as Twitter and better readability.

In social networks, people prefer to connect with similar people, known as the homophily effect. Homophily influences their communication patterns and the information dissemination among them [15, 16]. This homophily can emerge from similarities emerging from individual or collective characteristics and ideologies. For instance, on Twitter, users also have topical homophily where people having an interest in similar topics are more likely to follow and interact with each other [17]. However, popular and influential people who share similar statuses or who post about the same topic more frequently are less likely to retweet each other's tweets [18]. Thus, people tend to interact with people who share similar beliefs and backgrounds, but popularity plays a different effect, a rich-get-richer effect[19], where minorities in a topic disappear in the middle of influential and majority groups.

The impact of gender on topical homophily has not yet been fully explored. We are interested here in analyzing the gender gap in online communication. We focus our analysis on a topic that is widely spoken worldwide: Soccer (or Football). In fact, the last World Cup was watched by around 5 billion people out of a world population of almost 8 billion [20], indicating the worldwide popularity and importance of the topic that drives billions of communities and dollars. There are several additional reasons for selecting soccer as the focus of this case study. Unlike specialized domains such as STEM or politics, where participation often requires domain-specific knowledge, soccer represents a more universally accessible topic. This accessibility encourages broader engagement and interaction among diverse users, making it a valuable lens through which to explore gendered patterns of communication. Furthermore, we specifically chose to analyze data in both Portuguese and English to capture variations in levels of engagement and popularity across different linguistic and cultural contexts.

In this study, we aim to address the following key research questions: Are there gender-based differences in communication patterns depending on the popularity and familiarity of a topic? Are there significant differences in how women and men represent themselves on social media, and do they respond differently? Do women's emotions and sentiments vary across languages when discussing similar topics? Finally, how do communication networks and the roles of women and men within them differ in networks shaped by different levels of topic familiarity, and how does this influence varying levels of homophily?

Over three months, we collected soccer-related tweets using the Twitter API, building a dataset of 7 million English tweets from 2 million unique users and 2.6 million Portuguese tweets from 500 thousand unique users. We specifically included Portuguese because, in Brazil and Portugal, soccer is more than just a sport — it is a cultural backbone. People grow up immersed in it, leading to deeper engagement, appreciation and more active participation compared to regions where soccer holds less prominence.

Women make up almost half of the population and 38% of the soccer fan base [21], and yet we only observed 18% of tweets were from women. Across our dataset, women were found to be much less likely to talk about soccer compared to men and exhibited more retweeting behavior. This can be an evidence of a starker gender gap. Through

our study, we investigate various ways in which men and women behave differently in terms of communicating, such as type of content, frequency of content, linguistic style, and tweeting behavior in a community. We conduct three levels of analysis: text-level, network-level, and event-level, to gain insights and answer the research questions. We observe that the Portuguese network has lower homophily than the English network, showing that more interactions between men and women might be related to the popularity of a topic, making it more inclusive. For both languages, we observed that women and men can display unhealthy comments and reactions and that both genders are likely to display intense emotions, but the intensity is dependent on different types of emotions. For instance, women are more likely to show joy and anticipation, and men are more likely to show fear and disgust.

Twitter can play a significant role in creating safe spaces and communities for women by providing platforms for communication, support, and advocacy [22, 23]. Women can use Twitter to connect with others who share similar experiences, challenges, and interests, which can provide a sense of belonging and support. Twitter enables women to amplify their voices and share their stories with a broader audience [24], which can influence other women to join the trend and be vocal about their thoughts, opinions, and interests.

Here, we aim to promote a healthier and more equitable online environment, ensuring that everyone can express themselves freely without fear of judgment. We look into the factors that make a space comfortable for everyone to express freely without fearing judgment. UNESCO launched a #ChangeTheGame campaign to encourage gender equality in sports as they believe that football can act as an effective medium to reduce the gender gap and empower women worldwide [25]. Analyzing the evolution of online communication can help develop interventions that foster safer digital spaces and encourage broader participation—not only in soccer-related discussions but across various topics. This might ignite a change from the grassroots level and amplify women's involvement in the game across all levels, right from players, fans, sports experts, coaches, and positions in governance organizations. While our current analysis is limited to binary genders due to privacy, ethical, and computational constraints, we recognize the importance of expanding future research to encompass a broader gender spectrum.

Our paper is structured into six sections. Next, we review related works, followed by a detailed description of our dataset. Then, we present our analysis and key findings. The paper concludes with a discussion and suggestions for future research.

2 Related Work

Certain topics, such as science, technology, or politics, require domain expertise, which may discourage some individuals from participating. To ensure a broad and inclusive analysis, we focused on soccer — a globally popular sport with cultural significance in countries like Brazil and Portugal. This section reviews research on topic-specific social media datasets, explores gender biases in online discussions, and examines how emotions and sentiments differ by gender.

2.1 Sports and Social Media

Twitter has changed sports media relations by shifting the landscape from a one-to-many model to a dynamic many-to-many interaction, reshaping media roles and engagement between journalists and stakeholders [26–28]. Gibbs and Haynes [27] interviewed sports media professionals in the U.S. and Canada, revealing that Twitter has disrupted traditional sports hierarchies and become the most influential social media platform for sports.

Sports are among the most discussed topics on social media, following entertainment and event-related news [29]. Fan loyalty plays a key role in Twitter discussions, reflecting a commitment to teams or players. Fans engage with sports organizations, news channels, and athletes for live updates and interaction [30]. This loyalty is driven by team attraction, trust, and involvement, developing over time [31]. Retweeting sports news is influenced primarily by personal interest, outweighing factors like tweet style, informativeness, originality, and source credibility [32]. Similarly, sports Twitter consumption is mainly driven by information, entertainment, time pass, and fanship rather than accessibility, economics, skills, or social constraints [33], offering valuable insights for sports marketing research.

Twitter's filtered tweet stream can automatically identify key events in live sports broadcasts [34]. Researchers have examined how athletes use social media to engage with followers and promote their teams or brands. Most athlete-related tweets focus on interactivity (34%), with fewer posts dedicated to team promotion (15%) or personal brand endorsements (5%) [35]. Studies highlight that athletes primarily use Twitter for direct engagement rather than self-promotion, while sport-for-development nonprofits utilize the platform to share information, engage followers, and drive offline support [36, 37]. However, Svensson et al. [36] found that while interactive communication is common, most organizations rely on one-way messaging, with calls to action being rare. The use of Twitter for information sharing, engagement, and mobilization varies across organizations, showing no clear link to revenue, age, social issues, or operational reach.

2.2 Gender Biases on Social Media

People often use social networking sites for self-promotion. However, a survey-based study showed that women prioritize relational uses (e.g., maintaining close ties), whereas men focus on accessing general information while using these platforms [38]. Some fields and topics, such as sports, STEM (Science, Technology, Engineering, and Mathematics), and politics, are more commonly associated with men [7, 24], which can lead to their increased participation in these conversations. Thus, in these contexts, women can feel more hesitant to talk about such topics and more inclined to certain professions due to gender norms and constructs. Skewed narratives and perspectives get reinforced as a consequence and get embedded as a subconscious behavior. To move against this reinforcement, we need to be aware of gender differences to promote more diversity, inclusivity and fairness in online platforms. Studying the population on Twitter can give us a proxy of different demographic circumstances, which has

been shown, for instance, for many aspects of Politics [39–41]. This makes it a useful tool for gathering insights that would typically require surveys, which can be really hard to conduct, time consuming and expensive.

Research shows that women are actively engaged in soccer-related discussions, but gender-based differences remain. Yoon et al. [42] found that women use Twitter primarily for entertainment, while Clavio et al. [43] showed that female fans contribute more to sports team feeds than men, relating to in-game updates, game results, individual player news, contests, giveaways, and ticket discounts. They also use social media more at stadiums and engage more with team Tweets. Grace and Mueller [44] studied NFL Twitter data and found that women were underrepresented in mentions despite comprising 46% of the fanbase. These studies call for better representation and more inclusive communication in male-dominated spaces like soccer. For instance, Hu and Kearney [24] found that women demonstrate stronger group cohesion, tend to promote their tweets more than directly engaging, and focus more on family-related topics than men. These findings suggest that social media platforms can play a role on amplifying the voices and visibility of marginalized groups, providing a safe public sphere or reducing inequalities in a long term. Therefore, it is crucial to further explore gender-driven communication patterns on social media and work towards enhancing the sense of community online.

Emotions and sentiments can serve as valuable indicators for understanding the sense of community and fostering healthy online environments. Researchers have explored the expression of sentiments [45] and emotions [46] with different genders using simple and complex techniques [45–47]. For instance, Thelwall et al. [48] analyzed MySpace data and observed that women give and receive more positive comments than men, though no gender difference was observed for negative comments. Interestingly, women often express more positive emotions than men, especially when interacting with other women [49]. An empirical analysis of a large web forum dataset revealed that women express opinions more subjectively than men and are more likely to convey both positive and negative emotions [50]. If we talk about STEM, users' response towards women is generally positive; however, there are spikes in positivity during events celebrating women's achievements, such as International Day of Women and Girls in Science and International Women's Day, highlighting the impact of such campaigns on people's opinion [51]. In academic discussions on Twitter, it has also been observed that women tend to express stronger emotions and sentiments than men in response to specific events [52].

Happiness levels were consistently higher in women, and their topics of discussion reflected greater positivity. Several studies indicate that women are more expressive on social media, particularly in expressing positive emotions [49, 53–55]. Garcia-Rudolph et al. [56] found that women exhibited significantly higher levels of positive emotions, while men expressed more negative emotions, based on Twitter data from stroke survivors. In the context of soccer, Bagić Babac and Podobnik [57] found that male and female fans expressed intense emotions like anger and fear similarly, but there were notable gender differences in expressing softer emotions like joy and sadness. Our study aligns with existing literature, which indicates that women tend to express

stronger emotions and sentiments than men. However, we demonstrate that these intense emotional displays are often triggered by specific events, whereas men tend to express more consistent sentiments over time (see Figure 3). Additionally, we expand on this by showing that both women and men can engage in toxic online behavior, highlighting the need for improved mechanisms to foster healthier conversations.

3 Soccer-Tweet Dataset

We collected data from Twitter over a span of three months (March 7 to June 5, 2022) using the following pattern: "#Soccer OR football OR futebol OR #soccer OR #football OR #futebol OR #futball OR #futbol OR #fifa OR #CBF OR #brasileirao OR #CopaIntelbrasDoBrasil OR #Libertadores OR #Sudamericana OR #CampeonatoDoBrasileiro OR #CopaAmerica OR #ConfedCup OR #UEFACup OR #Supercup OR #WorldCup OR #LaLiga OR #CopadelRey OR #CoppaItalia OR #CoupedeFrance OR #premierleague OR #championsleague OR #FACUP OR Confederations Cup OR #confederationscup OR #ConfedCup OR Copa del Rey OR EFL cup". With this collected data, we infer the gender based on the first name of the users using Genderize [58] and Namepedia [59]. After gender detection, we discard tweets whose users cannot be exclusively assigned a gender (male or female); this is a common limitation when using public data [60].

After filtering, our dataset contains 6,957,598 tweets in English (refereed by ***en***): 5,767,122 (82.89%) from male users and 1,190,476 (17.11%) from female users. Over a three-month period, we identified 1,556,818 (78.68%) unique male users and 421,996 (21.32%) unique female users. For Portuguese (***pt***), we collected 2,572,247 tweets, with 2,011,286 (78.19%) from male users and 560,961 (21.81%) from female users. The dataset included 365,045 (71.08%) unique male users and 148,539 (28.92%) unique female users. Thus, for both languages, men are the majority in terms of users and tweets overall and per week (shown in Figure 1). This imbalance in user participation on social media varies depending on the topic, with female users more active in some discussions on popular profiles or trends [61], and health topics [62], while male users dominate conversations on political issues [63] and cryptocurrency [64]. Our dataset is available at https://github.com/akratiiet/Soccer-Twitter-Dataset.

3.1 Replies and Retweets

We also consider the retweets and replies to examine whether they have a distinct pattern from the original posted tweets. We observed that in English, 68.91% of the female tweets were retweets compared to 60.25% for males, and in Portuguese, 59.07% of the female tweets were retweets compared to 55.08% for males. For the replies, we observe in English, 15.87%(23.21%) of female(male) tweets, and in Portuguese, 20.67%(37.87%) of female(male) tweets. We then observe that women tend to retweet more than men, and the opposite is observed for replies.

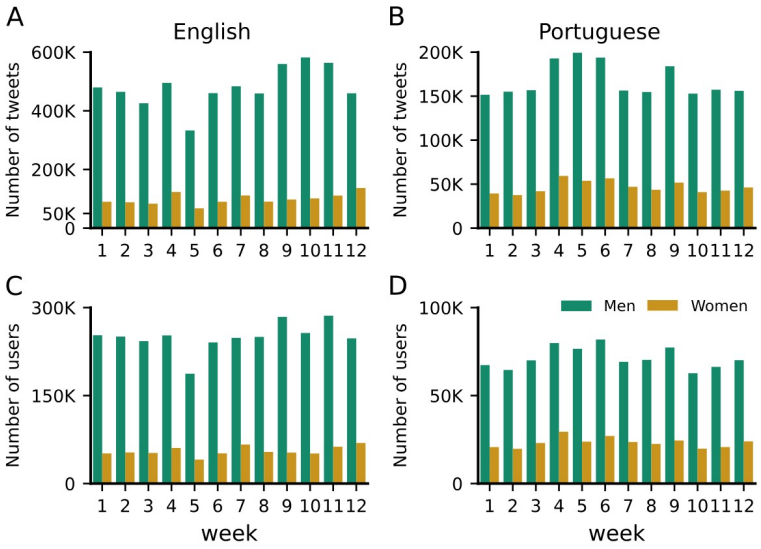


Fig. 1: Number of tweets and users per week disaggregated by gender. In our data, every week, 77-85% of tweets are from men, and 73-84% of users are men.

4 Experimental Analysis and Findings

We divide our analysis into two parts - analyzing the content of the tweets and analyzing the networks generated from the interaction between users through retweets and replies. We find consistent gender differences in online communication, and we highlight the need for a more inclusive and healthier online environment.

4.1 Text Analysis

Previous works have shown that women and men can express themselves differently through language, revealing distinct patterns in their communication styles [65, 66]. Women often use language that is more emotionally expressive, socially oriented, and warmer in tone. This means that women may tend to include more emotional words, focus on interpersonal relationships, and use language that conveys empathy and connection. In contrast, males typically use language that is more descriptive and impersonal, and prioritize conveying information in a straightforward and factual manner, focusing less on emotional expressions or social dynamics. In order to understand further how these disparities are translated in language usage for online platforms, we analyze tweets' text using emojis, hashtags, sentiments, emotions, and other linguistic attributes.

4.1.1 Emoji Usage Analysis

A picture is worth a thousand words. The emojis we use today in online communication are more than a tiny picture; they symbolize emotions. Emoji usage has increased rampantly so much that the Oxford Dictionary declared the emoji face with tears

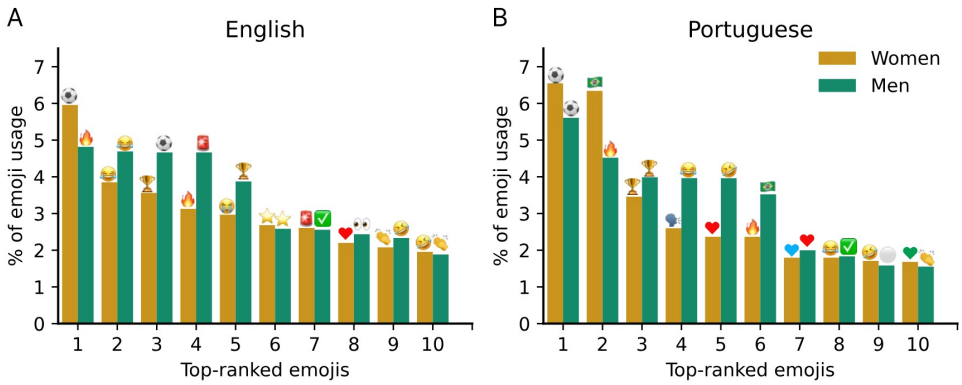


Fig. 2: Top 10 Emojis and their usage proportion by males and females

of joy (😂) as the word of the year for 2015 [67]. The scientific research community has also shown interest in analyzing emoji usage patterns in communication as they communicate an emotional key element of messages [68–71].

We use the Emot [72] library to extract emojis from the tweet text. After extraction, we analyze how emoji usage varies by the two genders. We found that in the English female tweet data around 32% tweets contained at least one emoji as compared to around 29.20% of the male tweets (Table 1). However, if we remove the retweets, quoted tweets, and replies, and only look at the original tweets, we find that 27.46% of tweets by females had emojis, while only 18.94% of original male tweets had emojis. The results indicate that emoji usage tends to be more prevalent among females in both languages (regardless of the differences in the percentage of emojis across languages where we observe higher levels for English in Figure 5). This finding aligns with the social norms [65, 66] and with the study by Chen et al.[71], where they report that female users include emojis in a larger proportion of text messages as compared to men.

We can also observe some differences in the top-ranked used emojis by each gender. Figure 2 shows the top 10 emojis and the percentage of tweets that used at least one particular emoji. This gives us insight into the type of emojis that are being used by males and females in the context of soccer communication. We see that for English tweets, the most widely used emoji by females is a ball of football, and for males, it is fire. For females, the same fire emoji is in the fourth position after the laugh and award emojis, which indicates a more optimistic perspective. For males, if we also read the tweets, we find more irony accompanied by the laugh. For Portuguese, we see that men and women use mostly the soccer ball, but still have the use of fire in second place for men. For Portuguese tweets, we see a high presence of the Brazilian flag for female tweets, indicating that this community might be big for Portuguese tweets. Comparing English and Portuguese tweets, we do see that for Portuguese, there is a more diverse set of emojis and the use of hearts, which aligns with the expected football passion for countries such as Portugal and Brazil.

Table 1: Statistics of tweet content shared by females and males in English and Portuguese. Values in parentheses represent averages including zeros, while values outside parentheses represent averages excluding zeros.

	Gender	Tweets	% with emoji	mean # of emojis	mean # hashtags	% of positive	% of negative	% of neutral	average sentiment
en	female	original	27.46%	2.41 (0.66)	2.69 (0.54)	20.88% (12.29%)	15.65% (5.60%)	82.11% (82.11%)	0.25 (0.18)
	female	all	32.01%	2.35 (0.75)	1.87 (0.43)	20.20% (11.72%)	15.63% (4.40%)	83.88% (83.88%)	0.28 (0.20)
	male	original	18.94%	2.32 (0.44)	2.65 (0.42)	19.94% (11.87%)	15.49% (5.84%)	82.28% (82.28%)	0.24 (0.17)
	male	all	29.20%	2.41 (0.70)	1.84 (0.37)	20.22% (11.63%)	15.42% (4.44%)	83.93% (83.93%)	0.29 (0.20)
	all	original	20.12%	2.34 (0.47)	2.66 (0.44)	20.07% (11.93%)	15.51% (5.81%)	82.26% (82.26%)	0.24 (0.17)
	all	all	29.68%	2.40 (0.71)	1.85 (0.38)	20.22% (11.65%)	15.46% (4.43%)	83.92% (83.92%)	0.29 (0.20)
pt	female	original	14.53%	4.10 (0.60)	2.54 (0.10)	25.43% (11.45%)	22.11% (7.64%)	76.22% (59.88%)	0.13 (0.08)
	female	all	22.59%	3.11 (0.70)	1.72 (0.13)	24.40% (10.35%)	20.81% (5.92%)	79.14% (61.10%)	0.18 (0.10)
	male	original	9.70%	3.22 (0.31)	2.20 (0.11)	23.09% (13.27%)	20.83% (9.77%)	77.35% (76.95%)	0.10 (0.08)
	male	all	18.03%	2.77 (0.50)	1.70 (0.15)	22.61% (12.52%)	20.12% (7.96%)	79.75% (79.51%)	0.15 (0.10)
	all	original	10.56%	3.43 (0.36)	2.25 (0.10)	23.44% (12.94%)	21.01% (9.39%)	77.19% (73.90%)	0.10 (0.08)
	all	all	19.06%	2.86 (0.55)	1.71 (0.15)	22.94% (12.03%)	20.24% (7.50%)	79.63% (75.35%)	0.15 (0.10)

Emojis are used further in sentiment and emotion analysis. We keep the emojis in their raw format while computing the sentiment using VADER API [73] as the model is capable of handling emojis. For emotion analysis, the emojis were converted to their equivalent emotion text so that they could be captured by the NRC-Emotion Intensity Lexicon API [74].

4.1.2 Hashtag Usage

Hashtags are a pivotal part of communication on Twitter and help people follow the content related to a topic easily using its specific hashtags [75]. In our dataset, from 7.53 million tweets, there are almost 3 million unique hashtags, and 20.77% of the tweets had at least one hashtag. 23.12% of the tweets with at least one hashtag were from females, as compared to 20.28% of tweets from males. This shows marginally higher hashtag usage behavior by women as compared to men (aligned with the high temporal variance for the percentage of hashtags across weeks in Figure5). However, both women and men are more likely to use one or two hashtags per tweet than several hashtags. The average number of hashtags across genders is similar; for females(males), it is 1.87(1.84) in English and 1.72(1.70) in Portuguese. In Figures 3-4, we observe

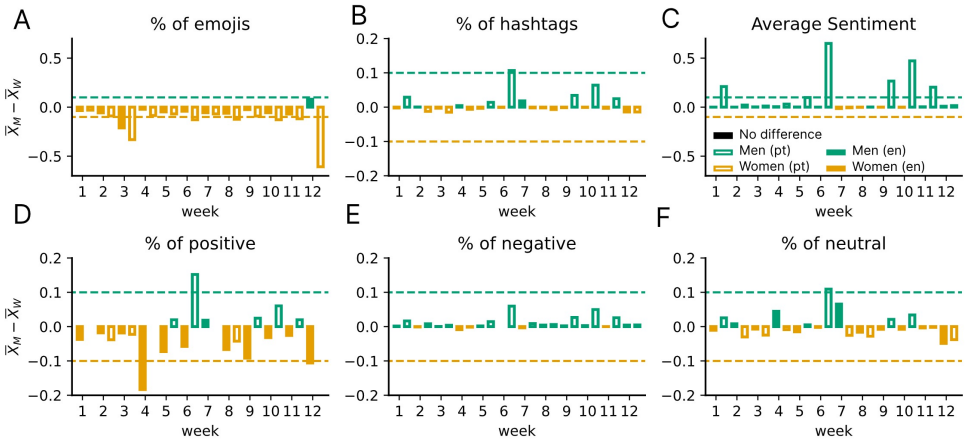


Fig. 3: Gender differences extracted from the content of the tweets per week. **A** Women tend to include a higher number of emojis over time. **B** There is no gender dominance in the use of hashtags over time. **C** Men tend to have higher average sentiment in comparison to women as they are more consistent. **D-F** Women have a higher percentage of positive tweets, and women post more neutral content over time. Men tend to share more negative content.

that gender dominance fluctuates weekly and hourly. Gender differences are calculated using a bootstrap technique, where 80% of the sample is resampled 1,000 times to compute average values. For each distribution (men and women), we calculate the difference in averages and perform a t-test to assess whether the averages are statistically different. We claim no significant difference between the groups only when the p-value exceeds 0.001. Women generally use more emojis and post a higher percentage of positive tweets, whereas men show greater use of hashtags and a higher proportion of negative tweets. On average, men express sentiment more frequently than women; however, as discussed in the following sections, women tend to display more intense, abrupt, short-lived emotional expressions, often linked to specific events, as illustrated in Figure 6.

4.1.3 Sentiment Analysis

Analyzing sentiments from content helps us to understand public opinions and preferences at a large scale for different topics, such as events, policies, and laws [76–78]. We use VADER (Valence Aware Dictionary for Sentiment Reasoning) API [73] to study how the tweets by men and women differ sentimentally. VADER is widely used for social network data due to its ability to handle texts with emojis, emoticons, extra punctuations, negations, use of contractions, randomized capitalization, use of degree modifiers, sentiment-laden slang words, and acronyms. VADER is a lexicon and rule-based tool to get sentiment scores across categories 'Positive' (compound score $\geq$ 0.05), 'Negative' (compound score $\leq$ -0.05) and 'Neutral' (compound score between

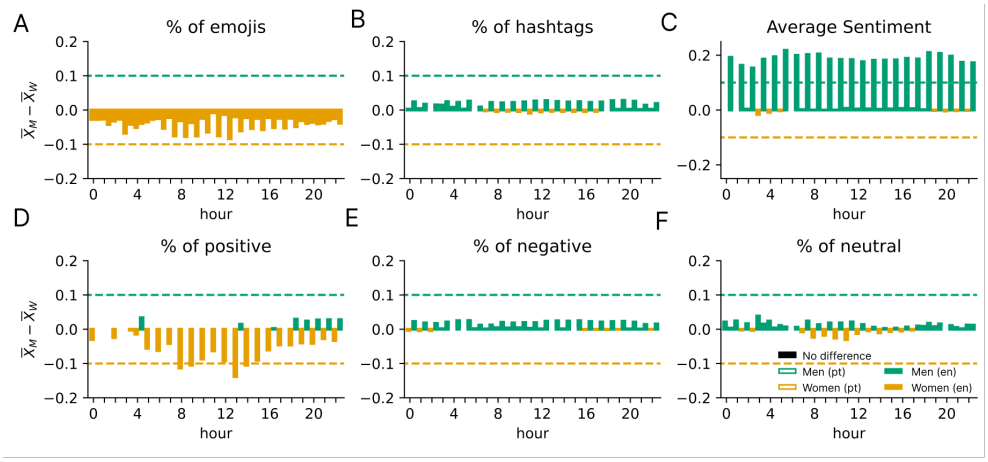


Fig. 4: Gender differences extracted from the content of the tweets per hour. **A** Women tend to include a higher number of emojis regardless of the time of the day. **B** Men are more likely to use hashtags hourly. **C** Men tend to have higher average sentiment in comparison to women. **D-F** Women have a higher percentage of positive tweets, and men tend to share more negative content.

-0.05 and 0.05) [73]. Therefore, each tweet is classified into one of the three categories: positive, negative and neutral.

Overall, the average sentiment scores of both genders are very similar (Table 1), but when we disaggregate the tweets per week, we observe that women post more positive tweets and also more neutral (Figure 3); for negative tweets, we did not observe any dominance. We do not observe any alignment with the temporal percentages across languages where English tend to have more positive and neutral tweets than Portuguese (Figure 5).

We further study how different genders show sentiments for events, and the results are interesting. We show event-based analysis for English in Figure 6. We observe that females tend to show more intense sentiments than males when some events happen. For instance, we highlight events corresponding to peaks of positive and negative sentiments, and the same events can also be connected with emotion analysis. The events corresponding to intense positive emotion in English include - *Real Madrid won the La Liga for the 35th time and Champion League final won by Real Madrid*, and for negative sentiments, they include - *Algeria Cameroon Match Controversy, Dwane Haskins, Americal Pittsburgh Steelers quarterback died in an accident and Ukraine lost making way for Wales in the world cup*. Similar results were observed for Portuguese.

4.1.4 Emotion Analysis

Figure 7 shows the word clouds of the most frequently used words to convey emotions. Females tend to use the words “Good”, “Best”, and “Love” more frequently than

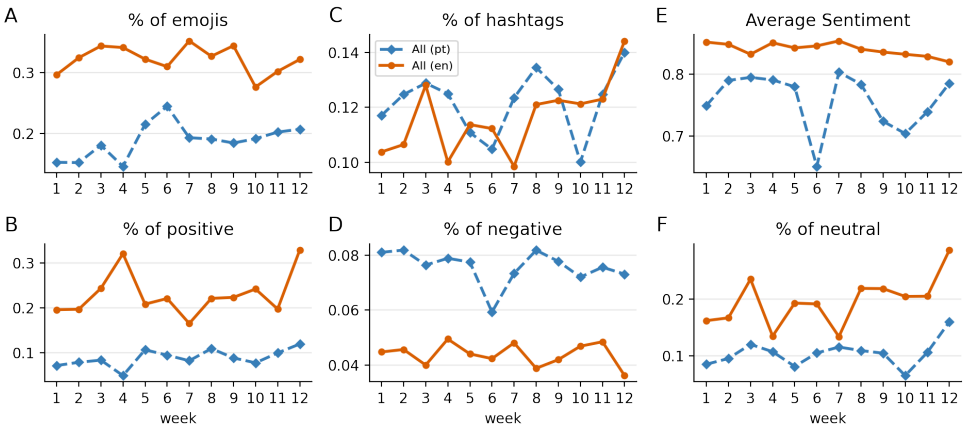
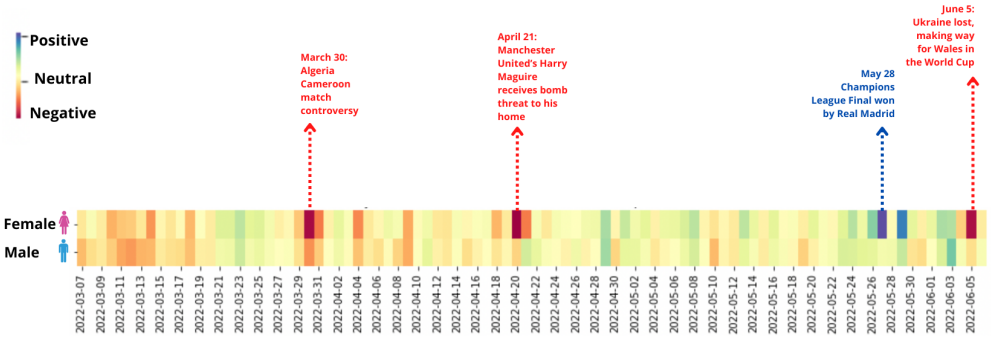


Fig. 5: Statistics from the content of the tweets. **A** Tweets in English tend to include more emojis than those in Portuguese. **B** The use of hashtags varies across time and language. **C** The average sentiment for English tweets tends to be higher than for Portuguese tweets. **D**–**F** English tweets are generally more positive than their Portuguese counterparts.

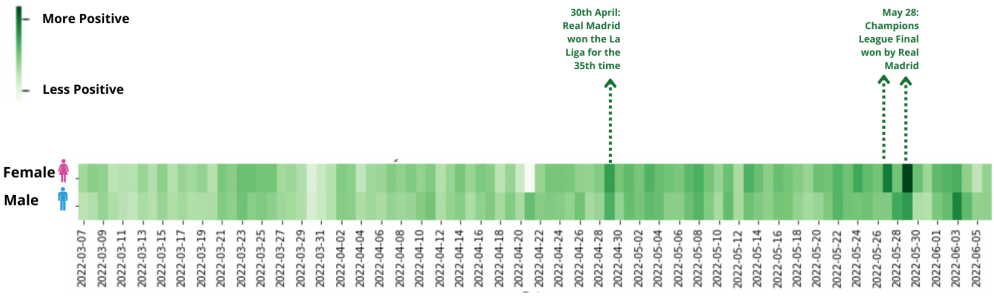
males. These choices of words by females indicate a proclivity towards conveying emotions that are more intense and perhaps more casual in nature.

Overall, the most prevalent emotions expressed by both males and females in soccer-related tweets were Joy (≈ 32% tweets), Anticipation (≈ 23% tweets), and Trust (≈ 22% tweets). To estimate gender differences, we analyzed the weekly and hourly distribution of emotions using bootstrapping (Figure 8 and 9). For each gender, we randomly select 80% of the tweets and compute the average value of a specific emotion. This process is repeated 1,000 times to ensure stable estimates. We then calculate the differences between the average of emotion values for men and women and compare their distributions using a t-test.The results indicate that females express higher levels of joy and anticipation compared to males, whereas males exhibit greater intensity in emotions such as fear, anger, and disgust. For surprise and sadness, no clear predominance is observed.

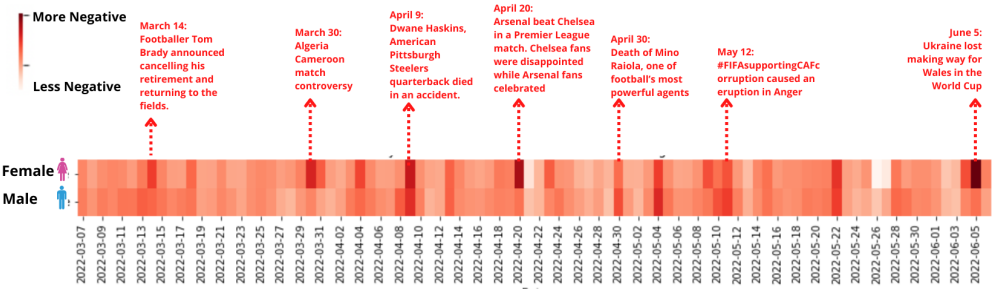
The analyses of the word clouds and emotions suggest that women and men tend to express themselves differently. An interesting nuance in this distinction is the intensity of the language chosen. Females not only use words associated with stronger emotions but also opt for more intense forms of these words compared to males. For instance, while both genders may use the word ‘Love,’ females are more prone to using its more intense variations. This nuanced difference underscores that females communicate emotions in a more vivid and intense manner compared to their male counterparts. Overall, different genders tend to use different words and expressions, and the intensity with which these words are expressed tends to be different, ultimately shaping how emotions are conveyed in communication.



(a) sentiment compound score



(b) positive sentiment



(c) negative sentiment

Fig. 6: Day wise distribution of sentiment scores across male's and female's tweets, and the events corresponding to these intense sentiments are highlighted (for English tweets).

4.1.5 Content Analysis

Our sentiment analysis reveals that irony and offensive language are not always accurately detected, highlighting the complexity of text interpretation [79]. Indeed, we observe a misalignment between the emotion of joy and the sentiment of positivity. Given these discrepancies, we extend our analysis to examine potential toxicity

Fig. 7: Word cloud of words used on the tweets for each language and gender. Words that are expected for the topic such as "football", " game" and "player" were removed to improve the context of the word cloud. Emotions are highly expressed when we remove topical words.

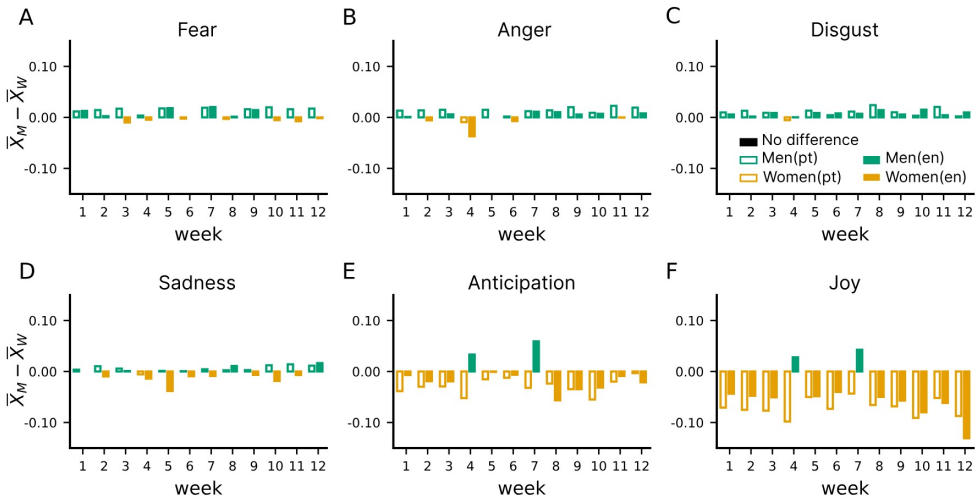


Fig. 8: Gender-based difference of emotion usage from tweets over time.

and abusive behavior in online interactions. We emphasize that our goal is not to assign blame to any gender or individuals but rather to present objective, data-driven insights. Ensuring an inclusive and healthy digital environment is essential to preventing mental health issues [80].

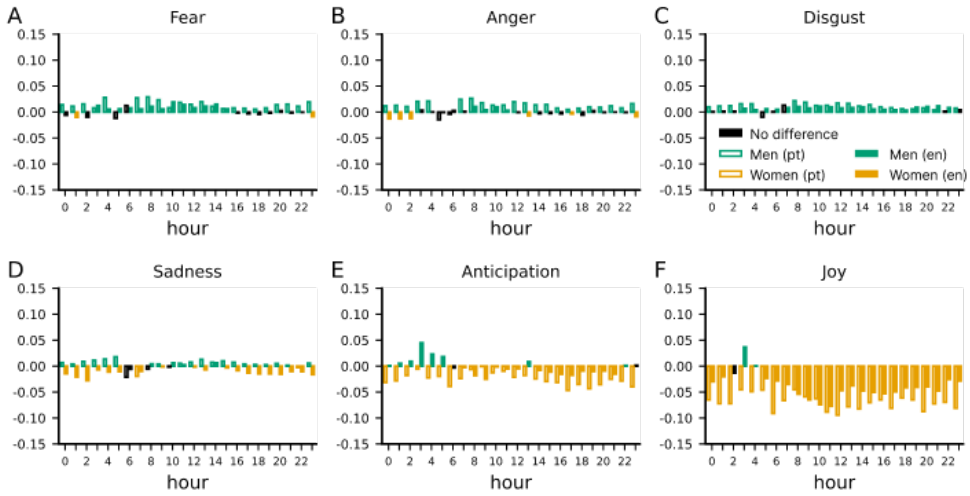


Fig. 9: Gender-based difference of emotion usage from tweets per hour over 12 weeks.

To assess different aspects of content toxicity, we use Google's Perspective API [73], which measures levels of severe toxicity, insult, attack on the author, threat, identity attack, and sexually explicit behavior. We apply an 80% bootstrapping subsampling method with 1,000 iterations and plot the differences in the average value of the distribution of toxicity metrics between genders (Figure 10). While the differences are generally small, they are statistically significant using the t-test. Our findings suggest that men are more likely to exhibit severe toxicity, insults, and aggressive behavior, whereas women are more prone to expressing negativity through threats, identity attacks, and sexually explicit language.

Online platforms, such as Twitter, offer more anonymity, allowing users to express themselves more openly. However, this anonymity has also led to the proliferation of profanity, creating a toxic environment at times. Toxicity is manifested in tweets that are impolite, disrespectful, or unreasonable, with the potential to drive individuals away from a discussion. Severe toxicity means a highly hateful, aggressive, or disrespectful comment, designed to strongly discourage a user from participating in a discussion or expressing their perspective. The high levels of toxicity displayed by some men in soccer communication on Twitter can stem from various factors. For instance, soccer is often deeply intertwined with masculinity in many cultures [81], leading to a heightened sense of competitiveness and aggression among male fans. This can manifest in toxic behavior such as insults, derogatory language, and aggression towards opposing fans or players.

In week 10, in Portuguese, women's communication shows higher toxicity and profanity (and sadness as shown in Figure 8) than men. In this period, the discussion from women was mainly about the game between Real Madrid 3-1 PSG, and close to this date, there were multiple complaints of racism towards a player of Real Madrid. For

the male tweets, we also observe the discussion of national teams and players from Portugal and Brazil, which were more intense than the discussions of the Champions League. This aligns with the results coming from the sentiment analysis, where women show more intense values in their sentiment, and men are more consistently participating. For women, we capture higher sentiments and emotions for particular events, and for men, we observe qualitatively a more diverse set of subtopics that are not always as unexpected or absurd.

Beyond toxicity, men tend to write in a manner that appears more aggressive toward both the author and commenters of a tweet. Men also post more insulting content; an insulting, inflammatory, or negative comment directed towards an individual or a group of people is considered for computing the insult. Men use more abusive and swear words in both languages, displayed in profanity. While not all profanity is intended to harm, words can impact the well-being of others. Some previous works have studied the correlation of gender with profanity. Previous works [82, 83] observed no relationship between gender and profanity, including swearing fluency and frequency. In contrast to this, Bamman et al. [84] have shown that males tend to use swear words more frequently than females and profanity usage is correlated with gender, though the correlation coefficient is not that high [85]. In our dataset, we observe similar results where males tend to have higher profanity than females, and this might be dependent on the topic. Women may be less comfortable expressing negative thoughts due to cultural norms that encourage caution in their communication.

Here, we show some of the tweets having high toxicity and insulting language (posted by male users) to give an idea of when a tweet is classified in these categories.

- *'Fuckin A. There is NO ONE worth $250 MILLION to play football or ANY sport. Obscene and fucking idiotic. #FuckClevelandBrowns'*
- *'Fuck football and especially fuck UEFA the dirty corrupt cunts.'*
- *'Fuck me, football fans are the worst aren't they? Only one openly gay player and the amount of disgusting shit that's being said about him and the wider community is infuriating. It's saddening that players feel the need to hide it but fuck me I can see why they do'*
- *'Fuck off and die, you fucking scum. Agents are the cancer of football. https://t.co/nBFmAOQFUr'*

These examples indicate the need to encourage a more healthier communication on social media.

Identity attack refers to making negative or hateful comments that specifically target someone based on their identity. In both languages, women show higher levels of identity attacks that might be due to the defensive stance when women feel their identity or belonging within the community is threatened. Additionally, women's experiences of discrimination, harassment, and marginalization within the sports community, both online and offline, can contribute to a heightened sensitivity to identity-related issues. This sensitivity may lead women to be more reactive to perceived attacks on their identity or credibility as soccer fans. However, an analysis of the attacking writing style

(Figure 10 G and H), reveals that men exhibit significantly higher levels compared to women. We observed that high values of attack and inflammation from women's tweets were written in a hypothetical and indirect way, such as "So would I be cheating on football if I become a fan of hockey?", and men tend to express directly their opinion such as "Since everyone is becoming a sports journalist nowadays, and since most of them are either misinformed or uninformed, I just decided to allocate some of my daily schedule to write articles and analyses about football like back" and "This guy only says the stupidest of things especially involving football. Unfortunately for us and him he's a football writer who loves utd.".

We further study the understanding of posted tweets to the audience and their relevance to soccer as shown in Figure 10 I, J, K, and L. We observe that women are very incoherent and share more spam in the posted content than males, and this might be seen due to the tendency of males to be more direct than women. Qualitatively, we can observe that women tend to be more ironic, indirect and hypothetical. Therefore, the incoherent text means that it is difficult to understand and nonsensical, and it can be related to them being considered spam.

On average, women's tweets were detected as more sexually explicit. We argue that, qualitatively, women more frequently address and challenge issues related to race, sex, and gender injustices, which may explain this observed dominance. For men, we qualitatively observe the use of sexuality to insult and swear at other people, which we can observe in the high values of insult, profanity, and attack on author/commenter.

Men have high toxicity and profanity, still the difference for inflammatory and obscene text is not very significant across gender. However, we observe that in Portuguese, men use more inflammatory language and in English, women use more. Further analysis is required to gain a comprehensive understanding of the underlying reasons. In general, soccer is a highly emotional and passionately followed sport, getting strong reactions from fans of all genders. The competitive nature of the sport, and strong support to favorite teams, can lead to heightened emotions and a propensity for inflammatory language.

From text analysis, one important conclusion is that the way men and women tweet and present their opinion and emotions tend to be consistent in both the languages, also over the time (weeks). The observation that men use more toxic, insulting, and profane language highlights the ongoing need to reduce abusive communication. This discrepancy suggests that efforts to promote respectful and constructive dialogue, particularly among male users, are essential for fostering a healthier and more inclusive online environment for everyone.

4.2 Network-Analysis

We build weighted weekly cumulative networks from the communication between the users (nodes), where the link weight is the number of times two users interacted with each other by replying or retweeting information. There are three types of networks for each language based on (i) retweets, (ii) replies, and (iii) both retweets and replies (mentioned as the combined network based on all communications). The size of the

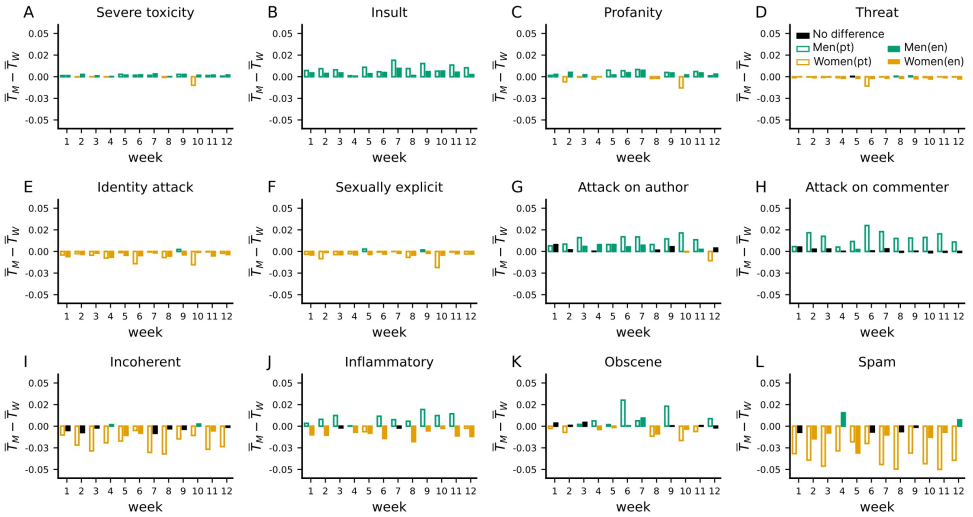


Fig. 10: Gender-based differences for different linguistic styles computed using tweet text. The differences are computed from the average value of a characteristic from men minus the one from women.

networks are - in English, there are 148,184 females (18.94%) and 634,366 males, and in Portuguese, there are 50,354 females (25.85%) and 144,441 males. In total, we have 4,308,009 interactions in English, where 1,348,557 are replies, and 2,959,452 are retweets, and in Portuguese, we have 2,117,105 interactions, where 986,288 are replies, and 1,130,817 are retweets. We analyze these networks to understand how information is shared for each language based on their structure and flow.

4.2.1 Communication Patterns

We first observed that the network from the communication across languages tends to be different ($NPD > 0.90$, results in Table 2) by computing the Network Portrait Divergence (NPD) between networks that estimate how different two networks are in relation to their structure (0 to 1 corresponds to being similar to different) [86]. The NPD method compares networks using the network portrait, which is a (l, k) size array containing the number of nodes that have k nodes at a distance l. Therefore, it provides an information-theoretic interpretation for comparing networks based on the structures of all scales and is well-generalized for weighted networks.

We observe that the networks tend to be different in structure, but when we compare the two networks built solely from the retweets, the NPD tends to be close to 0.4, indicating some degree of similarity. This means that the replies play a major role in making the interactions different. Next, we use NPD to compare the structures from the networks for women and men separately. Overall, women's and men's communications tend to emerge in different networks (values of NPD closer to 1). However, we see that for replies in English, the network structure tends to be more similar across

genders, indicating that the networks built from Portuguese tweets are more different in structure than the ones from English. Thus, the results of NPD already unveil heterogeneities coming from the communication of women and men.

Table 2: Network Portrait divergence (NPD) average (standard variation) values using a bootstrap technique. We compare the networks from 80% of the data built 1000 times, and compute the average and standard deviation values of NPD.

Tweets	Comparison	NPD
all	across languages	0.90 (0.007)
	English, across gender	0.97 (0.01)
	Portuguese, across gender	0.99 (0.001)
retweets	across languages	0.46 (0.04)
	English, across gender	0.39 (0.01)
	Portuguese, across gender	0.41 (0.02)
replies	across languages	0.91 (0.007)
	English, across gender	0.28 (0.02)
	Portuguese, across gender	0.94 (0.001)

4.2.2 Network Characteristics

Let us investigate further the characteristics of these networks to understand the patterns of women's and men's communication. We compare the average values of multiple metrics computed from the networks built from all the tweets in English and Portuguese (Figure 11). From the cumulative evolution over 12 weeks, we observe that the number of nodes and edges is higher for English, and the density and average clustering tend to be smaller. We argue that people tweeting in Portuguese might belong to more concentrated regions than in English (worldwide spoken language). Therefore, soccer might come as a more popular and well-connected local phenomenon in Brazil and Portugal, for instance, but for English tweets, we look at several communities together (worldwide spoken language). This distinction can be indicated by the higher connection between small groups (closer average clustering values) but not to the overall network (much smaller density) in Figure 11 C-D.

Even though English tweets tend to come from people from different regions, the fraction of women does not scale at the same rate as Portuguese. A much smaller community formed through Portuguese tweets contains a much larger fraction of women. In general, women tend to be a minority in sports-related topics [87, 88], but the popularity of soccer in smaller communities (Portuguese ones) might encourage women to participate more on the topic. This higher participation can also translate into a more active participation across genders. Women from Portuguese tweets tend to interact more with men than from English (lower values of gender homophily for Portuguese).

To make sure that these differences are not a group size effect, we compute the assortativity of both networks, and we still observe the same patterns. The assortativity is computed using the adjusted nominal assortativity, i.e., defined explicitly for networks having asymmetric mixing of nodes from different groups (genders in our case) [89]. Thus, the increase in the fraction of women in the Portuguese network over weeks impacts the homophily values, and with more women, the homophily and assortativity are reduced. We highlight that this finding is not necessarily expected as quantity does not always translate into more communication across majority and minority. In our case, this finding suggests that encouraging women to participate more – regardless of being a male dominant sport such as football – might be one way to improve inclusivity.

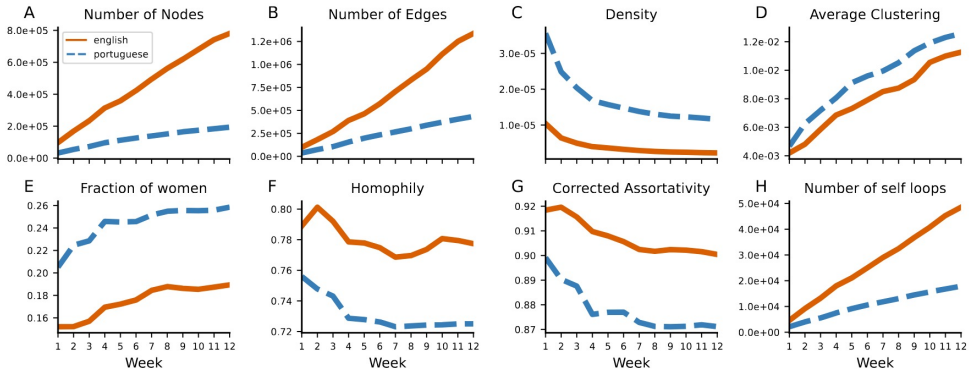


Fig. 11: Network metrics computed from the combined network of retweets and replies in Portuguese and English.

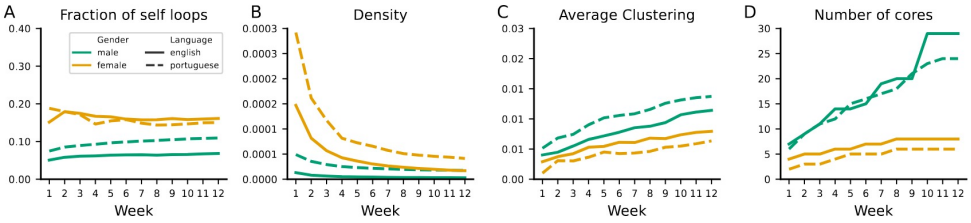


Fig. 12: Network metrics for the combined network (retweets + replies) disaggregated by gender for Portuguese and English.

Next, we look at the self-loops in the networks (Figure 11 H), and we see that English tend to have more self-loops than Portuguese, as it is expected by their group size. However, when we disaggregate these self-loops by gender (Figure 12 A), we see that around 10% of the edges coming from men's interactions are self-loops, which is half of

what we see from women's interactions. This indicates that women in our sample tend to retweet and reply more to their own content than men. By disaggregating other network metrics based on gender, we also see that women have denser networks, but men tend to have higher average clustering and higher core numbers. This indicates that women tend to be more connected in general, but men tend to act more as hubs and acquire more central positions (information spreaders).

The core number of a node shows its influential power, and the nodes at the very internal core (corresponding to the highest core number) are the top influential nodes [90]. The highest number of cores for English is 33, and for Portuguese is 28. At these cores, we analyze the networks in Portuguese (169 people) and English (214 people), and we found the women ratio of 14.79% and 12.61%, respectively. Even though women have a higher ratio of most active users in Portuguese than in English, the difference between the women ratio of the overall network ($f_{w,pt} = 25.85\%$, $f_{w,en} = 18.94\%$) and the core is decreased by 11.06% and 6.33%, respectively, indicating that women are not as influential as men in both languages.

4.2.3 Meso-scale Structures

We further study the organization of nodes in the network based on meso-scale structures, including core-periphery and communities. In Figure 13, we plot the evolution of network parameters over cores to analyze hierarchical structure in both networks based on all kinds of tweets; English has a higher core number than Portuguese. We identify the core-periphery using the k-shell decomposition method [90]. We first study how homophily varies as we move from periphery to core to understand the gender roles at different influential levels. Many previous works have shown that women do not acquire top positions in the network due to the glass-ceiling effect [91, 92]. Figure 13 shows the fraction of females in different layers as we move from periphery to core in both the networks, and we observe that when a core of Portuguese reaches a similar fraction of women inside it compared to English, they have a much lower homophily than in English. We also observe that homophily increases as we move from periphery to core, equal to 15 for both languages, a fraction of women decreases, and density continues to be similar across languages. However, the average clustering for Portuguese increases much more than for English, probably due to the fact that the people in the network belong to a specific region. Here, we then can observe that a higher women fraction in Portuguese is also associated with a higher women fraction at influential cores, and consequently, women and men are more likely to interact with each other.

Next, we study the embedding of females and males in different communities; the results are shown in Figure 14. Communities are extracted using the Leiden Community detection method due to its ability to find meaningful groups [93]. Analyzing the network of English tweets reveals that homophily increases as the group size decreases, and the biggest community reflects the smallest homophily value. This also happens in the Portuguese network, but the variance is much higher. Communities from English tweets are more homophilic (as shown in Figure 11). The increase of homophily does

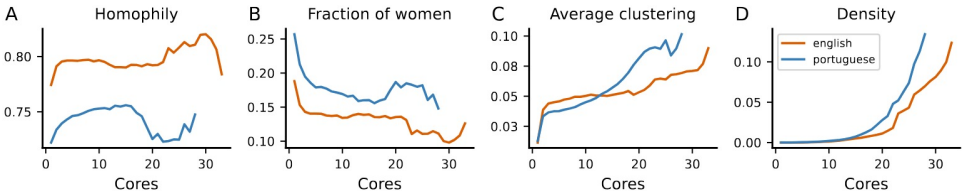


Fig. 13: Evolution of the metrics as we go from the core 1 to the highest numbered core.

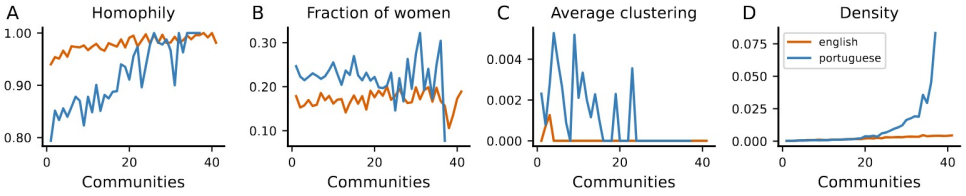


Fig. 14: Evolution of the metrics as we move from the largest (labeled as 1) to the smallest size community.

not correlate with the values of the fraction of women and average clustering, but for Portuguese, the density and homophily increase as the group size decreases.

5 Discussion

In this study, we explore gender dynamics in soccer-related Twitter discussions in English and Portuguese, focusing on a globally popular sport with immense cultural and economic significance. While prior research [10, 11, 94] has examined gender disparities in specialized fields like STEM and politics—where participation often requires domain expertise—soccer offers a more accessible space for engagement, fostering interaction across diverse audiences. Despite its widespread appeal, little is known about how gender shapes discourse in this space beyond isolated events, championships or specific figures [57, 95, 96]. Here, we take a broader approach, integrating text, network, emotion, and sentiment analysis to uncover deeper patterns in online soccer discussions, offering new insights into gendered communication on social media. Our findings also raise questions about how the popularity and familiarity of a topic may shape disparities in online communication. This paper does not aim to stereotype women and men, but to identify possible evidence when women or men do not feel welcome to interact on online platforms.

We collected soccer-related Twitter data for three months (March 7 to June 6, 2022) and identified the genders of users using Genderize [58] and Namepedia [59], similar to other papers [60, 97]. The scope of our study is confined to binary gender (women or men) and tweets from authors whose gender is explicitly identified. The final dataset

contains 7 million tweets in English (from 2 million users) and 2.5 million tweets in Portuguese (from 0.5 million users). For both languages, we looked at the content of tweets and how people interact to understand how a male-dominated environment affects communication for both women and men. We found that women are often a minority in terms of participation and representation for the topic of soccer, but when they do engage, they tend to express more positive emotions and stronger feelings compared to men.

There are some similarities in the online communication of women and men, such as the top 10 most frequently used emojis and the usage of hashtags in the tweets. Yet, women use more emojis than men, though this does not translate to the usage of hashtags. Men post more negative tweets, and women post more positive tweets. The average sentiment of men's tweets is generally higher than that of women, suggesting that, overall, their tweets tend to be more positive than negative in relation to the volume of posts. We can then further explore these positive and negative tweets, breaking them down into a more detailed and nuanced spectrum of emotions. Women tend to express higher levels of joy and anticipation than men in both languages, and disgust, anger, and fear tend to be more gender-neutral, with slightly higher levels for males. We found that differences are consistent weekly and hourly. Women display intense sentiments and emotions for any event as compared to men. Besides these, tweets posted by men have higher toxicity, profanity, insulting text, abusive language, and attacking writing style. However, women's tweets are more incoherent, sexually explicit, and have higher identity attacks. Interestingly, we did not find any notable emotional difference based on gender between English and Portuguese, and emotional responses appear to be unaffected by the overall network structure.

Using network analysis, we study how women and men form connections on Twitter and how these connections are maintained at a macro scale. We constructed multiple networks to study the role of language (i.e. English and Portuguese), gender (i.e. Women and Men) and communication type (i.e. retweet, reply and combined/overall) separately and together. The reply and combined networks are statistically different (computed using NPD) across genders, highlighting a significant difference in communication across genders based on replies. However, the retweet networks across languages and genders are not as different (NPD score is around 0.4), indicating that differences come more from active communication than passive, in line with the hypothesis that some people might not feel comfortable sharing directly their thoughts. The Portuguese network exhibits a higher proportion of women in interactions and lower homophily compared to the English network. This difference could be attributed to a regional focus in Portuguese tweets, mainly from Brazil and Portugal, where soccer is highly popular. Women's networks in Portuguese show denser structures, with higher average clustering and lower assortativity than those of men. Women appear less influential, and the ratio of women reduces from the periphery to the core of the network; the fraction is decreased by 11.06% and 6.33% in Portuguese and English, respectively. Small communities in both languages are more homophilic. In Portuguese, bigger communities are less homophilic, and homophily reduces with the community size, given that the fraction of women is maintained in most of the

communities. However, the difference in homophily across communities is not very significant in English.

Our observations indicate a correlation between a topic's popularity and lower homophily in the network, suggesting that greater popularity creates a safer space for communication between different genders. This, however, does not appear to influence how individuals express their sentiments and emotions. The significant communication gap highlighted in this study is eye-opening, emphasizing the need to focus on bridging it and creating online safe spaces for discussions on such topics. This paper does not propose intervention methods, nevertheless, we believe the observed outcomes can serve as a foundation. Our findings reveal persistent gender gaps in participation and communication styles within soccer-related discussions. Therefore, online platforms have an important role to play in fostering equitable engagement. Several design and policy interventions could help reduce these disparities. At the platform level, recommendation algorithms could be adjusted to promote content from underrepresented voices, mitigating visibility biases that reinforce existing inequalities. Platforms could also deploy moderation tools: particularly those attuned to gendered harassment and implicit bias—to reduce toxicity and make participation safer for marginalized groups. In addition, platforms might introduce features that encourage cross-group interaction (e.g., promoting diverse viewpoints in conversation threads, highlighting respectful exchanges) to help reduce homophily and broaden participation. These interventions should be developed with user input and regularly evaluated for their impact, ensuring that efforts to promote equity align with community needs and cultural contexts.

We acknowledge several limitations in this study. First, Twitter's user database may not fully represent the global population, as it tends to be predominantly composed of specific groups, such as white, male, and middle-upper-class individuals, which could impact the generalizability of our findings. We also focused only on two languages, Portuguese and English, which constrains the scope of the communities analyzed, and the results may not directly translate to other platforms (e.g., Instagram, Reddit, TikTok), where participation patterns, demographics, and community norms differ. Nevertheless, examining online social media platforms like Twitter provides valuable insights that can help make these spaces more inclusive via content moderation, reporting tools, inclusive design features, and support for victims of harassment.

Second, some methodologies, particularly the APIs used for gender inference and sentiment analysis, still require refinement. In this study, we employ the Genderize [58] and Namepedia [59] APIs to infer users' gender, retaining only those cases where gender is detected with high probability. Instances with uncertain classification are excluded, thereby restricting the analysis to a binary gender framework. This introduces potential biases, such as the misclassification of users with unisex names, the exclusion of non-binary individuals, and disproportionate filtering of underrepresented cultural or linguistic groups, which may limit a complete understanding of communication styles. In terms of sentiment, sarcasm and irony were occasionally misclassified as positive, a known limitation of NLP techniques [98, 99]. To mitigate this, we adopted a multi-faceted approach, cross-validating results with manual checks, which improved accuracy.

Third, our analysis is static and therefore cannot capture the dynamic, rapidly evolving nature of online data. Changes in platform policies, user behavior, and trending topics mean that our findings reflect only a specific snapshot in time. A longitudinal study would provide a more comprehensive understanding of temporal variation.

redFinally, a more comprehensive network analysis would have been possible with access to full follower–followee relationships over the same time period. Our dataset, however, was constrained to user activity and tweets extracted through the default API. Additionally, the presence of bots or automated accounts may still have influenced observed patterns, although filtered as much as possible.

Our study goes beyond the exploration of gender variations in communication patterns, highlighting potential misperceptions of free speech on social media. Despite soccer being more popular among Portuguese speakers, we observe similarities to English speakers (such as influential positions for women) as well as variations (such as differences in network structures) in communication patterns across languages. In the future, we would like to study further the correlation of popularity and ease of topics with the homophily of their network and writing style. If similar results are observed, then one can use awareness and education programs to bridge these gaps. Similar techniques could be used for topics that are very domain-specific, such as STEM. We also highlight that women and men may use the Internet and social media for different purposes and in different ways [100], meaning that variations in their communication may also stem from their underlying intentions. We will also investigate the explainability of patterns extracted from the networks and their communities further. Additionally, we plan to compare these patterns with a sport or activity predominantly followed by women, such as ballet. We posit that different sports may influence people to communicate in distinct manners.

6 Conclusion and Future Directions

This work studies gender-based communication patterns on Twitter within the context of soccer in English and Portuguese. Our findings reveal notable gender differences in sentiment, emotional expression, language use, and network structures. Women exhibit more intense sentiments, with greater levels of joy and anticipation, while men's tweets are more likely to contain toxic and offensive language. Women's tweets also show more incoherence and a higher incidence of identity attacks. Network analysis indicates that the Portuguese network has higher female participation and lower gender homophily compared to the English network, likely influenced by the strong regional presence of soccer in Portuguese-speaking communities. Women's influence also diminishes toward the core of the network, in contrast to men's, highlighting structural challenges in visibility and engagement.

These results suggest that the popularity and familiarity of a topic can reduce homophily and encourage safer cross-gender communication, though emotional expression patterns remain unaffected. Our findings lay the groundwork for future research aimed at bridging communication gaps and promoting equitable participation in online discussions. They underscore the importance of fostering inclusive online spaces

where all genders feel equally welcome. Future research should examine how topic familiarity and accessibility influence homophily and communication styles in other domains, such as STEM. Additionally, exploring the reasons behind network evolution and the socio-cultural factors shaping engagement could yield valuable insights. Comparing gendered communication patterns in other sports and events with higher female engagement, such as ballet, could further enrich our understanding.

Declarations

Competing Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. interest.

Authors Contribution Statement

Mariana Macedo: Conceptualization, Methodology, Analysis, Writing. **Akrati Saxena:** Conceptualization, Methodology, Analysis, Writing.

Ethical and informed consent for data used

This study does not contain any studies with human or animal subjects performed by any of the authors.

Data Availability and Access

Our dataset is available at https://github.com/akratiiet/Soccer-Twitter-Dataset.

References

[1] Williams, J., Chinn, S.J., Suleiman, J.: The value of twitter for sports fans. Journal of Direct, Data and Digital Marketing Practice **16**(1), 36–50 (2014) https://doi.org/10.1057/dddmp.2014.36

[2] Wold., H.M., Vikre., L., Gulla., J.A., Özgöbek., Su., X.: Twitter Topic Modeling for Breaking News Detection. SciTePress (2016). https://doi.org/10.5220/0005801902110218

[3] Lingiardi, V., Carone, N., Semeraro, G., Musto, C., D'Amico, M., Brena, S.: Mapping twitter hate speech towards social and sexual minorities: a lexicon-based approach to semantic content analysis. Behaviour & Information Technology **39**(7), 711–721 (2020) https://doi.org/10.1080/0144929X.2019.1607903

[4] Holmberg, K., Hellsten, I.: Gender differences in the climate change communication on twitter. Internet research (2015)

[5] Garcia, D., Weber, I., Garimella, V.: Gender asymmetries in reality and fiction: The bechdel test of social media. Proceedings of the International AAAI Conference on Web and Social Media **8**(1), 131–140 (2014)

[6] May, A., Wachs, J., Hannák, A.: Gender differences in participation and reward on stack overflow. Empirical Software Engineering **24**(4), 1997–2019 (2019)

[7] Wagner, C., Garcia, D., Jadidi, M., Strohmaier, M.: It's a Man's Wikipedia? Assessing Gender Inequality in an Online Encyclopedia. arXiv (2015). https://doi.org/10.48550/ARXIV.1501.06307 . https://arxiv.org/abs/1501.06307

[8] Thelwall, M., Foster, D.: Male or female gender-polarized youtube videos are less viewed. Journal of the Association for Information Science and Technology **72** (2021) https://doi.org/10.1002/asi.24529

[9] Messias, J., Vikatos, P., Benevenuto, F.: White, man, and highly followed: Gender and race inequalities in twitter. In: Proceedings of the International Conference on Web Intelligence, pp. 266–274 (2017)

[10] Nilizadeh, S., Groggel, A., Lista, P., Das, S., Ahn, Y.-Y., Kapadia, A., Rojas, F.: Twitter's glass ceiling: The effect of perceived gender on online visibility. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 10, pp. 289–298 (2016)

[11] Manzano, C., Sánchez-Giménez, J.A.: Women, gender and think tanks: political influence network in twitter 2018 (2019)

[12] Chaplin, T.: Gender and emotion expression: A developmental contextual perspective. Emotion review : journal of the International Society for Research on Emotion **7**, 14–21 (2015) https://doi.org/10.1177/1754073914544408

[13] Sainz-de-Baranda, C., Adá-Lameiras, A., Blanco-Ruiz, M.: Gender differences in sports news coverage on twitter. International Journal of Environmental Research and Public Health **17**(14), 5199 (2020)

[14] Munoz, A.M.M., Salido-Fernandez, J.: Gender biases in the social networks of regional public media: The case of@ csurnoticias' twitter. Revista Latina de Comunicacion Social **82**, 46–61 (2024)

[15] De Choudhury, M., Sundaram, H., John, A., Seligmann, D.D., Kelliher, A.: "birds of a feather": Does user homophily impact information diffusion in social media? arXiv preprint arXiv:1006.1702 (2010)

[16] De Choudhury, M.: Tie formation on twitter: Homophily and structure of egocentric networks. In: 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing, pp. 465–470 (2011). IEEE

[17] Kang, J.-H., Lerman, K.: Using lists to measure homophily on twitter. In: Workshops at the Twenty-sixth AAAI Conference on Artificial Intelligence (2012)

[18] Macskassy, S., Michelson, M.: Why do people retweet? anti-homophily wins the day! In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 5, pp. 209–216 (2011)

[19] Lin, H., Wang, Y., Kim, Y.: The rich get richer and the poor get poorer? the effect of news recommendation algorithms in exacerbating inequalities in news engagement and social capital. New Media & Society, 14614448231168572 (2023)

[20] Article, W.: FIFA World Cup. https://en.wikipedia.org/wiki/FIFA_World_Cup (Online accessed on 30-Apr-2024)

[21] Lange, D.: Soccer fans by gender 2019 (2020). https://www.statista.com/statistics/1134065/share-soccer-fans-gender/

[22] Sanderson, J., Gramlich, K.: "you go girl!": Twitter and conversations about sport culture and gender. Sociology of Sport Journal **33**(2), 113–123 (2016)

[23] Antunovic, D., Hardin, M.: Activism in women's sports blogs: Fandom and feminist potential. International Journal of Sport Communication **5**(3), 305–322 (2012)

[24] Hu, L., Kearney, M.W.: Gendered tweets: Computational text analysis of gender differences in political discussion on twitter. Journal of Language and Social Psychology **40**(4), 482–503 (2021)

[25] Schischlik, A.: Women and football: #ChangeTheGame - towards gender equality in sports. Online accessed on 14-01-2023 (2023). https://www.unesco.org/en/articles/women-and-football-changethegame-towards-gender-equality-sports

[26] Pegoraro, A.: Twitter as disruptive innovation in sport communication. Communication & Sport **2**(2), 132–137 (2014)

[27] Gibbs, C., Haynes, R.: A phenomenological investigation into how twitter has changed the nature of sport media relations. International Journal of Sport Communication **6**(4), 394–408 (2013)

[28] Hull, K., Lewis, N.P.: Why twitter displaces broadcast sports media: A model. International Journal of Sport Communication **7**(1), 16–33 (2014)

[29] Matsa, K.E., Mitchell, A.: Key takeaways about social media and news. Pew Research Center. Journalism and media (2014)

[30] Gibbs, C., O'Reilly, N., Brunette, M.: Professional team sport and twitter: Gratifications sought and obtained by followers. International Journal of Sport Communication **7**(2), 188–213 (2014)

[31] Yoon, S., Petrick, J.F., Backman, S.J.: Twitter power and sport-fan loyalty: The moderating effects of twitter. International Journal of Sport Communication **10**(2), 153–177 (2017)

[32] Boehmer, J., Tandoc, E.C.: Why we retweet: Factors influencing intentions to share sport news on twitter. International Journal of Sport Communication **8**(2), 212–232 (2015)

[33] Witkemper, C., Lim, C.H., Waldburger, A.: Social media and sports marketing: Examining the motivations and constraints of twitter users. Sport Marketing Quarterly **21**(3) (2012)

[34] Lanagan, J., Smeaton, A.: Using twitter to detect and tag important events in sports media. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 5, pp. 542–545 (2011)

[35] Hambrick, M.E., Simmons, J.M., Greenhalgh, G.P., Greenwell, T.C.: Understanding professional athletes' use of twitter: A content analysis of athlete tweets. International Journal of Sport Communication **3**(4), 454–471 (2010)

[36] Svensson, P.G., Mahoney, T.Q., Hambrick, M.E.: Twitter as a communication tool for nonprofits: A study of sport-for-development organizations. Nonprofit and voluntary sector quarterly **44**(6), 1086–1106 (2015)

[37] Thorpe, H., Rinehart, R.: Action sport ngos in a neo-liberal context: The cases of skateistan and surf aid international. Journal of sport and social issues **37**(2), 115–141 (2013)

[38] Krasnova, H., Veltri, N.F., Eling, N., Buxmann, P.: Why men and women continue to use social networking sites: The role of gender differences. The Journal of Strategic Information Systems **26**(4), 261–284 (2017)

[39] Graells-Garrido, E., Baeza-Yates, R., Lalmas, M.: How representative is an abortion debate on twitter? In: Proceedings of the 10th ACM Conference on Web Science, pp. 133–134 (2019)

[40] Wu, H.H., Gallagher, R.J., Alshaabi, T., Adams, J.L., Minot, J.R., Arnold, M.V., Welles, B.F., Harp, R., Dodds, P.S., Danforth, C.M.: Say their names: Resurgence in the collective attention toward black victims of fatal police violence following the death of george floyd. Plos one **18**(1), 0279225 (2023)

[41] Peña-Fernández, S., Larrondo-Ureta, A., Morales-i-Gras, J.: Feminism, gender identity and polarization in tiktok and twitter. Comunicar: Media Education

Research Journal **31**(75), 47–58 (2023)

[42] Yoon, J., Smith, C., Hyung Kim, A.C., Clavio, G., Witkemper, C., Pedersen, P.M.: Gender effects on sport twitter consumption: Differences in motivations and constraints. Journal of Multidisciplinary Research (1947-2900) **6**(3) (2014)

[43] Clavio, G., Walsh, P., Coyle, P.: The effects of gender on perceptions of team twitter feeds. Global Sport Business Journal **1**(1) (2013)

[44] Grace, A.N., Mueller, T.S.: Gender bias in sport media: a critical analysis of twitter content and the national football league's carolina panthers. Journal of gender studies **28**(3), 363–370 (2019)

[45] Saxena, A., Reddy, H., Saxena, P.: Introduction to sentiment analysis covering basics, tools, evaluation metrics, challenges, and applications. Principles of social networking: the new horizon and emerging challenges, 249–277 (2022)

[46] Hakak, N.M., Mohd, M., Kirmani, M., Mohd, M.: Emotion analysis: A survey. In: 2017 International Conference on Computer, Communications and Electronics (COMPTELIX), pp. 397–402 (2017). IEEE

[47] Hong, T.-P., Serrano-Estrada, L., Saxena, A., Biswas, A.: Deep Learning for Social Media Data Analytics. Springer, Cham (2022)

[48] Thelwall, M., Wilkinson, D., Uppal, S.: Data mining emotion in social network communication: Gender differences in myspace. Journal of the American Society for Information Science and Technology **61**(1), 190–199 (2010)

[49] Kivran-Swaine, F., Brody, S., Diakopoulos, N., Naaman, M.: Of joy and gender: emotional expression in online social networks. In: Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work Companion, pp. 139–142 (2012)

[50] Zhang, Y., Dang, Y., Chen, H.: Research note: Examining gender emotional differences in web forum communication. Decision Support Systems **55**(3), 851–860 (2013)

[51] Fouad, S., Alkooheji, E.: Sentiment analysis for women in stem using twitter and transfer learning models. In: 2023 IEEE 17th International Conference on Semantic Computing (ICSC), pp. 227–234 (2023). IEEE

[52] Panchendrarajan, R., Saxena, H., Saxena, A.: Social media and academia: How gender influences online scholarly discourse. arXiv preprint arXiv:2505.03773 (2025)

[53] Parkins, R.: Gender and Emotional Expressiveness: An Analysis of Prosodic Features in Emotional Expression. Griffith University Nathan, QLD, ??? (2012)

[54] Rangel, F., Hernández, I., Rosso, P., Reyes, A.: Emotions and irony per gender in facebook. In: Proceedings of Workshop ES3LOD, LREC, pp. 1–6 (2014)

[55] Reychav, I., Inbar, O., Simon, T., McHaney, R., Zhu, L.: Emotion in enterprise social media systems. Information Technology & People **32**(1), 18–46 (2019)

[56] Garcia-Rudolph, A., Laxe, S., Saurí, J., Bernabeu Guitart, M.: Stroke survivors on twitter: sentiment and topic analysis from a gender perspective. Journal of medical Internet research **21**(8), 14077 (2019)

[57] Bagić Babac, M., Podobnik, V.: A sentiment analysis of who participates, how and why, at social media sport websites: How differently men and women write about football. Online Information Review **40**(6), 814–833 (2016)

[58] Genderize.io: https://genderize.io/ ([Online accessed on 2-Aug-2022])

[59] namepedia.org: http://www.namepedia.org/ ([Online accessed on 2020-11-11])

[60] Macedo, M., Jaramillo, A.M., Menezes, R.: Academic mobility as a driver of productivity: A gender-centric approach. In: International Workshop on Complex Networks, pp. 120–131 (2023). Springer

[61] Burger, J.D., Henderson, J., Kim, G., Zarrella, G.: Discriminating gender on twitter. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pp. 1301–1309 (2011)

[62] Yang, Y.-C., Al-Garadi, M.A., Love, J.S., Perrone, J., Sarker, A.: Automatic gender detection in twitter profiles for health-related cohort studies. JAMIA open **4**(2), 042 (2021)

[63] Saeed, R.: Twitter and development of social capital: A gender perspective. Telematics and Informatics Reports **9**, 100037 (2023)

[64] Sarsam, S.M., Alzahrani, A.I., Al-Samarraie, H., Alblehai, F.: Gender preferences in cryptocurrency systems: Sentiment analysis and predictive modelling. Investment Analysts Journal, 1–15 (2024)

[65] Newman, M.L., Groom, C.J., Handelman, L.D., Pennebaker, J.W.: Gender differences in language use: An analysis of 14,000 text samples. Discourse processes **45**(3), 211–236 (2008)

[66] Park, G., Yaden, D.B., Schwartz, H.A., Kern, M.L., Eichstaedt, J.C., Kosinski, M., Stillwell, D., Ungar, L.H., Seligman, M.E.: Women are warmer but no less assertive than men: Gender and language on facebook. PloS one **11**(5), 0155885 (2016)

[67] Oxford University Press, A.: Word of the Year 2015. https://languages.oup.com/word-of-the-year/2015/. [Online accessed on 2-Aug-2022] (2015)

[68] Robertson, A., Magdy, W., Goldwater, S.: Black or white but never neutral: how readers perceive identity from yellow or skin-toned emoji. Proceedings of the ACM on Human-Computer Interaction **5**(CSCW2), 1–23 (2021)

[69] Robertson, A., Magdy, W., Goldwater, S.: Identity signals in emoji do not influence perception of factual truth on twitter. arXiv preprint arXiv:2105.03160 (2021)

[70] Robertson, A., Liza, F.F., Nguyen, D., McGillivray, B., Hale, S.A.: Semantic journeys: quantifying change in emoji meaning from 2012-2018. arXiv preprint arXiv:2105.00846 (2021)

[71] Chen, Z., Lu, X., Ai, W., Li, H., Mei, Q., Liu, X.: Through a gender lens: Learning usage patterns of emojis from large-scale android users. In: Proceedings of the 2018 World Wide Web Conference, pp. 763–772 (2018)

[72] Shah, N., Rohilla, S.: Emot API. https://pypi.org/project/emot/. [Online; accessed 2-Aug-2022] (2022)

[73] Hutto, C., Gilbert, E.: Vader: A parsimonious rule-based model for sentiment analysis of social media text. Proceedings of the International AAAI Conference on Web and Social Media **8**(1), 216–225 (2014)

[74] Mohammad, S.M.: Word affect intensities. In: Proceedings of the 11th Edition of the Language Resources and Evaluation Conference (LREC-2018), Miyazaki, Japan (2018)

[75] Twitter: How to use hashtags. https://help.twitter.com/en/using-twitter/how-to-use-hashtags ([Online accessed on 18-July-2022])

[76] Stracqualursi, L., Agati, P.: Tweet topics and sentiments relating to distance learning among italian twitter users. Scientific Reports **12** (2022) https://doi.org/10.1038/s41598-022-12915-w

[77] Cheng, I., Heyl, J., Lad, N., Facini, G., Grout, Z.: Evaluation of twitter data for an emerging crisis: an application to the first wave of covid-19 in the uk. Scientific Reports **11**(1), 1–13 (2021)

[78] Matalon, Y., Magdaci, O., Almozlino, A., Yamin, D.: Using sentiment analysis to predict opinion inversion in tweets of political communication. Scientific Reports **11** (2021) https://doi.org/10.1038/s41598-021-86510-w

[79] Farías, D.I.H., Patti, V., Rosso, P.: Irony detection in twitter: The role of affective content. ACM Transactions on Internet Technology (TOIT) **16**(3), 1–24 (2016)

[80] O'reilly, M., Dogra, N., Whiteman, N., Hughes, J., Eruyar, S., Reilly, P.: Is

social media bad for mental health and wellbeing? exploring the perspectives of adolescents. Clinical child psychology and psychiatry **23**(4), 601–613 (2018)

[81] Hughson, J.: The boys are back in town: Soccer support and the social reproduction of masculinity. Journal of Sport and Social Issues **24**(1), 8–23 (2000)

[82] Thelwall, M.: Fk yea i swear: cursing and gender in myspace. Corpora **3**(1), 83–107 (2008)

[83] Stephens, R., Zile, A.: Does emotional arousal influence swearing fluency? Journal of Psycholinguistic Research **46**, 983–995 (2017)

[84] Bamman, D., Eisenstein, J., Schnoebelen, T.: Gender identity and lexical variation in social media. Journal of Sociolinguistics **18**(2), 135–160 (2014)

[85] Wong, S.C., Teh, P.L., Cheng, C.-B.: How different genders use profanity on twitter? In: Proceedings of the 2020 4th International Conference on Compute and Data Analysis, pp. 1–9 (2020)

[86] Bagrow, J.P., Bollt, E.M.: An information-theoretic, all-scales approach to comparing networks. Applied Network Science **4**(1), 1–15 (2019)

[87] Farrell, T., Fernandez, M., Novotny, J., Alani, H.: Exploring misogyny across the manosphere in reddit. In: Proceedings of the 10th ACM Conference on Web Science, pp. 87–96 (2019)

[88] Thelwall, M., Stuart, E.: She's reddit: A source of statistically significant gendered interest information? Information processing & management **56**(4), 1543–1558 (2019)

[89] Karimi, F., Oliveira, M.: On the inadequacy of nominal assortativity for assessing homophily in networks. Scientific Reports **13**(1), 21053 (2023)

[90] Kitsak, M., Gallos, L.K., Havlin, S., Liljeros, F., Muchnik, L., Stanley, H.E., Makse, H.A.: Identification of influential spreaders in complex networks. Nature physics **6**(11), 888–893 (2010)

[91] Avin, C., Keller, B., Lotker, Z., Mathieu, C., Peleg, D., Pignolet, Y.-A.: Homophily and the glass ceiling effect in social networks. In: Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science, pp. 41–50 (2015)

[92] Stoica, A.-A., Riederer, C., Chaintreau, A.: Algorithmic glass ceiling in social networks: The effects of social recommendations on network diversity. In: Proceedings of the 2018 World Wide Web Conference, pp. 923–932 (2018)

[93] Traag, V.A., Waltman, L., Van Eck, N.J.: From louvain to leiden: guaranteeing

well-connected communities. Scientific reports **9**(1), 5233 (2019)

[94] Amarasekara, I., Grant, W.J.: Exploring the youtube science communication gender gap: A sentiment analysis. Public Understanding of Science **28**(1), 68–84 (2019)

[95] Patel, R., Passi, K.: Sentiment analysis on twitter data of world cup soccer tournament using machine learning. IoT **1**(2), 14 (2020)

[96] Demir, Y., Ayhan, B.: History or vanity? the first "female" referees in the history of the fifa world cup and its reflections on twitter. Feminist Media Studies **25**(1), 54–71 (2025)

[97] Jaramillo, A.M., Macedo, M., Menezes, R.: Reaching to the top: the gender effect in highly-ranked academics in computer science. Advances in Complex Systems **24**(03n04), 2150008 (2021)

[98] González-Ibánez, R., Muresan, S., Wacholder, N.: Identifying sarcasm in twitter: a closer look. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pp. 581–586 (2011)

[99] Saxena, A., Reddy, H., Saxena, P.: Recent developments in sentiment analysis on social networks: techniques, datasets, and open issues. Principles of Social Networking: The New Horizon and Emerging Challenges, 279–306 (2022)

[100] Jackson, L.A., Ervin, K.S., Gardner, P.D., Schmitt, N.: Gender and the internet: Women communicating and men searching. Sex roles **44**, 363–379 (2001)