Understanding the spread of toxicity in Eredivisie online communities: A phenomenon investigation using NLP and Social Network Analysis"

Tim Zijlstra

Master thesis

Table of Contents

[1. Introduction 2](#_Toc200011238)

[1.2 Main question 3](#_Toc200011239)

[1.3 Sub questions 3](#_Toc200011240)

[2. Literature review 4](#_Toc200011241)

[2.1 Conceptualizing online toxicity 4](#_Toc200011242)

[2.2 Detecting toxicity with Natural Language Processing 5](#_Toc200011243)

[2.3 Toxicity as a social phenomenon 6](#_Toc200011244)

[2.4 Online toxicity in sports and football 7](#_Toc200011245)

[2.5 Toxicity-Triggering Events in Football Discourse 9](#_Toc200011246)

[2.6 Research gap and theoretical synthesis 10](#_Toc200011247)

[3. Methodology 12](#_Toc200011248)

[3.1 Research design 12](#_Toc200011249)

[3.2 Research execution 12](#_Toc200011250)

[3.2.1 Literature Study 12](#_Toc200011251)

[3.2.2 Primary Research 13](#_Toc200011252)

[3.2.3 Secondary Research 13](#_Toc200011253)

[3.3 Data Collection Strategy 14](#_Toc200011254)

[3.3.1 Event Selection Framework 14](#_Toc200011255)

[3.3.2 Data Source and Scraping Tools 15](#_Toc200011256)

[3.3.3 Preprocessing and Filtering 16](#_Toc200011257)

[3.3.4 Data Coverage 16](#_Toc200011258)

[3.3.5 Ethical Considerations 17](#_Toc200011259)

[3.4 Tool Architecture and Development Plan 17](#_Toc200011260)

[3.5 Summary table 19](#_Toc200011261)

[4. Results & analysis 20](#_Toc200011262)

[4.1 Tool overview 20](#_Toc200011263)

[4.2 NLP module performance 21](#_Toc200011264)

[4.2.1 Dutch Classifiers (Fine-Tuned RobBERT Models) 21](#_Toc200011265)

[4.2.2 English classifiers (pre-trained) 22](#_Toc200011266)

[4.2.3 Language coverage and exclusions 22](#_Toc200011267)

[4.3 SNA module & community detection 23](#_Toc200011268)

[4.3.1 Network construction 23](#_Toc200011269)

[4.3.2 Node labelling and toxicity aggregation 24](#_Toc200011270)

[4.3.3 Community detection 25](#_Toc200011271)

[4.3.4 Network metrics and comparison 27](#_Toc200011272)

[4.3.5 Toxicity distribution across clubs 28](#_Toc200011273)

[4.4 Event-based visualizations 28](#_Toc200011274)

[5. Conclusions 29](#_Toc200011275)

[References 30](#_Toc200011276)

[Appendix 35](#_Toc200011277)

# Introduction

In recent years, social media platforms have radically transformed the landscape of public discourse, enabling real-time communication and the formation of online communities centred around shared interests. Within the realm of football, platforms like X (formerly Twitter) have become essential tools for fan engagement, media commentary, and real-time reactions to matches and events (Vermeer & Araujo, 2020; Sheth, Shalin, & Kursuncu, 2022). However, this digital engagement also facilitates the amplification of harmful behaviour, including hate speech, racism, and harassment particularly during emotionally charged moments such as match losses, controversial decisions, or high-stakes derbies (Vidgen et al., 2022; Kearns et al., 2023).

The consequences of such toxicity extend beyond the digital realm. Online abuse directed at players can lead to psychological distress, social withdrawal, and diminished on-field performance (Akar, 2025; Ye et al., 2023). For football clubs and governing bodies, toxic fan behaviour poses reputational risks, weakens brand identity, and can drive away sponsors and fans, particularly those from underrepresented or younger groups (Kearns et al., 2023; Salminen et al., 2020). As a result, managing online toxicity is no longer only a moderation challenge for platforms — it is a strategic priority for clubs and institutions tasked with safeguarding the integrity of the sport.

While NLP has become central to efforts to detect online abuse, most models remain narrowly focused on the classification of isolated messages (Jahan & Oussalah, 2023; Caselli et al., 2021). This approach overlooks the relational and temporal dimensions of toxicity — how toxic content is amplified through replies, retweets, and mentions, and how influence and exposure are shaped by network dynamics (Saveski, Roy, & Roy, 2021; Ribeiro et al., 2023). Research shows that a small subset of highly connected users, often pseudonymous or unaffiliated fan accounts, disproportionately drive the spread of abuse (Schoch, 2021; Cinelli et al., 2020). Traditional moderation systems are ill-equipped to capture this behaviour.

In the Dutch football context, toxicity is further complicated by linguistic and cultural specificity. The term “Joden,” for instance, is used by Ajax fans as a reclaimed identity marker but by rival supporters as a slur, making its classification highly context-dependent (Seijbel et al., 2022). Similarly, Dutch-language insults — such as disease-based profanity (“kankerscheidsrechter”) — may not be detected by classifiers trained on English corpora (Gevers et al., 2022; Burtenshaw & Kestemont, 2021). Dutch football discourse also features frequent code-switching between Dutch and English, sarcastic phrasing, and region-specific slang, adding further complexity to toxicity detection.

To address these challenges, this project proposes the development of a retrospective, event-driven software tool that integrates NLP and SNA. Rather than detecting toxicity in isolation, the tool will analyse how toxic content diffuses across social interaction networks — and how this varies across specific types of emotionally charged football events, such as red cards, derby matches, or controversial referee decisions. By revealing when and how toxic surges occur, and which users or communities are involved in their amplification, the tool will help football clubs, the KNVB, and potentially social media platforms to better understand the structure and triggers of online abuse. This insight can support more strategic interventions, data-informed content moderation, and greater accountability in digital fan engagement.

## 1.2 Main question

“How can a software tool that combines NLP and Social Network Analysis be developed to detect, track, and visualize the spread of toxicity in Eredivisie-related Twitter communities during emotionally charged football events?”

## 1.3 Sub questions

* What types of football events are most likely to trigger toxic surges in Eredivisie-related online discourse?
* How can toxicity in Dutch football-related tweets be accurately detected using a domain-adapted NLP model?
* How can social interaction data (replies, mentions, retweets) be transformed into meaningful network structures for toxicity analysis?
* Which network metrics and diffusion patterns best capture the spread and amplification of toxic behavior during toxic events?
* What are the key design requirements and functionalities of the tool to support visualization and interpretation of toxicity dynamics across events?
* What practical insights can the tool offer to football clubs or governing bodies in understanding and mitigating online toxicity?

# Literature review

To understand how toxicity emerges and spreads within Eredivisie-related social media communities, this chapter explores the academic literature on online toxicity, its detection through natural language processing, and its propagation through social networks. By reviewing both content-based and network-based perspectives, this chapter provides the theoretical foundation for the research methodology and sub-questions. The review begins with an exploration of how toxicity is conceptualized in digital environments, followed by an overview of NLP-based classification approaches and their limitations. It then turns to social network dynamics that shape the diffusion of toxic messages, and concludes with a discussion of literature focused on online toxicity in the context of football and sports more broadly. This chapter identifies key gaps in current research and justifies the integrated use of NLP and Social Network Analysis in this study.

## 2.1 Conceptualizing online toxicity

Online toxicity refers to hostile or disruptive communication in digital environments, including hate speech, cyberbullying, trolling, and verbal aggression (Sheth et al., 2022). It can take explicit forms such as threats or slurs, or more implicit forms like sarcasm, coded language, and rhetorical hostility (Suler, 2004; Ghosh et al., 2020; Almerekhi et al., 2019). While definitions vary, most agree that toxicity involves content that degrades conversation quality or discourages participation (Salminen et al., 2020; Hardaker, 2013). Factors such as platform design, anonymity, and lack of face-to-face accountability contribute to disinhibition and moral disengagement, making users more likely to express harmful views (Suler, 2004; Lapidot-Lefler & Barak, 2012).

Toxicity is not static or universal; rather, it is shaped by discourse context, audience, and speaker intent. Linguistic cues such as profanity, adversarial phrasing, and intensifiers are often ambiguous, with their interpretation depending on tone, community norms, and platform culture (Ghosh et al., 2020; Joshi et al., 2023). This presents a significant challenge for computational detection systems, particularly in informal, multilingual environments where users frequently engage in code-switching, sarcasm, and indirect speech.

Toxic communication can be categorized into overlapping subtypes, including hate speech, abusive language, harassment, trolling, and incivility (Salminen et al., 2020; Pavlopoulos et al., 2020). These forms differ in target and tone: hate speech typically targets individuals or groups based on identity characteristics such as race, religion, or gender, while abusive language includes insults or profanities that may not be identity-based (Ghosh et al., 2021). Harassment involves repeated or targeted abuse, trolling is defined by its intent to provoke, and incivility refers to general rudeness or disrespect in public discourse. These categories often overlap and interact in complex ways, particularly in online communities characterized by performative antagonism or status competition.

Cultural and linguistic context adds further complexity. In the Dutch football discourse, for example, the term “Joden” (Jews) is used both by Ajax fans as a reclaimed identity marker and by rival supporters as a slur—making its toxicity highly context-dependent (Seijbel et al., 2022). Illness-based profanity, such as “kankerscheidsrechter” (literally “cancer referee”), is also common in Dutch fan language, yet such expressions may not be detected by English-centric classifiers (Gevers et al., 2022). These examples highlight the importance of accounting for not only literal word meaning but also speaker-audience dynamics and socio-linguistic conventions (Burtenshaw & Kestemont, 2021).

The psychological consequences of online toxicity are well-documented. Victims of abuse report higher levels of anxiety, depression, and social withdrawal, with youth and marginalized users being especially vulnerable (Ye et al., 2023; Behera et al., 2022). The harm is not only individual but also communal, as toxicity may suppress participation, polarize discourse, and distort public dialogue. Moreover, studies suggest that user behaviour is fluid: individuals may become more radicalized over time or, conversely, adopt more pacifying tones depending on social feedback and network exposure (Mall et al., 2020). These dynamics challenge any view of toxicity as a fixed or isolated phenomenon and instead position it as a process shaped by social and communicative factors (Ribeiro et al., 2023; Suler, 2004).

Perceptions of toxicity also vary across cultural and demographic lines. A phrase considered offensive in one setting may be benign in another, depending on shared norms and group expectations. Research shows meaningful variation in how toxicity is perceived based on gender, nationality, and political orientation (Salminen et al., 2020; van Geel et al., 2024). As such, conceptualizing online toxicity requires attention to linguistic variation, cultural context, identity dynamics, and broader communication patterns. These elements form the foundation for any attempt to detect, classify, or understand toxic behaviour in a nuanced and socially aware manner.

## 2.2 Detecting toxicity with Natural Language Processing

The detection of toxic content at scale has become a central challenge in computational linguistics and content moderation. Natural Language Processing (NLP) enables automated systems to identify hate speech, harassment, insults, and other harmful forms of communication by recognizing linguistic patterns in user-generated text (Jahan & Oussalah, 2023; Abbasi et al., 2022). These approaches aim to flag content that may violate community standards, threaten user safety, or degrade public discourse.

Initial detection systems were rule-based, relying on curated keyword lists or hand-crafted linguistic rules. While these methods were relatively transparent and computationally inexpensive, they lacked nuance, leading to high false-positive rates—especially when processing sarcasm, irony, or indirect speech (Davidson et al., 2017). These limitations motivated the development of supervised learning approaches trained on annotated corpora, which could better capture variation in tone and phrasing.

Supervised classifiers such as logistic regression and random forest models were soon followed by more sophisticated deep learning architectures. These models, trained on datasets such as Wikipedia Detox, Civil Comments, and the Dutch Abusive Language Corpus (DALC), provided substantial improvements in detecting offensive content across domains (Caselli et al., 2021; Burtenshaw & Kestemont, 2021). However, model performance remains highly dependent on the quality and representativeness of the training data. Datasets that are skewed toward certain topics, demographics, or dialects may result in biased or incomplete detection.

Transformer-based language models such as BERT, mBERT, and RobBERT have become the dominant approach in toxicity classification tasks. Pre-trained on large corpora and fine-tuned for specific applications, these models are capable of capturing syntactic and semantic nuance, making them particularly well-suited for identifying subtle and embedded forms of toxicity (Delobelle et al., 2020). When fine-tuned on domain-specific data, they offer strong performance across a range of classification types, including multilabel tasks that account for overlapping categories such as hate speech, threats, and insults.

Nevertheless, several challenges persist. One of the most fundamental is the subjectivity of toxicity labels. What is considered toxic can vary widely across annotators based on their cultural background, political stance, or tolerance thresholds (Sap et al., 2019; van Aken et al., 2018). This subjectivity leads to inconsistencies in labeled datasets, which in turn undermines classifier reliability. Studies have also shown that standard toxicity classifiers are more likely to flag content from marginalized dialects or minority users as offensive, raising concerns about fairness and algorithmic bias (Ghosh et al., 2021; Davani et al., 2023). Recent work advocates for the use of fairness-aware training procedures and more inclusive annotation guidelines (Sahoo et al., 2022).

Another methodological limitation is that most classifiers operate on individual messages without considering conversational context. Yet toxic behaviour often unfolds over time and is shaped by prior interactions or group dynamics. Models that incorporate thread structure, user history, or temporal metadata tend to achieve higher accuracy, especially when distinguishing isolated outbursts from coordinated harassment or pile-on behaviour (Almerekhi et al., 2019; Ribeiro et al., 2023). Hybrid approaches that combine NLP with graph-based social features or sentiment trajectories offer promising avenues for capturing the relational nature of online toxicity (Wang et al., 2022; Qian et al., 2023).

Finally, while transformer models can be fine-tuned for specific domains, performance often drops when applied across cultural or linguistic contexts different from their training data. Classifiers trained on English-language corpora may struggle with the informal, multilingual, and highly contextual nature of football-related discourse—particularly in communities where slang, irony, or subcultural references are common. This underscores the need for domain-adapted models that are trained on locally relevant corpora and sensitive to the linguistic norms of the communities under study (Burtenshaw & Kestemont, 2021; Gevers et al., 2022).

In sum, NLP has made substantial progress in enabling scalable toxicity detection, but remains limited by bias, context blindness, and cross-cultural sensitivity. These limitations highlight the importance of combining linguistic analysis with contextual signals, and of tailoring models to the specific discourse environments in which toxicity occurs.

## 2.3 Toxicity as a social phenomenon

While much of the research on online toxicity focuses on message content, recent work highlights the importance of examining how harmful discourse spreads within social structures. On platforms such as X (formerly Twitter), users are embedded in interaction networks formed through replies, mentions, and retweets. These structures influence which content gains traction, which users become influential, and how sentiment—whether supportive or antagonistic—circulates and escalates (Saveski, Roy, & Roy, 2021; Ribeiro et al., 2023).

Social Network Analysis (SNA) offers a robust framework for studying these relational dynamics. Originally developed in sociology and now widely adopted in computational social science, SNA focuses on the connections between actors rather than the attributes of individuals alone (Wasserman & Faust, 1994). In the context of online toxicity, SNA makes it possible to map communication flows, identify structural hubs or bridges, and detect communities that may reinforce harmful behaviour (Cinelli et al., 2020; Himelboim et al., 2017).

Users do not interact in isolation. Research has shown that individuals are more likely to post provocative content when they anticipate positive reinforcement—such as likes, retweets, or replies—from peers (Ribeiro et al., 2023). These dynamics are intensified in tightly knit networks, where toxic norms can become entrenched and toxic messages may be amplified more quickly. Clusters that function as echo chambers can limit exposure to opposing views, reinforcing in-group attitudes and legitimizing antagonism toward out-groups (Cinelli et al., 2020).

SNA provides a range of tools for quantifying these structures. Degree centrality highlights users who are most active or visible; betweenness centrality identifies users who bridge otherwise disconnected communities; and clustering coefficients measure how tightly knit particular user groups are (Himelboim et al., 2017; Weller et al., 2014). These metrics help to identify not only the originators of toxicity but also the users who serve as accelerators or structural enablers of its spread. Community detection algorithms such as modularity-based clustering can isolate subgroups where toxicity may persist, evolve, or intensify over time.

Studies applying SNA to political discourse and sports communities alike have found that toxic content often spreads along clear social lines, following shared interests, identities, or affiliations (Buongiovanni et al., 2023; Ribeiro et al., 2023). In football-related contexts, this often manifests through fan clusters who engage in rivalry-driven discourse, retweeting inflammatory posts, or piling on against a target user. Influential accounts such as fan pages or prominent commentators can serve as ignition points for these cascades by signalling blame or mocking individuals.

Despite its analytical strengths, SNA also has limitations. Social media network data is often incomplete, particularly when collected retrospectively or when relying on public APIs that provide only partial access to platform content (Morstatter et al., 2013; Weller et al., 2014). Additionally, most network constructions rely on visible interaction data—such as replies, retweets, or mentions—which exclude passive participants: users who observe toxic conversations but do not actively engage. These "lurkers" can still be influenced by toxic discourse or silently support it, but remain invisible in most datasets (Gyarmati & Trinh, 2010). Finally, centrality measures require contextual interpretation. High degree or betweenness centrality does not always indicate influence or culpability; prominent users may serve as moderators, targets, or neutral conduits rather than perpetrators of toxicity (Kiddle et al., 2024).

Nevertheless, SNA remains a powerful approach for analysing the relational dynamics of online harm. By shifting the unit of analysis from individual messages to interaction patterns, SNA captures how toxicity is embedded within broader social processes. In the context of football communities, it offers critical insight into the amplification, normalization, and clustering of toxic content—complementing content-level analysis and enabling a more holistic understanding of digital aggression.

## 2.4 Online toxicity in sports and football

The world of professional football has become a prominent site of online toxicity, where players, referees, clubs, fans, and journalists are regularly subjected to verbal abuse on platforms such as Twitter, Instagram, and YouTube. Unlike toxicity in political discourse, which tends to be sustained and ideological, football-related toxicity is often reactive, emotionally charged, and triggered by specific match events such as controversial decisions, poor performances, or incidents involving race, gender, or protest (Vidgen et al., 2022; Kilvington & Price, 2020; Behera et al., 2022).

In elite leagues such as the English Premier League, the scale and severity of online abuse are well documented. Vidgen et al. (2022) found that 68% of players were subjected to toxic tweets during the 2021–2022 season, with racial abuse making up a substantial portion of the overall harm. Abuse often intensifies following high-stakes matches, missed penalties, or disciplinary decisions, especially when involving high-profile players. Similar trends are found in the treatment of match officials: Mojtahedi et al. (2024) show that referees are frequent targets of online harassment, with abuse increasing in visibility and intensity after contentious calls or VAR controversies.

Toxicity is not limited to actors on the pitch. Female sports journalists and fans are disproportionately targeted with gendered insults and sexualized threats. Johnson et al. (2024) demonstrate that although both male and female reporters face hostility, women experience significantly more personalized and violent abuse. Research by Kearns et al. (2023) highlights how misogynistic backlash toward women’s football and female supporters has grown, particularly on platforms where anonymity and group dynamics enable aggressive posting.

Football communities are shaped by strong identities and tribal loyalties, which influence how fans respond to abuse or accusations of misconduct. Cable et al. (2022) found that when faced with criticism, many fans deflect or rationalize toxic behaviour as a form of club loyalty or banter. These responses often downplay harm and foster a culture of impunity, where targeted harassment is normalized and even celebrated in certain online subcultures.

Social media fan communities often operate as echo chambers, reinforcing in-group norms and marginalizing dissenting views. This is especially visible in football Twitter spaces, where pseudonymous users adopt club-themed avatars and personas to engage in emotionally charged posting (Cleland, 2014; Mall et al., 2020). While much of this content is framed as humorous or performative, it can easily slip into harassment or hate speech—particularly after emotionally intense matches. Toxic discourse is further amplified by prominent accounts such as fan pages, influencers, or journalists, who can shape narrative direction, assign blame, and trigger reply-chain cascades (Ribeiro et al., 2023; Buongiovanni et al., 2023).

These abuse cascades often follow predictable social patterns. Influential accounts—intentionally or not—can instigate "pile-on" dynamics by signalling outrage or drawing attention to a specific incident. As retweets and replies multiply, the volume of abuse escalates rapidly, echoing the logic of digital mobs that are loosely coordinated but emotionally synchronized (Mall et al., 2020). Platform algorithms that prioritize virality can unintentionally reward such behaviour, further increasing its visibility and reach.

In the Dutch context, football-related toxicity exhibits its own linguistic and cultural dynamics. Seijbel et al. (2022) explored how antisemitic language—historically associated with stadium chants—has migrated online in the form of coded slurs and meme-based expressions. Their research shows that Ajax’s perceived Jewish identity is often weaponized by rival fans, especially during matches against Feyenoord. A follow-up study found that such rhetoric worsened during the COVID-19 lockdowns, when physical fan engagement moved online (Seijbel et al., 2023). These findings highlight the importance of culturally informed models when analysing online football discourse in the Netherlands.

Despite increasing media attention to racism and abuse in Dutch football, academic research on the structure and diffusion of online toxicity in the Eredivisie remains limited. Most existing studies focus on stadium behaviour or general societal polarization, leaving a gap in understanding how online toxicity is triggered, spread, and normalized in football-specific digital environments. High-profile campaigns by clubs such as Ajax to combat online hate (NOS, 2025) demonstrate growing awareness of the issue, but research that integrates network analysis and NLP remains scarce.

By situating football toxicity within broader discursive, emotional, and structural frameworks, this section underscores the need for analytical tools that capture both content and context. Understanding toxicity in this domain requires attention not only to linguistic features but also to rivalry dynamics, fan subcultures, and the social architecture of interaction. This is especially relevant in the Eredivisie, where intense match events and deep-rooted cultural narratives converge to shape how toxic messages are expressed and received online.

## 2.5 Toxicity-Triggering Events in Football Discourse

A defining characteristic of online toxicity in football communities is its temporal and emotional volatility. Rather than being evenly distributed, spikes in abusive content typically follow emotionally charged events—moments that activate intense fan sentiment, collective frustration, or perceived injustice. These surges often involve sudden increases in message volume, heightened negativity, and the rapid spread of hostile narratives (Vidgen et al., 2022; Mall et al., 2020; Ribeiro et al., 2023). Understanding which types of events trigger these spikes is essential for analysing football-related toxicity as an event-driven and context-sensitive phenomenon.

Empirical studies in the Premier League and broader sports discourse confirm that toxicity escalates during specific moments, especially around performance failures, disciplinary decisions, and identity-related incidents. Vidgen et al. (2022) found that missed penalties, red cards, and losses were associated with sharp increases in abusive tweets, with the most toxic reactions occurring after controversial match outcomes. Mall et al. (2020) similarly observed that online toxicity intensified within minutes of emotionally salient match events, such as unfair referee calls or last-minute goals. These spikes were amplified when fans perceived the outcome as unjust or when expectations were violated.

Drawing on this research, the following categories of toxicity-triggering events can be distinguished:

1. **Red cards and suspensions**

Disciplinary incidents, especially when seen as unjust or inconsistent, often provoke outrage. Fans may direct abuse at referees, players, or governing bodies, leading to spikes in aggressive language and antagonistic retweets (Mojtahedi et al., 2024; Mall et al., 2020).

1. **Referee decisions and VAR controversies**

Officiating decisions—such as disallowed goals, controversial penalties, or VAR delays—are among the most common catalysts for online harassment, particularly toward officials. These moments frequently spark viral hashtags, meme-sharing, and reply-chain aggression (Kearns et al., 2023; Mojtahedi et al., 2024).

1. **Derby matches and historic rivalries**

Games like Ajax vs. Feyenoord carry symbolic and emotional weight. These fixtures often feature identity-based insults, national or regional slurs, and intensified online rivalries, reflecting both sporting tension and cultural conflict (Seijbel et al., 2022; Buongiovanni et al., 2023).

1. **Player errors or perceived lack of effort**

Missed chances, own goals, or visible frustration can prompt fans to scapegoat individual players, especially if they are polarizing figures or hold leadership roles. Abuse may take the form of direct insults, memes, or threats (Vidgen et al., 2022).

1. **Incidents involving racism, sexism, or protest**

Events involving political gestures (e.g., players kneeling), anti-racism campaigns, or symbolic apparel (e.g., armbands) often polarize communities. Studies show that such incidents activate echo chambers and spur sharp increases in toxic content (Buongiovanni et al., 2023; Kilvington & Price, 2020).

These event types are not only well-documented triggers of online abuse, but they also reflect deeper emotional and identity-based mechanisms. Toxicity tends to escalate when fan expectations are violated or when symbolic meaning is attached to particular match moments—whether through cultural rivalries, disciplinary injustice, or broader social tensions (Mall et al., 2020; Buongiovanni et al., 2023). Moreover, the way toxicity emerges in response to these incidents is shaped by platform-specific dynamics, such as the speed of diffusion via retweets and the amplification of emotional content by engagement-based algorithms (Ribeiro et al., 2023).

In the Dutch Eredivisie, these dynamics are likely to be influenced by local linguistic and cultural factors. Previous studies highlight the prevalence of illness-based profanity, coded antisemitism, and rivalry-specific slurs, particularly during high-stakes or emotionally intense matches (Seijbel et al., 2022; Gevers et al., 2022). Such findings suggest that toxicity in Dutch football communities is not only reactive but also deeply embedded in culturally specific forms of expression.

Taken together, the literature underscores that football toxicity is event-driven, emotionally charged, and context-dependent, calling for analytical approaches that combine temporal, textual, and social dimensions. These insights lay the foundation for understanding when toxicity is most likely to appear and how it may evolve within the structure of football discourse networks.

## 2.6 Research gap and theoretical synthesis

The preceding sections have shown that online toxicity is a multifaceted phenomenon, shaped by linguistic ambiguity, cultural norms, social dynamics, and platform-specific interactions. Toxicity does not emerge in isolation but is embedded in complex communicative and relational contexts. In football communities, these dynamics are particularly pronounced due to the intensity of fan identity, rivalry, and emotional investment in match events.

Section 2.1 established that toxicity can take both explicit and implicit forms and that its interpretation is highly context-dependent. Cultural and linguistic variation—especially in multilingual environments like Dutch football Twitter—complicates efforts to detect toxicity using generic or monolingual tools. Section 2.2 further highlighted the strengths and limitations of current NLP approaches, particularly the challenge of capturing sarcasm, code-switching, and local slang without domain-specific training. These limitations underline the importance of using culturally adapted and context-aware models, rather than relying on off-the-shelf classifiers.

While NLP provides insight into the content of toxic messages, Section 2.3 demonstrated the value of Social Network Analysis in understanding the relational structures that support their spread. Toxicity often gains traction not solely because of its linguistic content, but because it is embedded in networks of amplification, validation, and emotional contagion. SNA offers tools for identifying key actors, subcommunities, and interaction patterns that contribute to the normalization and diffusion of harmful discourse.

Sections 2.4 and 2.5 narrowed the focus to the domain of football, where online toxicity is not only frequent but also highly event-driven. Football discourse is marked by emotional spikes that follow key match events—such as red cards, referee decisions, or symbolic protests—and these moments often trigger surges in targeted abuse. The literature emphasizes that such toxicity is reinforced by platform dynamics and cultural narratives specific to the footballing context, particularly in the Dutch Eredivisie.

Taken together, these findings suggest that existing research often addresses either content or network structure, but rarely both in an integrated and event-sensitive manner. This creates a gap in the literature: while we know that toxicity is context-dependent and socially mediated, few studies offer a combined methodological approach that reflects this complexity.

This thesis responds to that gap by synthesizing these perspectives into a multi-dimensional framework for detecting and analysing toxicity. It does so by integrating:

* **Natural Language Processing**, to detect toxic language with sensitivity to Dutch football-specific discourse;
* **Social Network Analysis**, to trace how toxicity spreads across interaction networks; and
* **Event-based analysis**, to understand when toxicity emerges and what types of incidents are most likely to trigger it.

By combining these dimensions, the study aims to provide a more comprehensive understanding of online toxicity within Eredivisie football communities—one that accounts for both the content of harmful messages and the social architecture through which they are produced, validated, and amplified.

# Methodology

## 3.1 Research design

This thesis adopts a tool development–oriented research design, aimed at building a prototype that combines Natural Language Processing (NLP) and Social Network Analysis (SNA) for the retrospective analysis of online toxicity in Dutch football-related Twitter communities. Rather than providing a theoretical explanation of how toxicity emerges, the tool is designed to detect toxic messages and visualize their interaction dynamics in a network structure.

The tool focuses on event-based analysis, applying the pipeline to Eredivisie matches known to provoke emotionally charged reactions, such as red cards, controversial referee decisions, derby matches, or player errors. It enables retrospective examination of toxicity trends surrounding these moments, as well as the structure of user interactions involved in amplifying toxic content.

The design is rooted in principles from computational social science and applied NLP, emphasizing reproducibility, modularity, and practical usability for stakeholders such as researchers, moderators, or football institutions. The pipeline follows a clear input–processing–output structure: it collects tweets, classifies them using a Dutch-language toxicity model, constructs interaction graphs, and computes network metrics to support interpretation.

The system does not aim to explain intent or social meaning, but rather to provide structured outputs that allow for further analysis and reflection on how toxic discourse materializes during emotionally salient football events.

## 3.2 Research execution

### 3.2.1 Literature Study

The methodological approach was grounded in three key areas of academic literature that shaped both the rationale and structure of the tool.

First, studies on toxic content detection through NLP (see Section 2.2) emphasized the limitations of English-only models when applied to localized, multilingual contexts such as Dutch football Twitter. Based on this, the Dutch-language transformer model RobBERT was selected and fine-tuned using the Dutch Abusive Language Corpus (DALC). This dataset was chosen for its cultural fit and multilabel structure, allowing detection of various forms of toxicity (e.g., insults, identity attacks, threats), though it was acknowledged that it may not capture all forms of football-specific or sarcastic abuse.

Second, literature on toxicity as a social phenomenon (Section 2.3) motivated the use of Social Network Analysis. Prior work has shown that toxic content is not only expressed individually but amplified through replies, retweets, and tightly clustered user groups. To capture these relational dynamics, the tool includes graph-based modelling and network metrics that highlight structural aspects of toxicity diffusion.

Third, research on emotionally charged events in football discourse (Section 2.5) informed the event selection strategy. Prior studies have shown that online abuse frequently spikes during matches involving controversial decisions, disciplinary actions, or strong rivalries. These findings justified the use of retrospective, event-based scraping windows, focusing on moments most likely to generate toxic reactions.

Taken together, the literature did not only offer theoretical grounding but also directly informed technical design choices from classifier selection and training, to network modelling, to the event-driven structure of the data pipeline.

### 3.2.2 Primary Research

The primary research involved designing, implementing, and testing a modular software pipeline for detecting and analysing online toxicity. The tool processes real-world Twitter data surrounding Dutch Eredivisie football matches, transforming it into structured outputs for NLP and SNA analysis.

The implementation followed an iterative development cycle, consisting of the following stages:

* A **scraping system** built using Selenium and Nitter, used to collect tweets, replies, and second-level replies from club accounts and surrounding fan interactions.
* A **toxicity classification model**, created by fine-tuning RobBERT on DALC, to label tweets as toxic or non-toxic across multiple dimensions.
* A **graph construction module**, which converts reply metadata into directed interaction networks per event, with nodes as users and edges as interactions.
* A **network analysis module**, applying standard SNA metrics such as centrality and clustering to identify amplification patterns.

Each module was tested independently before being integrated into an end-to-end pipeline. The tool was applied to multiple match events during the 2024–2025 Eredivisie season, demonstrating functionality across varied discourse contexts. Modularity allowed for flexible updates (e.g., improved filters, stricter thresholds) without needing to rebuild the entire system.

### 3.2.3 Secondary Research

Secondary data sources were used to support two core aspects of the methodology: event selection and classifier development.

To select relevant matches, the project consulted football news outlets, fan commentary, and live coverage (e.g., NOS, Voetbal International, ESPN). These sources helped identify emotionally charged events, such as red cards, referee controversies, derby matches, and high-profile errors ensuring that scraping windows aligned with moments most likely to provoke toxic reactions.

For the NLP component, the Dutch Abusive Language Corpus (DALC) was chosen as the training and validation dataset. While larger multilingual corpora were considered, DALC was more culturally appropriate for the Dutch-language content typical of Eredivisie Twitter discourse. This choice prioritized linguistic fit over dataset size, accepting that some edge cases (e.g., niche slang, memes) may fall outside the model’s coverage.

Together, these secondary sources ensured that the tool operated in alignment with both real-world football discourse and domain-relevant toxicity patterns.

## 3.3 Data Collection Strategy

### 3.3.1 Event Selection Framework

This study adopts an event-driven data collection strategy, focusing on Eredivisie football matches from the 2024–2025 season that were most likely to trigger toxic online discourse. Rather than collecting data continuously, events were selected based on empirical categories identified in the literature (see Section 2.5), which shows that spikes in online abuse often follow emotionally charged incidents.

Initially, five types of toxicity-triggering events were considered:

1. Red cards and disciplinary incidents
2. Controversial referee or VAR decisions
3. Derby matches and long-standing rivalries
4. Player errors or perceived underperformance
5. Social protest events (e.g., symbolic gestures)

However, during the 2024–2025 season, no notable social protest incidents occurred, so this category was excluded. The final selection included only the first four event types.

To ensure accuracy in scheduling, all fixtures were obtained from Sofascore, a live football tracking platform that reflects real-time changes to match dates. This was necessary because the official KNVB calendar often does not account for rescheduled matches due to European competitions or recovery periods.

The process of identifying relevant matches was conducted manually, using different sources and criteria for each event type:

* **Red cards** were identified by reviewing match reports on Sofascore.
* **Controversial referee or VAR decisions** were flagged based on coverage in Dutch media outlets such as NOS, AD, and Voetbal International. Because there are no fixed criteria for what qualifies as controversial, this step required subjective judgment based on patterns in post-match reporting and headlines.
* **Derby matches and rivalries** were selected using a predefined list from ESPN, which identified fixtures with historical or cultural significance.
* **Player errors or blame incidents** were identified through manual review of match summaries and media reactions highlighting individual mistakes, such as missed penalties or own goals.

Each selected event was assigned a fixed 36-hour scraping window:

* 24 hours before kick-off, to capture anticipation, line-up reactions, and pre-match discourse
* 12 hours after full time, to capture emotional reactions and early fallout

This window reflects prior findings that toxicity is most likely to peak immediately around emotionally intense events and tends to decay rapidly after.

While the manual selection approach introduced an element of subjectivity, especially for categories like referee controversy or player blame, it allowed for context-sensitive, real-time alignment with the narrative of each match. Automated detection of toxic incidents remains unreliable and outside the scope of this tool-focused project, making manual selection a pragmatic and appropriate choice.

### 3.3.2 Data Source and Scraping Tools

To collect tweets related to the selected matches, a custom scraping pipeline was developed using Selenium in Python. Instead of using the official X (formerly Twitter) API, which had become rate-limited and commercially restricted, the scraper accessed data through Nitter, an open-source, privacy-friendly frontend for X.

Nitter was selected because it offered a technically feasible and ethically sound alternative to the API. First, it allowed stable access to both timeline tweets and threaded replies without requiring authentication or API credentials, essential for large-scale scraping. Second, by relying solely on public content, it ensured that the project remained within the bounds of ethical research standards and platform policies. Finally, unlike the official API, which had shifted key endpoints behind a paywall, Nitter provided unrestricted access suitable for retrospective, event-based analysis.

The scraping process targeted:

* Timeline tweets posted by Eredivisie club accounts (e.g., @AFCAjax, @PSV, @Feyenoord).
* Replies to those tweets, posted by fans or other users.
* Second-level replies, i.e. replies to fan replies, when they had visible comment counts.

To ensure consistent execution, scraping was automated using a match metadata file, which contained all relevant identifiers and timestamp windows for each event. The system operated in batch mode and followed three sequential stages: first, scraping the timeline tweets from each club during the 36-hour event window; second, collecting replies to those tweets; and third, scraping second-level replies if the first-level messages showed signs of continued discussion.

Data was saved in structured CSV format, with separate folders for each layer (timelines, replies, and second-level replies) to support modular preprocessing and analysis.

A key limitation of the scraping process is that it captures only two layers of reply depth. Early exploration showed that third-level replies were rare and sparsely distributed, while dramatically increasing scraping time. Given the project's focus on the most immediate forms of toxic interaction, particularly fan reactions to club tweets, this limitation was accepted as a practical and analytically sound trade-off.

In sum, the scraping setup provided a scalable, compliant, and context-sensitive solution for collecting Dutch football discourse tied to emotionally charged Eredivisie matches.

### 3.3.3 Preprocessing and Filtering

After scraping tweets, replies, and second-level replies for each match, the data was prepared for classification and network analysis through a structured preprocessing pipeline. The focus of this step was not to clean or alter the content itself, but to merge, label, and filter the data consistently while preserving the original language and emotional tone of each message.

The first step involved merging all collected tweets into a single dataset per match. This included timeline tweets from club accounts, first-level replies from fans, and second-level replies when available. Each message was tagged with metadata specifying the match ID, the club account it originated from, and the interaction type (timeline, reply, second-level reply). This unified structure made it possible to link each message to a specific event, team, and context for analysis.

Language filtering was applied selectively. While timeline tweets from verified club accounts were kept in full, replies and second-level replies were filtered to include only messages written in Dutch or English. This was done to align with the capabilities of the toxicity classifier, which had been fine-tuned using Dutch-language data but could also reasonably handle English-language input, a common occurrence in Dutch football discourse. Messages in other languages or those with low textual content were excluded.

Importantly, no content cleaning or token-level preprocessing was applied to the tweets. Elements such as hashtags, emojis, user mentions, punctuation, URLs, and capitalization were intentionally preserved. This decision reflects the nature of football-related toxicity, which often relies on slang, sarcasm, intensifiers, and stylized expression. Stripping these elements would have weakened the classifier’s ability to detect subtle or culturally specific forms of abuse.

To enable temporal and event-based analysis, timestamps were standardized and compared against the predefined scraping windows for each match (24 hours before kick-off to 12 hours after). Any messages that fell outside of this window were discarded. This ensured that all retained data reflected discourse directly surrounding the match itself, rather than unrelated or outdated content.

Finally, to uphold ethical research standards, all personally identifiable user information, including usernames and direct tweet URLs, was anonymized or excluded from analysis outputs. No private, deleted, or restricted tweets were collected at any point in the process.

In summary, the preprocessing step focused on structuring, filtering, and labelling the data while intentionally preserving its original form. This approach ensured the dataset was linguistically and contextually appropriate for both toxicity classification and network-based analysis, without stripping away the expressive features that define real football discourse online.

### 3.3.4 Data Coverage

Data collection focused on matches from the 2024–2025 Eredivisie season that had already been played at the time of scraping. In total, 29 match weeks were available, covering the majority of the season up to the point of data extraction.

From these match weeks, a total of 79 matches were selected for scraping. Each match was manually identified as featuring at least one of the four toxicity-triggering event types defined in the literature: red cards, controversial referee decisions, derby rivalries, or player blame incidents. The inclusion of only these types ensured that the dataset reflected emotionally charged football moments, where spikes in online toxicity were most likely to occur.

For each of the 79 matches, tweets were scraped from both the home and away team accounts, along with replies and second-level replies. All data was tied to a specific match ID and stored with full temporal and structural metadata, making it possible to filter and compare across matches.

While the scraping did not include the entire Eredivisie season, this targeted approach provided a high-density sample of events likely to provoke toxic reactions. It allowed the tool to be tested across a range of match types and emotional contexts, including derbies, high-stakes games, and moments of perceived injustice, rather than spread thinly over uneventful fixtures.

As expected, the volume of tweets and replies varied per match. High-profile derbies and disciplinary controversies generated significantly more fan interaction than quieter matchdays. However, this variability is a meaningful feature of the dataset: it enables comparison between matches with different levels of toxic activity and demonstrates the tool's ability to handle both high- and low-volume contexts.

By focusing only on matches associated with literature-backed event types, and collecting data during fixed scraping windows, the dataset remains balanced, relevant, and event-specific. It offers sufficient breadth to test the full functionality of the NLP and SNA modules, while maintaining analytical focus on the moments where toxicity is most likely to appear.

### 3.3.5 Ethical Considerations

This research project follows strict ethical standards in the collection, processing, and use of social media data. All tweets included in the dataset were publicly accessible and collected using open-source tools that respect platform policies and do not circumvent privacy settings. No private, protected, or deleted content was scraped at any point.

To protect individual users, all identifying information, including usernames, user IDs, and direct tweet URLs, was anonymized or removed during preprocessing. The analysis focuses on message content and interaction patterns, not individual actors.

All data was stored securely and used solely for academic research purposes in accordance with the General Data Protection Regulation (GDPR). Ethical approval procedures followed university guidelines. In reporting results, care is taken to avoid the public attribution of toxicity to specific users, fan groups, or clubs.

While social media data is publicly visible, the project recognizes the importance of minimizing harm and avoiding unnecessary exposure of individuals. The analysis aims to inform strategies for moderation and community health, not to stigmatize users.

## 3.4 Tool Architecture and Development Plan

To meet the research objectives, this project resulted in the development of a modular software tool that integrates Natural Language Processing (NLP) and Social Network Analysis (SNA) for the retrospective analysis of toxic discourse in Dutch football-related Twitter communities. The tool is designed to process public Twitter data surrounding emotionally charged Eredivisie matches and produce structured outputs to support classification, network modelling, and event-based interpretation.

The architecture consists of five interconnected modules, each aligned with a specific stage of the analysis pipeline:

1. **Data collection module**

Tweets are collected using a custom-built scraping pipeline implemented in Python with Selenium, operating through the Nitter frontend. This module collects timeline tweets from Eredivisie club accounts, first-level replies from fans, and second-level replies when relevant. Scraping is based on fixed event windows, using match metadata (e.g., club handles, match IDs, timestamps) to drive batch-mode scraping.

1. **Toxicity classification module**

A RobBERT-based Dutch-language transformer model, fine-tuned on the Dutch Abusive Language Corpus (DALC), is used to classify tweets as toxic or non-toxic. The classifier supports multilabel outputs, identifying forms such as insults, identity attacks, and general abusive language. It preserves the original text (including hashtags, emojis, and slang) to maximize contextual sensitivity. Classifier outputs include toxicity labels and confidence scores per message.

1. **Network construction module**

Interaction data (replies, mentions) is transformed into directed graphs where each node represents a user and each edge an interaction (e.g., replying to or mentioning another user). Graphs are generated separately for each match or event, enabling temporal and structural comparisons. Edges are enriched with metadata such as timestamp, interaction type, and toxicity status.

1. **Network analysis module**

This module applies key SNA metrics to the constructed graphs, including:

* + Degree centrality to identify active or visible users
  + Betweenness centrality to detect users bridging subcommunities
  + Clustering coefficients to measure tightness of interaction clusters
  + Community detection to identify user groups engaged in shared toxic behaviour

These metrics support the detection of influential users and structural amplification of toxic content, without implying causality or blame.

1. **Visualization and reporting module**

The final module prepares outputs for interpretation using Python-based tools such as Plotly or Dash. It enables the display of time-based toxicity spikes, network visualizations per match, and filtered graphs by toxicity type or user group. These visuals are intended to support non-technical stakeholders such as digital moderators, researchers, or football organizations.

The tool was developed using an iterative workflow, starting with standalone prototypes for scraping and classification, followed by integration and refinement of graph-based components. Each module was tested independently before being validated as part of the full pipeline. The final system was applied to a set of 79 matches to evaluate performance under varied match contexts.

Throughout development, emphasis was placed on modularity, traceability, and interpretability, enabling the system to scale, adapt, and produce outputs that are both technically robust and actionable for future use.

## 3.5 Summary table

|  |  |
| --- | --- |
| **Sub-questions** | **Method(s) used** |
| What types of football events are most likely to trigger toxic surges in Eredivisie-related online discourse? | Literature review on toxicity triggers in football |
| How can toxicity in Dutch football-related tweets be accurately detected using a domain-adapted NLP model? | Fine-tuning of RobBERT using DALC |
| How can social interaction data (replies, mentions, retweets) be transformed into meaningful network structures for toxicity analysis? | Graph construction using NetworkX/igraph |
| Which network metrics and diffusion patterns best capture the spread and amplification of toxic behavior during toxic events? | SNA |
| What are the key design requirements and functionalities of the tool to support visualization and interpretation of toxicity dynamics across events? | Interface prototyping and dashboard design using visualization tools |
| What practical insights can the tool offer to football clubs or governing bodies in understanding and mitigating online toxicity? | Integration of NLP and SNA outputs |

# Results & analysis

## 4.1 Tool overview

This thesis documents the development of a modular software tool designed to support the detection, visualization, and structural analysis of toxic discourse in Dutch football-related Twitter communities. Rather than conducting a substantive investigation of toxicity itself, the primary aim is to build a functional and extensible tool that enables researchers, moderators, and football institutions to examine toxic interactions surrounding emotionally charged Eredivisie matches.

The tool integrates multiple components into a cohesive analytical pipeline, combining tweet collection, toxicity classification, interaction network construction, and interactive visualization. It is designed to operate retrospectively on match-specific social media data, allowing users to explore how toxic content unfolds during selected football events.

The architecture consists of four main modules:

**Scraping module**

Tweets are collected using a Selenium-based scraper that accesses data through the Nitter frontend. This approach enables the retrieval of timeline tweets from official Eredivisie club accounts, along with first- and second-level replies from fans and other users. Data is collected within 36-hour windows around matches that are likely to provoke strong reactions, such as those involving red cards or controversial decisions. These scraping parameters reflect prior research showing that online toxicity tends to spike in response to specific match events (Vidgen et al., 2022; Mall et al., 2020).

**Toxicity classification module**

The classification module uses a RobBERT-based transformer model that has been fine-tuned on the Dutch Abusive Language Corpus (Burtenshaw & Kestemont, 2021). The model performs multilabel classification to identify different forms of toxicity, including insults, identity attacks, and threats. It is optimized for the informal and linguistically diverse nature of Dutch football discourse, retaining social features such as emojis and hashtags to preserve contextual nuance (Caselli et al., 2021; Ghosh et al., 2020).

**Network construction and analysis module**

Reply interactions are used to construct directed graphs in which nodes represent users and edges represent reply relationships. Each graph corresponds to a single match event and includes metadata such as timestamps, toxicity labels, and user roles. The tool supports the application of standard network metrics—such as degree centrality, betweenness centrality, and clustering coefficients—as well as modularity-based community detection. These features allow end users to examine structural patterns in how toxicity may be clustered or amplified across subgroups (Himelboim et al., 2017; Ribeiro et al., 2023).

**Interactive visualization module**

To ensure usability for a variety of stakeholders, the tool includes an interactive dashboard that allows users to explore match-level toxicity trends, inspect interaction graphs, and filter content by user type or toxicity level. The interface is designed to be accessible to both technical and non-technical users, with visual summaries that support interpretation and decision-making.

Taken together, these modules provide a comprehensive and reusable framework for analysing toxic discourse in football-related social media contexts. The tool is designed with scalability and modularity in mind, making it suitable for future adaptation to other domains, languages, or social media platforms.

## 4.2 NLP module performance

The classification component of the tool is designed to detect toxic language in replies to Eredivisie club tweets. It operates by applying two sets of transformer-based models: one set fine-tuned for Dutch-language classification, and another set used in pre-trained form for English-language content. This dual structure accommodates the multilingual nature of football discourse in the Netherlands, which often includes a mix of Dutch, English, and informal or code-switched expressions.

### 4.2.1 Dutch Classifiers (Fine-Tuned RobBERT Models)

For Dutch tweets, two separate transformer models were fine-tuned using a manually labelled dataset of 500 tweets: one for detecting offensive language, and the other for abusive language. The dataset was developed to reflect the linguistic style of Dutch football-related discourse, capturing elements such as illness-based profanity, rivalry-specific slurs, and sarcastic phrasing. The two models were trained using a weighted cross-entropy loss to address class imbalance, ensuring that minority toxic examples had stronger influence during optimization.

Classification was restricted to replies and second-level replies, not club-authored timeline tweets, to focus on fan discourse and interaction. Each tweet identified as Dutch was passed through both classifiers. A message was labelled as toxic if either model produced a probability ≥ 0.7. This conservative threshold was selected to ensure high precision, based on the right-skewed distribution of predicted scores, where most non-toxic tweets scored near 0.1 and toxic tweets typically exceeded 0.8.

Table 4.1 presents the performance metrics of both models on a validation subset, including accuracy, precision, recall, and class-specific F1-scores.

***Table 4.1 Performance of Dutch toxicity classifiers on validation set***

|  |  |  |
| --- | --- | --- |
| **Metric** | **Offensive model** | **Abusive model** |
| Accurcay | 0.81 | 0.79 |
| Precision | 0.70 | 0.70 |
| Recall | 0.63 | 0.53 |
| F1-score | 0.67 | 0.60 |
| Toxic F1 | 0.67 | 0.60 |
| Non-toxic F1 | 0.87 | 0.86 |
| True Positives | 19 | 16 |
| False Positives | 11 | 14 |

Across the full dataset, 8,447 tweets were classified as Dutch. Of these, 2,433 tweets (28.8%) were labelled as toxic. Table 4.2 summarizes the descriptive statistics of the predicted probabilities for both models.

***Table 4.2 Dutch classifier probability summary***

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Mean probability | Std. Dev. | Min | Max |
| Offensive (NL) | 0.242 | 0.380 | 0.0006 | 0.9997 |
| Abusive (NL) | 0.234 | 0.362 | 0.0009 | 0.9987 |

### 4.2.2 English classifiers (pre-trained)

For English-language tweets (N = 2,097), the tool integrates four publicly available transformer models from the cardiffnlp/twitter-roberta-base family, each trained on a specific type of toxicity: hate speech, offensive language, abusive language, and threats. These models were applied in their original, pre-trained form due to the absence of a suitable English-language football-specific dataset for fine-tuning.

Only the hate speech model was used to generate binary toxicity labels. A threshold of 0.7 was again applied, and any tweet scoring above this threshold was labelled as toxic. This cautious approach was motivated by the generally low confidence scores observed in the other three classifiers, which were insufficiently distinct from background noise to justify binary labelling. In total, 1,595 tweets (76.1%) were flagged as toxic based on the hate speech classifier, while 502 were labelled as non-toxic. Summary statistics of probability distributions for all four models are presented in Table 4.3.

***Table 4.3 English classifier probability summary***

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Mean probability | Std. Dev. | Min | Max |
| Hate speech (EN) | 0.712 | 0.117 | 0.0217 | 0.857 |
| Offensive (EN) | 0.213 | 0.220 | 0.020 | 0.948 |
| Abusive (EN) | 0.036 | 0.030 | 0.0002 | 0.109 |
| Threat (EN) | 0.489 | 0.116 | 0.140 | 0.842 |

### 4.2.3 Language coverage and exclusions

Before classification, all tweets were passed through an automated language identification step. Out of 16,701 replies, 8,447 were classified as Dutch, 2,097 as English, and 6,157 as either unknown, ambiguous, or in unsupported languages. These excluded tweets included messages that were too short, composed entirely of emojis, or written in informal slang that could not be reliably mapped to a specific language.

This filtering step helps ensure that toxicity classification is only performed where language-specific models are applicable. Prior research has shown that language identification on social media is error-prone in short or non-standard texts (Jauhiainen et al., 2019), and that applying NLP models trained on one language to another can lead to substantial misclassification (Ghosh et al., 2021). To avoid this, the tool excludes unsupported messages altogether, trading recall for greater precision and interpretability.

As a result, approximately 63% of the full dataset was retained for classification: 8,447 Dutch tweets and 2,097 English tweets. The remaining 6,157 were excluded. This conservative filtering strategy prioritizes methodological soundness and aligns with best practices in multilingual NLP, which recommend using language-specific models where possible rather than applying generic classifiers across heterogeneous inputs.

By combining fine-tuned Dutch classifiers, thresholded English models, and a conservative language filtering approach, the tool produces toxicity labels that are appropriate for multilingual football discourse. These outputs form the basis for further network-level analysis in the subsequent module.

## 4.3 SNA module & community detection

This section outlines the social network analysis (SNA) capabilities of the tool, which were developed to map and interpret toxicity within online football discourse. While previous sections addressed content-level detection using NLP, this module shifts the perspective to interaction-level dynamics. By modelling reply behaviour as directed networks, the tool supports community-level interpretation of how toxic communication emerges and spreads within and across Eredivisie fanbases.

### 4.3.1 Network construction

To analyse the structure and circulation of toxicity in Dutch football discourse on X, the tool constructs directed reply networks based on user-to-user interactions. These networks form the backbone of the Social Network Analysis (SNA) module, transforming raw conversations into structured graphs suitable for computational analysis, filtering, and visualization. This approach allows users of the tool to trace how toxic content unfolds across temporally bounded events, revealing structural patterns that may otherwise remain hidden in unstructured text. The use of directed graphs to model reply behaviour is well-established in SNA research, particularly within studies of online discourse and sports-related digital publics (Himelboim et al., 2017; Vermeer & Araujo, 2019).

Each network is constructed from data collected through a custom scraper that targets posts by Eredivisie football clubs. Specifically, the tool collects replies and second-level replies to timeline tweets posted by official club accounts. As a result, every conversation in the network originates from a verified club post. This ensures that each network is grounded in a real-world match-related stimulus and mirrors practices in prior studies that position official accounts as central broadcasting hubs (Bruns et al., 2014).

Nodes in the network represent unique user accounts, while directed edges signify the act of replying. A directed edge from user A to user B denotes that A replied to a tweet authored by B. This preserves the directional logic of interaction and makes it possible to trace chains of influence, targeted responses, or cascades of toxicity. Both direct and second-level replies are included, thereby capturing not only responses to clubs but also interactions among fans within the same discussion thread. This depth of interaction is important for identifying emergent toxic clusters and potential echo chambers (Cinelli et al., 2021).

The tool builds two sets of reply networks: one at the match level and one at the club level. Match-level networks are constructed for each of the 79 Eredivisie fixtures in the dataset. These graphs are temporally bounded and include only those interactions posted around a specific matchday. They allow users to isolate how online toxicity and engagement evolve in response to discrete football events, such as controversial goals or derby matches. In contrast, club-level networks aggregate all interactions directed at a club’s official account throughout the full scraping period. These broader graphs offer insight into persistent user behaviour, recurring toxic clusters, and overall engagement structures across an entire season.

Graph construction begins by extracting the author and parent author fields from each reply. Parent information is inferred using tweet metadata and reply chain structure. The tool then creates a directed edge from the author to the parent author, representing the act of replying. Usernames are normalized by converting all handles to lowercase and removing extraneous symbols to ensure consistency and avoid duplication due to minor formatting differences. Timeline tweets authored by official club accounts are retained as isolated nodes to preserve their structural role, even if they are not directly replied to.

Once all edges are defined, the tool filters out self-loops and structurally invalid relationships. The resulting graphs are stored in memory as DiGraph objects using the NetworkX library. This format supports both analysis and visualization and provides compatibility with downstream modules in the tool pipeline. The inclusion of second-level replies ensures that the networks not only capture direct responses to clubs but also reflect how toxicity may propagate through user-to-user chains. This layered structure aligns with findings from prior research on social media echo chambers and network polarization, where toxicity is known to cluster and spread within tightly connected user communities (Cinelli et al., 2021).

Ultimately, these reply graphs serve as foundational data structures within the tool. They enable functions such as user-level labelling, community detection, and metric computation, and can be explored visually or exported for further use depending on the analytical needs of the user.

### 4.3.2 Node labelling and toxicity aggregation

After constructing the interaction networks, the tool enriches each user node with metadata derived from the NLP module and match-level metadata. This step transforms the networks from purely structural graphs into annotated social systems, in which users can be filtered, grouped, and visualized according to their toxic behaviour, role, and language use.

The most important addition to each user node is a toxicity label, derived from the tweet-level classifications generated earlier. To assign this label, the tool calculates the proportion of each user's replies that were classified as toxic. If 25% or more of a user’s replies exceed the toxicity threshold (set at 0.7), the user is labelled as toxic. This threshold reflects a balanced approach: it avoids falsely labelling users who post an isolated harmful comment, while still capturing patterns of repeated negative behaviour. Users with fewer than four classified replies, or whose messages were excluded from NLP classification (e.g., due to unsupported language), are assigned the label unknown.

In addition to toxicity status, each user is also classified by account type. In this tool, all conversations originate from timeline tweets posted by official Eredivisie club accounts. These accounts form the starting point of every network, as the data collection process is based on scraping replies and second-level replies to these club-authored posts. As such, club accounts always appear in the network as initial nodes or conversation anchors to which fans respond directly or indirectly. These official accounts are automatically identified and flagged based on a predefined list of verified handles extracted from the match metadata. All other users in the graph are considered general participants and are labelled as fans. Because the dataset is sourced exclusively from club-initiated content, the role of club accounts is structurally consistent across all match and club networks.

Each user is also associated with a language category, which reflects the dominant language in their classified replies. This label captures whether a user primarily posted in Dutch, English, or another language. It supports optional filtering or comparative analysis across linguistic groups and is especially useful in understanding how toxicity differs between multilingual communities.

By attaching these attributes to every user node, the tool enables flexible filtering, sorting, and visualization of toxic behaviour within online football communities. For instance, toxic users can be visualized in red, non-toxic users in blue, and unknown participants in grey. Official club accounts are highlighted in yellow to distinguish their structural role. One example of a labelled network is presented in Figure 4.3, which displays the annotated interaction graph for Match M014. This visualization clearly illustrates how toxic, non-toxic, and unlabelled users are distributed around the initiating club accounts, and how toxic interactions often cluster within the broader discussion.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 1 Toxicity interaction network for Match M014.

This node-level annotation process lays the groundwork for community detection and metric analysis in subsequent stages of the tool. It ensures that the networks generated are not only structurally valid but also semantically meaningful, enabling richer, more context-aware interaction analysis for users of the tool.

### 4.3.3 Community detection

To explore how toxicity concentrates within football-related discussions, the tool incorporates a community detection module that partitions each interaction network into tightly connected user clusters. In this context, a community is defined as a group of users who are more densely connected to one another than to users outside their group. Mapping these communities reveals the underlying social structure of football discourse on X, where toxicity may not be evenly distributed but rather clustered in echo chambers or isolated toxic subgroups. This aligns with previous SNA studies of sports communities, which have shown that user influence and sentiment often concentrate within distinct fan clusters (van der Nagel & Scholten, 2019).

The tool applies the greedy modularity community detection algorithm (Clauset, Newman, & Moore, 2004), a widely used approach for detecting communities in large graphs. This algorithm partitions each network into a set of non-overlapping communities by maximizing modularity—a metric that quantifies the density of links within groups compared to links between them. The method is deterministic, computationally efficient, and well-suited for large directed networks like those constructed from Eredivisie match replies.

Community detection is performed separately on each match-level network, providing a snapshot of how interactions group together during specific events. In these networks, each user is assigned a community label, allowing for aggregated statistics such as the number of toxic users per community or the overall toxicity percentage within each cluster. An example output is shown in table 4.4, which lists communities identified in Match M077, along with their size and share of toxic users. Although the community detection algorithm identified 14 clusters for this match, only 12 communities are included in the table. Communities with extremely low user counts (e.g., one or two users) were excluded from reporting, as they lacked sufficient data to compute meaningful toxicity ratios. Several of the reported clusters contain high concentrations of toxic users, for instance, Community 10 has 7 users, 6 of whom are labelled toxic (85.7%), highlighting how toxicity can emerge in isolated subgroups rather than spread uniformly throughout the network.

***Table 4.4: Community statistics for Match M077***

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Community ID | No. Of users | Toxic users | Non-toxic users | Unknown users | % Toxic |
| 0 | 41 | 17 | 15 | 9 | 41.5% |
| 1 | 29 | 15 | 10 | 4 | 51.7% |
| 2 | 29 | 7 | 19 | 3 | 24.1% |
| 3 | 28 | 17 | 9 | 2 | 60.7% |
| 4 | 19 | 13 | 5 | 1 | 68.4% |
| 5 | 19 | 6 | 6 | 7 | 31.6% |
| 6 | 16 | 6 | 7 | 3 | 37.5% |
| 7 | 12 | 4 | 1 | 7 | 33.3% |
| 8 | 9 | 3 | 3 | 3 | 33.3% |
| 9 | 9 | 4 | 3 | 2 | 44.4% |
| 10 | 7 | 6 | 0 | 1 | 85.7% |
| 11 | 5 | 4 | 0 | 1 | 80% |

This level of analysis enhances the tool’s ability to detect coordinated or reinforcing toxic behaviour, especially in scenarios where harmful users primarily engage with each other. These patterns often go undetected in purely content-based classification systems, which treat users in isolation.

While community detection is especially useful for match-level snapshots, the same approach is also applied to club-level networks. Because these graphs aggregate interactions from all matches involving a club, they can reveal more persistent toxic clusters that form around a club’s identity over time. This is particularly relevant for identifying accounts or subgroups that regularly contribute to toxic discourse throughout the season.

Each community’s composition—including its size, user types, toxicity ratio, and language breakdown—is stored for further filtering, visualization, or export. These outputs allow users of the tool to interpret not just where toxicity occurs, but how it is socially organized and structurally sustained. By embedding community detection in the SNA workflow, the tool provides a deeper, group-level understanding of harmful behaviour in Dutch football communities online.

### 4.3.4 Network metrics and comparison

Beyond identifying user communities, the tool also computes structural network metrics that provide quantitative insight into how interaction patterns differ across matches and clubs. These metrics allow users to compare the scale, density, and cohesion of conversations, and to observe how toxicity unfolds within different structural configurations. The inclusion of these network metrics is grounded in their established value within social network analysis for capturing influence, cohesion, and information flow (Vivek et al., 2025; Ertem et al., 2016).

For each match-level and club-level network, the tool calculates the following:

* **Number of nodes:** the total number of unique users in the network
* **Number of edges:** the number of reply interactions recorded
* **Average degree centrality:** the average proportion of users each node is directly connected to. Degree centrality is widely used to identify active or prominent individuals in online social networks, as it reflects the number of direct connections each user maintains (Vivek et al., 2025).
* **Average betweenness centrality:** the average extent to which users lie on the shortest communication paths between others. This metric highlights users who serve as bridges between clusters and can significantly influence the flow of information or toxicity (Vivek et al., 2025).
* **Average clustering coefficient:** the degree to which users form tightly connected groups. This measure has been applied in practice to detect cohesive subgroups within social networks, especially in the context of community formation and echo chambers (Ertem et al., 2016).

These metrics are automatically computed during graph construction and are stored alongside each network for export or comparison. They provide a structural overview of how active and cohesive the conversation is around each match or club, and whether toxic behaviour emerges in fragmented or highly interconnected contexts.

Table 4.5 shows the structural metrics for five selected match networks. The largest of these, Match M058, features 965 users and 1,355 reply edges. Although it has high engagement, the average degree centrality remains relatively low, suggesting that users are often replying to isolated accounts rather than participating in broad conversations. Conversely, matches like M010 and M035 show higher clustering coefficients, indicating more tightly connected discussion groups.

***Table 4.5: Match-level structural metrics***

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Match ID | No. of users | No. of edges | Avg. degree centrality | Avg. betweenness | Avg. clustering |
| M058 | 965 | 1355 | 0.0029 | 0.00011 | 0.0253 |
| M075 | 592 | 808 | 0.0046 | 0.00005 | 0.0225 |
| M029 | 534 | 736 | 0.0052 | 0.00052 | 0.0071 |
| M010 | 478 | 611 | 0.0054 | 0.00003 | 0.0338 |
| M035 | 452 | 580 | 0.0057 | 0.00003 | 0.0220 |

On the club level, differences in structure are even more pronounced. As shown in table 4.6, clubs with large national fanbases, such as Ajax and Feyenoord, generate the most interactions. AFC Ajax's network includes over 2,200 users and more than 4,200 reply edges. However, size alone does not determine structure: clubs like FC Twente and AZ Alkmaar display significantly higher clustering coefficients, implying that their online discussions are more localized and internally cohesive. These variations suggest that toxicity may manifest differently depending on whether interactions are broad and loosely connected or occur within close-knit groups.

***Table 4.6: Club-level structural metrics***

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Club | No. of users | No. of edges | Avg. degree centrality | Avg. betweenness | Avg. clustering |
| AFC Ajax | 2269 | 4234 | 0.0016 | 0.00015 | 0.0596 |
| Feyenoord | 1508 | 2764 | 0.0024 | 0.00034 | 0.0485 |
| PSV | 1348 | 2132 | 0.0023 | 0.00031 | 0.0414 |
| FC Twente | 747 | 1178 | 0.0042 | 0.00043 | 0.0747 |
| AZ Alkmaar | 343 | 506 | 0.0086 | 0.00042 | 0.0472 |

Together, these metrics provide valuable context for interpreting both user behaviour and network structure. High clustering may coincide with echo chambers of toxic reinforcement, while high betweenness centrality may reveal key users through whom toxicity spreads. The ability to compare networks across time (match-level) and fanbases (club-level) gives users of the tool the flexibility to explore toxicity as both a structural and situational phenomenon.

### 4.3.5 Toxicity distribution across clubs

While match-level analysis reveals how toxicity unfolds during individual football events, broader behavioural patterns emerge when data is aggregated at the club level. This functionality within the SNA module enables the tool to compare toxic engagement across Eredivisie fanbases by calculating and visualizing both the total number of toxic users and the proportion of toxicity within each club’s online interaction network.

The tool computes club-level toxicity statistics by aggregating all classified interactions across the matches in which a club’s official X account was involved. For each club, it calculates the total number of users with sufficient classification data, the number of those users labelled as toxic, and the resulting toxicity ratio. To ensure reliability, only users with at least four classified replies are included in this aggregation step, thereby excluding occasional users whose behaviour may not be meaningfully evaluated.

Two complementary visualizations are provided. The first, shown in **Figure 4.5**, ranks clubs by the absolute number of toxic users. As expected, clubs with larger and more active online fanbases dominate this ranking. AFC Ajax, Feyenoord, and PSV each contribute several hundred toxic users, reflecting their broader reach and interaction volume. These findings align with earlier SNA work on Dutch football networks, which shows that interaction density tends to concentrate around large club accounts (Vermeer & Araujo, 2019).

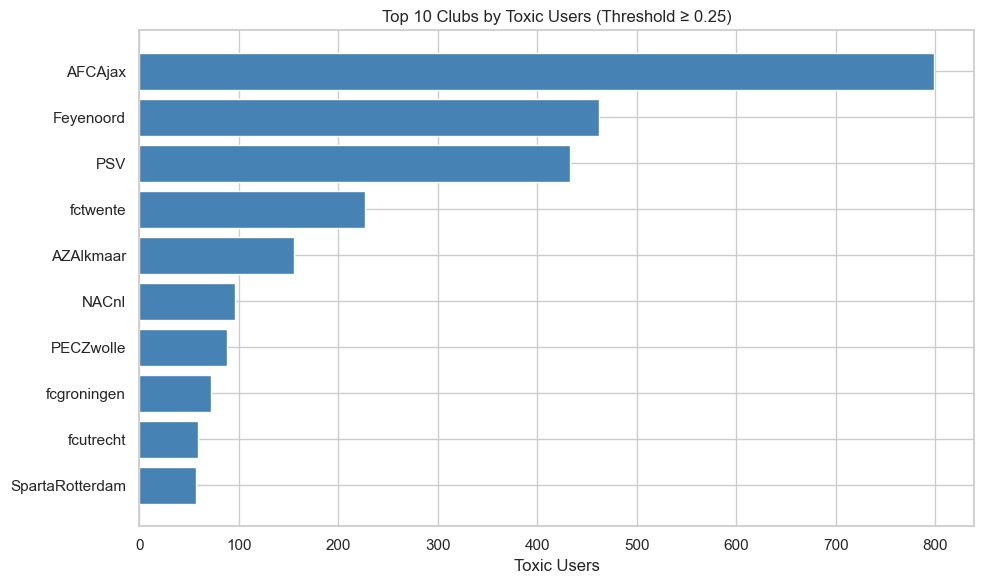


Figure 2 Top 10 Clubs by Number of Toxic Users

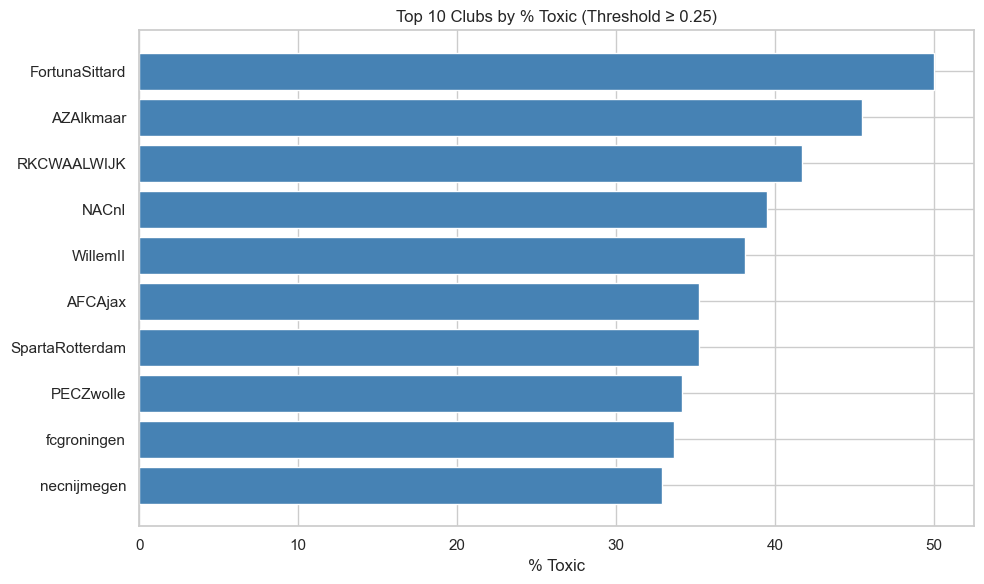
However, when toxicity is normalized as a percentage of the total classified user base, a different picture emerges. As shown in **Figure 4.6**, clubs such as Fortuna Sittard, AZ Alkmaar, and RKC Waalwijk show the highest relative proportions of toxic users, with some exceeding 45%. These clubs, despite having smaller overall communities, appear to foster more concentrated toxicity. This supports prior findings in online behavioural research suggesting that smaller, more insular groups may intensify harmful discourse through repeated exposure and reinforcement (Cinelli et al., 2021; Chandrasekharan et al., 2017)

Figure 3 Top 10 Clubs by Percentage of Toxic Users

Together, these two views—absolute volume and relative concentration—offer a more nuanced understanding of how toxicity is distributed across the league. High total numbers of toxic users may indicate the scale and virality of discourse, while high percentages point to structural conditions that enable toxicity to thrive within smaller but tighter-knit communities. Both are important for clubs, moderators, or researchers seeking to evaluate reputational risk, community dynamics, or the effectiveness of intervention strategies.

### 4.3.6 Timeline visualizations

To support temporal exploration of toxicity dynamics, the tool includes functionality for generating interaction timeline plots. These visualizations allow users to examine how toxic and non-toxic replies fluctuate over time, particularly in the hours surrounding Eredivisie matches. This feature is designed to help users identify spikes in interaction volume and assess when toxicity tends to emerge during the course of an event.

Each timeline plot displays the number of classified interactions, grouped into toxic and non-toxic categories, over a continuous time axis. The timeline is annotated with vertical markers to indicate the kick-off and full-time moments of a match, providing a temporal anchor for interpreting changes in user activity. These annotations make it easier to observe whether interaction peaks align with key sporting moments, such as the start of a match, goals, or controversial decisions.

An example is shown in **Figure 4.7**, which visualizes the volume of toxic and non-toxic replies surrounding the match between Ajax and AZ (Match M071). In this case, interaction density builds rapidly just before kick-off, peaks during the early moments of the match, and declines shortly after full-time. Both toxic and non-toxic messages are represented in stacked colour bands to show how sentiment shifts across time.

A graph with numbers and lines

AI-generated content may be incorrect.

Figure 4 Toxic vs Non-Toxic Interaction Timeline – Ajax vs AZ (Match M071)

These plots are not intended to support detailed content interpretation within the tool itself, but rather to offer an accessible, high-level view of interaction rhythms. By comparing timelines across matches, users can detect consistent patterns, outliers, or unusual bursts of activity that may warrant deeper examination. This functionality is particularly valuable for stakeholders such as clubs, moderators, or media teams who need to identify high-engagement moments or potential risk periods without reading thousands of individual messages.

## 4.4 Summary

This chapter has demonstrated how the tool leverages Social Network Analysis to model and visualize the structure of toxic interactions in online football communities. Building on the output of the NLP module, the tool constructs directed reply networks at both the match and club level, labels users based on behavioural attributes, detects densely connected communities, and generates structural metrics and timeline-based visualizations.

The design prioritizes modularity, scalability, and interpretability, allowing users to explore toxicity not just at the individual message level, but also within broader social and temporal contexts. By combining structural insights with linguistic classification, the tool offers a holistic view of online interaction dynamics, applicable across matches, clubs, and time windows.

These capabilities enable downstream use cases such as flagging high-risk communities, monitoring temporal flashpoints, or exporting structured data for further analysis. The following chapter reflects on the broader contributions of the tool, its current limitations, and potential pathways for future development and implementation.

# Conclusions

This final chapter provides a reflection on the tool developed in this thesis and its relevance for supporting the structured analysis of toxicity in online football communities. While Chapter 4 detailed the technical components and outputs of the tool, the current chapter discusses its broader contribution, key limitations, and potential applications. It also outlines opportunities for future work that could enhance the tool’s functionality, usability, or scope. Rather than offering a conclusion in the form of theoretical interpretation, this chapter remains consistent with the thesis aim: to design, implement, and document a modular tool for toxic content analysis using natural language processing and social network analysis techniques.

### 5.1 Tool contribution & summary

The primary contribution of this thesis lies in the development of a modular software tool that enables the detection, classification, and structural analysis of toxic interactions within Dutch football-related discussions on X. The tool integrates natural language processing (NLP) with social network analysis (SNA) to support users in mapping how toxicity is distributed, clustered, and sustained within temporally and socially bounded online conversations. Rather than aiming to investigate toxicity as a social phenomenon, the focus of this thesis has been on building an operational and extensible tool that others—such as researchers, moderators, or football organizations—can use to carry out targeted analyses.

The tool is structured around five core components: a custom scraping module, a Dutch-English toxicity classifier, a reply-based network builder, a community detection pipeline, and a visual interface for exploration. Each component is designed to operate independently while contributing to a larger analytical workflow. For instance, the NLP module assigns toxicity labels to replies in two languages, the network module transforms reply chains into directed graphs, and the community detection module groups users into cohesive clusters to allow for structural analysis of toxic engagement. Output from each module can be visualized, filtered, or exported for external use.

This modular design ensures flexibility in both academic and applied settings. Users may choose to focus on one dimension—such as the evolution of toxicity over time, or the structure of toxic communities—or use the full workflow to examine how linguistic toxicity interacts with network position. The tool operates offline and uses scraped data, which allows for reproducibility and event-based analysis without requiring continuous data access from the X platform. While the underlying methods are computational, the interface and output formats are structured to be interpretable by non-technical users.

In sum, the tool contributes a replicable and well-documented framework for analysing toxic behaviour in Dutch football discourse. It combines recent advances in NLP and SNA within a single workflow and demonstrates how these methods can be brought together in a way that supports real-world use cases. The next sections will outline the current limitations of the tool and discuss avenues for further development.

### 5.2 Limitations

While the tool developed in this thesis provides a robust and modular framework for analysing toxic interactions in online football discourse, several limitations remain that influence both its current performance and its broader applicability. These limitations relate to data coverage, language support, classifier design, network construction, and domain specificity.

A key limitation lies in the scope of data collection. The scraping component of the tool captures only replies and second-level replies to timeline tweets posted by Eredivisie club accounts. Although this design supports clean network boundaries and ensures that all conversations originate from a verifiable football-related context, it omits other significant forms of social interaction. Quote tweets, retweets, and standalone mentions are not collected. This restricts the tool’s ability to capture indirect or viral forms of toxicity, which may spread outside of reply chains or originate in broader football discourse. Toxicity expressed through resharing or amplification, for example, is structurally invisible in the current implementation, even though such behaviours are highly relevant in reputational harm and platform dynamics.

The tool also has language coverage constraints. At present, toxicity classification is only applied to tweets identified as either Dutch or English. Messages in other languages, or tweets composed largely of emojis, code-switching, or slang not recognized by the language detector, are excluded from analysis. In total, approximately 6,500 tweets could not be classified due to language mismatch or detection failure. While this filtering preserves model accuracy, it introduces a systematic blind spot in toxicity measurement, particularly in multilingual fanbases or during emotionally expressive interactions that rely on visuals and symbols rather than text. This limitation reflects a trade-off between linguistic precision and inclusivity.

The classification component also introduces important limitations. For Dutch-language tweets, a manually labelled dataset of 500 examples was constructed and used to fine-tune the toxicity classifier. This manual effort adds contextual sensitivity to the model and ensures that football-specific expressions of offense or aggression are represented in training. However, the size of the dataset remains relatively small, limiting the diversity of linguistic patterns the model can learn. More critically, the English-language classifier was not fine-tuned at all. Due to time constraints, it relies solely on a pre-trained transformer without any domain-specific adaptation. This means the English component lacks calibration to football discourse, Dutch cultural context, or platform-specific slang—factors known to influence how toxicity manifests in sports communities. Expanding both language models through additional annotation, particularly across diverse matches, user types, and linguistic forms, would significantly enhance classification quality and reduce false positives and false negatives.

The binary thresholding strategy used for labelling messages as toxic or non-toxic also introduces constraints. A reply is marked as toxic if it exceeds a fixed probability threshold of 0.70. While this approach increases precision and reduces over-flagging of ambiguous content, it may exclude subtler or more context-dependent forms of toxicity. Sarcasm, passive-aggression, or coded language may not reach the threshold despite being harmful in practice. Additionally, the threshold is fixed across both languages and contexts, without dynamic calibration. Future work could explore the benefits of soft labelling, multi-class schemes, or context-aware thresholding based on match type, user behaviour, or message length.

There are also limitations related to network construction and representation. The tool models user-to-user interaction through directed reply networks, where edges represent the act of replying. While this preserves conversational structure and allows for reply-depth analysis, it excludes other forms of social ties, such as follow relationships, shared retweet behaviour, or mutual interaction over time. This creates a partial view of the online community. Influential users who post toxic content without replying directly to others, for example, may remain peripheral in the reply graph despite having substantial reach. Similarly, users who quote toxic messages or spread them through secondary channels are not captured, limiting the model’s ability to represent toxic diffusion or influence propagation.

Finally, the tool is tailored to a domain-specific context. It is designed around the structure of Dutch professional football, with scraping logic, club metadata, and user labelling functions that assume the presence of official Eredivisie accounts. While this allows for high accuracy within the domain, it constrains the generalizability of the tool. Applying it to other sports leagues, cultural contexts, or non-sport domains would require adaptation of scraping targets, account whitelists, language settings, and possibly model retraining. The tool is extensible in principle, but not fully plug-and-play across domains.

These limitations should not be seen as weaknesses but rather as indicators of the project’s scope and realism. Within its current parameters, the tool performs reliably and offers significant value as a research and monitoring interface. However, acknowledging its boundaries is essential for responsible use and forms the basis for identifying concrete improvements, which are outlined in Section 5.4.

# References

Abbasi, A., Javed, A. R., Iqbal, F., Kryvinska, N., & Jalil, Z. (2022). Deep learning for religious and continent-based toxic content detection and classification. *Scientific Reports*, *12*(1), 17478. <https://doi.org/10.1038/s41598-022-22523-3>

Ahmed, Z., Vidgen, B., & Hale, S. A. (2022). Tackling racial bias in automated online hate detection: Towards fair and accurate detection of hateful users with geometric deep learning. *EPJ Data Science*, *11*(1), 1–22. <https://doi.org/10.1140/epjds/s13688-022-00319-9>

*NOS* (2025, April 4). *Ajax vraagt met speciale shirts aandacht voor strijd tegen haat op social media*. NOS. <https://nos.nl/artikel/2562301-ajax-vraagt-met-speciale-shirts-aandacht-voor-strijd-tegen-haat-op-social-media>

Akar, E. (2025). Exploring the impact of social network structures on toxicity in online mental health communities. *Computers in Human Behavior*, *146*, 108542. <https://doi.org/10.1016/j.chb.2024.108542>

Almerekhi, H., Kwak, H., Jansen, B. J., & Salminen, J. (2019). Detecting Toxicity Triggers in Online Discussions. *Proceedings of the 30th ACM Conference on Hypertext and Social Media*, 291–292. <https://doi.org/10.1145/3342220.3344933>

Harteveld, E. (2022, February 22). *Polarisatie in Nederland: Hoe verdeeld zijn we?* Universiteit van Amsterdam. <https://www.uva.nl/shared-content/faculteiten/nl/faculteit-der-maatschappij-en-gedragswetenschappen/nieuws/2022/02/polarisatie-in-nederland-hoe-verdeeld-zijn-we.html>

Behera, R. K., Bala, P. K., Rana, N. P., & Kayal, G. (2022). Self-promotion and online shaming during COVID-19: A toxic combination. *International Journal of Information Management Data Insights*, *2*(2), 100117. <https://doi.org/10.1016/j.jjimei.2022.100117>

Bryman, A. (2016). *Social research methods* (Fifth Edition). Oxford University Press.

Buongiovanni, C., Candusso, R., Cerretini, G., Febbe, D., Morini, V., & Rossetti, G. (2023). Will You Take the Knee? Italian Twitter Echo Chambers’ Genesis During EURO 2020. In H. Cherifi, R. N. Mantegna, L. M. Rocha, C. Cherifi, & S. Miccichè (Eds.), *Complex Networks and Their Applications XI* (Vol. 1077, pp. 29–40). Springer International Publishing. <https://doi.org/10.1007/978-3-031-21127-0_3>

Burtenshaw, B., & Kestemont, M. (2021). A Dutch Dataset for Cross-lingual Multilabel Toxicity Detection. In R. Rapp, S. Sharoff, & P. Zweigenbaum (Eds.), *Proceedings of the 14th Workshop on Building and Using Comparable Corpora (BUCC 2021)* (pp. 75–79). INCOMA Ltd. <https://aclanthology.org/2021.bucc-1.10/>

Cable, J., Kilvington ,Daniel, & and Mottershead, G. (2022). ‘Racist behaviour is interfering with the game’: Exploring football fans’ online responses to accusations of racism in football. *Soccer & Society*, *23*(8), 880–893. <https://doi.org/10.1080/14660970.2022.2109804>

Caselli, T., Schelhaas, A., Weultjes, M., Leistra, F., van der Veen, H., Timmerman, G., & Nissim, M. (2021). DALC: The Dutch Abusive Language Corpus. In A. Mostafazadeh Davani, D. Kiela, M. Lambert, B. Vidgen, V. Prabhakaran, & Z. Waseem (Eds.), *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)* (pp. 54–66). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.woah-1.6>

Chakrabarty, N. (2020). A Machine Learning Approach to Comment Toxicity Classification. In A. K. Das, J. Nayak, B. Naik, S. K. Pati, & D. Pelusi (Eds.), *Computational Intelligence in Pattern Recognition* (pp. 183–193). Springer. <https://doi.org/10.1007/978-981-13-9042-5_16>

Cinelli, M., Morales, G. D. F., Galeazzi, A., Quattrociocchi, W., & Starnini, M. (2020). *Echo Chambers on Social Media: A comparative analysis* (No. arXiv:2004.09603). arXiv. <https://doi.org/10.48550/arXiv.2004.09603>

Davani, A. M., Atari, M., Kennedy, B., & Dehghani, M. (2023). Hate Speech Classifiers Learn Normative Social Stereotypes. *Transactions of the Association for Computational Linguistics*, *11*, 300–319. <https://doi.org/10.1162/tacl_a_00550>

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated Hate Speech Detection and the Problem of Offensive Language. *Proceedings of the International AAAI Conference on Web and Social Media*, *11*(1), Article 1. <https://doi.org/10.1609/icwsm.v11i1.14955>

Engelhardt, P. E. (2013). Are listeners ‘fully’ Gricean?: A comment on Davies and Katsos (2013). *Journal of Pragmatics*, *49*(1), 107–113. <https://doi.org/10.1016/j.pragma.2013.01.001>

Fan, L., Li, L., & Hemphill, L. (2024). Toxicity on social media during the 2022 Mpox public health emergency: Quantitative study of topical and network dynamics. *JMIR Infodemiology*, *4*, e52997. <https://doi.org/10.2196/52997>

Fenton, A., Ahmed, W., Hardey, M. (Maz), Boardman, R., & Kavanagh, E. (2024). Women’s football subculture of misogyny: The escalation to online gender-based violence. *European Sport Management Quarterly*. <https://www.tandfonline.com/doi/abs/10.1080/16184742.2023.2270566>

Gevers, I., Markov, I., & Daelemans, W. (2022b). Linguistic Analysis of Toxic Language on Social Media: 32nd Meeting of Computational Linguistics in the Netherlands, CLIN 2022. *Computational Linguistics in the Netherlands*, *12*, 33–48.

Ghosh, S., Baker, D., Jurgens, D., & Prabhakaran, V. (2021). *Detecting Cross-Geographic Biases in Toxicity Modeling on Social Media* (No. arXiv:2104.06999). arXiv. <https://doi.org/10.48550/arXiv.2104.06999>

Haddow, B., Bawden, R., Barone, A. V. M., Helcl, J., & Birch, A. (2022). Survey of Low-Resource Machine Translation. *Computational Linguistics*, *48*(3), 673–732. <https://doi.org/10.1162/coli_a_00446>

Helal, N. A., Hassan, A., Badr, N. L., & Afify, Y. M. (2024). A contextual-based approach for sarcasm detection. *Scientific Reports*, *14*(1), 15415. <https://doi.org/10.1038/s41598-024-65217-8>

Jahan, M. S., & Oussalah, M. (2023). A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, *546*, 126232. <https://doi.org/10.1016/j.neucom.2023.126232>

Johnson, R. G., Al-khateeb, S., Forbes, A., & Cupido, K. (2024). Targeted Social Media Harassment: A Comparative Analysis of Toxicity Directed at Men and Women Sports Reporters. *Communication & Sport*, *12*(3), 443–465. <https://doi.org/10.1177/21674795231213330>

Kearns, C., Sinclair, G., Black, J., Doidge, M., Fletcher, T., Kilvington, D., & others. (2023). A scoping review of research on online hate and sport. *International Review for the Sociology of Sport*. <https://doi.org/10.1177/21674795221132728>

Kiddle, R., Törnberg, P., & Trilling, D. (2024). Network toxicity analysis: An information-theoretic approach to studying the social dynamics of online toxicity. *Journal of Computational Social Science*, *7*(1), 305–330. <https://doi.org/10.1007/s42001-023-00239-2>

Kilvington, D., & Price, J. (2021). The ‘beautiful game’ in a world of hate: Sports journalism, football and social media abuse. In *Insights on Reporting Sports in the Digital Age*. Routledge.

Kok, D. de, & Pütz, T. (2020). Self-distillation for German and Dutch dependency parsing. *Computational Linguistics in the Netherlands Journal*, *10*, 91–107.

Lapidot-Lefler, N., & Barak, A. (2012). Effects of anonymity, invisibility, and lack of eye-contact on toxic online disinhibition. *Computers in Human Behavior*, *28*(2), 434–443. <https://doi.org/10.1016/j.chb.2011.10.014>

Mall, R., Nagpal, M., Salminen, J., Almerekhi, H., Jung, S.-G., & Jansen, B. J. (2020). Four Types of Toxic People: Characterizing Online Users’ Toxicity over Time. *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society*, 1–11. <https://doi.org/10.1145/3419249.3420142>

Mathew, B., Saha, P., Tharad, H., Rajgaria, S., Singhania, P., Maity, S. K., Goyal, P., & Mukherje, A. (2019). *Thou shalt not hate: Countering Online Hate Speech* (No. arXiv:1808.04409). arXiv. <https://doi.org/10.48550/arXiv.1808.04409>

Ministerie van Volksgezondheid (2022, December 29). *Driekwart van de Nederlanders denkt dat polarisatie toeneemt—Nieuwsbericht—Sociaal en Cultureel Planbureau*. Ministerie van Volksgezondheid, Welzijn en Sport. <https://www.scp.nl/actueel/nieuws/2022/12/29/ergernis-over-harde-toon-en-extreme-uitingen-in-politieke-en-publieke-debat>

Mohanty, F., Rup, S., Dash, B., Majhi, B., & Swamy, M. N. S. (2020). An improved scheme for digital mammogram classification using weighted chaotic salp swarm algorithm-based kernel extreme learning machine. *Applied Soft Computing*, *91*, 106266. <https://doi.org/10.1016/j.asoc.2020.106266>

Mojtahedi, D., Webb, T., Leadley, C. B., & Jones, M. (2024). Match officials and abuse: A systematic review. *European Sport Management Quarterly*, *24*(1), 199–221. <https://doi.org/10.1080/16184742.2022.2099441>

Mullah, N. S., Zainon, W. M. W., & Wahab, M. N. (2024). Transfer learning approach for identifying negative sentiment in tweets directed to football players. *Engineering Applications of Artificial Intelligence*, *127*, 108377. <https://doi.org/10.1016/j.engappai.2024.108377>

Pavlopoulos, J., Sorensen, J., Dixon, L., Thain, N., & Androutsopoulos, I. (2020). Toxicity Detection: Does Context Really Matter? In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 4296–4305). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-main.396>

Sahoo, N., Gupta, H., & Bhattacharyya, P. (2022). Detecting Unintended Social Bias in Toxic Language Datasets. In A. Fokkens & V. Srikumar (Eds.), *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)* (pp. 132–143). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.conll-1.10>

Salminen, J., Sengün, S., Corporan, J., Jung, S., & Jansen, B. J. (2020). Topic-driven toxicity: Exploring the relationship between online toxicity and news topics. *PLOS ONE*, *15*(2), e0228723. <https://doi.org/10.1371/journal.pone.0228723>

Sap, M., Card, D., Gabriel, S., Choi, Y., & Smith, N. A. (2019). The Risk of Racial Bias in Hate Speech Detection. In A. Korhonen, D. Traum, & L. Màrquez (Eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 1668–1678). Association for Computational Linguistics. <https://doi.org/10.18653/v1/P19-1163>

Saveski, M., Roy, B., & Roy, D. (2021). The Structure of Toxic Conversations on Twitter. *Proceedings of the Web Conference 2021*, 1086–1097. <https://doi.org/10.1145/3442381.3449861>

Seijbel, J., van Sterkenburg, J., & Oonk, G. (2022). Expressing rivalry online: Antisemitic rhetoric among Dutch football supporters on Twitter. *Soccer and Society*. <https://archive.jpr.org.uk/object-2953>

Seijbel, J., van Sterkenburg, J., & Spaaij, R. (2023). Online Football-Related Antisemitism in the Context of the COVID-19 Pandemic: A Multi-Method Analysis of the Dutch Twittersphere. *American Behavioral Scientist*, *67*(11), 1304–1321. <https://doi.org/10.1177/00027642221118286>

Sheth, A., Shalin, V. L., & Kursuncu, U. (2022). Defining and detecting toxicity on social media: Context and knowledge are key. *Neurocomputing*, *489*, 279–293. <https://doi.org/10.1016/j.neucom.2021.11.095>

Suler, J. (2004). The Online Disinhibition Effect. *CyberPsychology & Behavior*, *7*(3), 321–326. <https://doi.org/10.1089/1094931041291295>

van der Vegt, I. (2024). Gender differences in online abuse: The case of Dutch politicians. *Crime Science*, *13*(1), 5. <https://doi.org/10.1186/s40163-024-00203-z>

van Tulder, R. (2019). *Getting all motives right: A holistic approach to research in business and society*. RSM Erasmus University.

Vermeer, S., & Araujo, T. (2020). Keep the ball rolling: Information diffusion within large sports-related networks through social mediators. *Social Media + Society*, *6*(2), 1–13. <https://doi.org/10.1177/2167479519841868>

Vidgen, B., Chung, Y.-L., Johansson, P., Park, H.-R., Williams, A., & Hale, S. A. (2022). Tracking abuse on Twitter against football players in the 2021–22 Premier League season. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.4403913>

Ye, Z., Wu, D., He, X., Ma, Q., Peng, J., Mao, G., Feng, L., & Tong, Y. (2023). Meta-analysis of the relationship between bullying and depressive symptoms in children and adolescents. *BMC Psychiatry*, *23*(1), 215. <https://doi.org/10.1186/s12888-023-04681-4>

# Appendix