NOVA IMS Information Management School

MDSAA

Master Degree Program in

Data Science and Advanced Analytics

Portuguese Digital Ecosystem:

A Cross-Platform Comparison of Toxicity, Sentiment, and Moral Language in Online Discourse

Cláudia Maria Brito de Sampaio Beiral

Master Thesis

presented as partial requirement for obtaining a Master's Degree in Data Science and Advanced Analytics

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

Portuguese Digital Ecosystem:

A Cross-Platform Comparison of Toxicity, Sentiment, and Moral Language in Online Discourse

by

Cláudia Maria Brito de Sampaio Beiral

Master Thesis as partial requirement for obtaining the Master's degree in Data Science and Advanced Analytics, with a specialization in Business Analytics

Supervised by

Professor Doutor Flávio Luís Portas Pinheiro, NOVA IMS

July, 2025

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Cláudia Beiral

Lisbon, July 2025

ACKNOWLEDGMENTS

I would like to express my heartfelt gratitude to my parents and my sister for their unconditional patience, support and encouragement throughout my academic journey. Your belief in me has been a constant source of strength, especially during times when I doubted myself. Thank you for making me feel I can always count on you.

To my dear friends, thank you for keeping me grounded, making me laugh, and reminding me that life exists beyond deadlines and drafts. I feel very grateful for your friendship, presence and support.

Lastly but certainly not least, I thank my supervisor and professor, Flávio Pinheiro, for his guidance, availability and advice given throughout this journey.

ABSTRACT

Social media platforms have fundamentally reshaped the landscape of public discourse, creating dynamic environments where people express opinions, share and consume information, and engage in discussion. However, these platforms are not neutral spaces. Their design, algorithms, and cultural norms influence how users interact, what content is seen, and therefore what themes are discussed and the tone of those conversations. Furthermore, much like in offline settings, people adapt their communication to the norms of their environment. As each platform serves a distinct purpose and cultivates its own atmosphere, users may express themselves differently across them. This thesis explores how online discourse varies across platforms by comparing conversations in Portuguese on Instagram, Reddit, X (formerly Twitter), and TikTok. By collecting a dataset of over 150,000 social media comments, we were able to analyse the sentiment expressed in conversations, the toxicity level of the language used, and the moral framing to understand the tone and character of discussions on each platform. The findings show that tone and style shift significantly depending on the platform: X consistently hosts more toxic and emotionally negative conversations; TikTok fosters a more neutral and less confrontational tone; Reddit contains the highest use of moral language; and Instagram occupies a middle ground across most dimensions. The findings suggest social media platforms foster different communication environments that shape how discussions unfold, influencing both the tone and content of user interactions.

KEYWORDS: Comparative Analysis, Moral Foundations Theory, Natural Language Processing, Online Toxicity, Social Media

Sustainable Development Goals (SDG):

TABLE OF CONTENTS

List of Figures

List of Tables

List of Abbreviations and Acronyms

ANOVA	Analysis of Variance
CLES	Common Language Effect Size
Fcporto	Futebol Clube do Porto
MFD	Moral Foundations Dictionary
MFT	Moral Foundations Theory
PRAW	Python Reddit API Wrapper
PSD	Partido Social Democrata
PS	Partido Socialista
Slbenfica	Sport Lisboa e Benfica
Sportingcp	Sporting Clube de Portugal

1. INTRODUCTION

Social Media Platforms have fundamentally reshaped the landscape of public discourse, creating dynamic environments where people express opinions, share and consume information, and engage in discussion. Digital platforms, such as Instagram, Reddit, X (formerly Twitter), LinkedIn, Facebook and TikTok, have become central channels of communication, offering unprecedented opportunities for connectivity and the exchange of information, innovations, gossip, job opportunities, news, and ideas.

However, social media platforms are not neutral spaces, as they heavily influence how users interact, what content is seen, and the tone of conversations. The architecture, algorithms, moderation practices, and business models of each platform contribute to shaping the dynamics of discourse, fostering particular types of engagement while discouraging others (Bossetta, 2018; Kent & Taylor, 2021; Papacharissi, 2009). As a result, these environments can foster positive connections and facilitate the sharing of information. Still, they can also facilitate problematic phenomena such as misinformation, echo chambers, polarisation (Barberá, 2020; Cinelli et al., 2021; Vicario et al., 2018), toxic language, and other forms of antisocial behaviour (Matamoros-Fernández & Farkas, 2021; Siegel, A. A., 2020). These virtual environments not only mould online communication but also provide unprecedented public access to conversations between users on a wide range of topics. Together with advances in methods for analysing such content, they offer valuable data for studying human communication and behaviour at scale (Edelmann et al., 2020).

Past research tends to focus on single platforms or specific aspects of discourse, limiting the ability to draw comparative conclusions about how the same cultural and linguistic group adapts their communication across diverse digital spaces. Dunbar et al. (2015) concluded that online communities have very similar structural characteristics to offline face-to-face networks. Thus, we may assume that much like in offline settings, people might express themselves differently according to the environment in which they find themselves. Indeed, different platforms serve distinct purposes (Alhabash & Ma, 2017) and are shaped by unique contextual norms that may influence users' perceptions of what is acceptable or appropriate to share (Waterloo et al., 2018). However, just as in the offline world, different communities offer distinct experiences to their participants. What distinguishes different online environments?

This study addresses this question by comparing the discourse characteristics of user communication across social media platforms. We collect user comments from Instagram, Reddit, X, and TikTok, with a focus on discussions about politics and football in Portuguese. These comments are then analysed to assess the expressed sentiment, language toxicity, and moral framing through the lens of Moral Foundations Theory (MFT) to understand if the tone and framing of discussions vary across platforms. This work makes two contributions: first, it provides a comparative analysis of discourse characteristics across multiple platforms, examining how online discourse varies between digital environments; second, it focuses on the

Portuguese online context, addressing an underexplored linguistic and cultural setting in the study of social media discourse. Although research has been done based on data in Portuguese, it remains relatively limited in comparison to the substantial volume of studies based on English-language content, especially in terms of multiplatform comparative research. This disparity consequently constrains our understanding of whether certain phenomena and discourse dynamics generalize across different linguistic and cultural contexts. By analysing discussions in Portuguese across diverse platforms, this study contributes to a more inclusive and representative understanding of online communication.

2. RELATED WORK

Over the past two decades, social media platforms have emerged as a central channel for media sharing and consumption, attracting an ever-growing number of users worldwide and becoming primary sources for news, entertainment, and personal communication, as well as forums for public discussion. With the rise of these digital spaces, scholarly interest has grown in understanding their broader effects on society. As a result, the data generated on these platforms has become a valuable resource for studying a wide range of collective behaviours and communication dynamics (Fan et al., 2022).

2.1 Online Discourse

The problem of online toxicity has drawn significant attention in research on online communication. Defined broadly as hostile, aggressive or uncivil language, studies have addressed the prevalence and the consequences of toxicity in social media discourse, supported by advances in machine learning methods for detecting toxic language (Anjum & Katarya, 2024; Fortuna & Nunes, 2019). For example, Salehabadi et al. (2022) studied how toxic tweets or direct replies affect subsequent exchanges. Their findings indicate that while toxic conversations overall attract less engagement than their non-toxic counterparts, anonymous users are more likely to participate in toxic discussions. Within toxic threads, higher levels of toxicity are associated with increased user engagement. Recuero (2024) further argues that social media platforms' infrastructures and algorithms enable toxic discourse, framing toxicity as an effect of platforms mediating social interactions. Furthermore, within the broader domain of online toxicity, scholars have also examined related phenomena such as online hate speech and cyberbullying, focusing on how harmful language is expressed, propagated and sustained in digital platforms (Castaño-Pulgarín et al., 2021; Chatzakou et al., 2017; ElSherief et al., 2018; Salawu et al., 2020).

In addition to toxicity, online discourse is influenced by echo chambers, polarisation, and the spread of misinformation. These dynamics often overlap with toxic communication and contribute to its amplification.

Online users often gravitate toward content with which they agree and that confirms their beliefs, while disregarding or avoiding opposing viewpoints. This tendency reflects in the development of echo chambers, environments in which individuals with similar perspectives cluster together and reinforce shared narratives, limiting exposure to diverse perspectives (Del Vicario, Vivaldo, et al., 2016; Garimella et al., 2018). Del Vicario, Bessi, et al. (2016) highlights the close relationship between echo chambers, polarisation, and the proliferation of misinformation. They observe that users cluster into like-minded communities, reinforcing confirmation bias and segregation, which in turn fuels polarization. Moreover, this dynamic facilitates the circulation of biased narratives based on unverified claims, since social homogeneity, rather than information reliability, is shown to be the primary driver of content diffusion. In fact, research has shown that users are more likely to spread fake news stories than truthful ones (Vosoughi et al., 2018). Notably, Barberá et al. (2015) found that these patterns are particularly pronounced in political discussions, where users tend to form ideologically aligned networks and have limited exposure to opposing viewpoints. Bail et al. (2018) on the other hand, state that exposing people to perspectives opposite what they believe does not necessarily avoid polarisation and may even exacerbate existing beliefs. By contrast, Barberá (2014) suggested that social media use may reduce mass political polarisation over time, indicating that the relationship between social media and polarisation is complex and context-dependent.

Van Bavel et al. (2024) argue that social media often contributes to the spreading of existing moral dynamics, amplifying outrage, status seeking, and intergroup conflict while also fostering positive aspects of morality, such as social support and collective action. As people increasingly use social media platforms to share their beliefs and opinions, moral content has become widespread (Crockett, 2017; McLoughlin et al., 2021), mainly because these platforms enable messages to spread instantly across vast networks. Consequently, users are often exposed to moral values and transgressions of others, including those with very different views, which can contribute to environments that are morally charged or even hostile (Crockett, 2017). Moreover, morally framed or emotionally charged language enhances the likelihood of content being shared (Brady et al., 2017; Valenzuela et al., 2017), which may incentivize users to employ moral language to increase their reach and visibility.

To better understand the nature of this prevalent moral content on social media, the MFT provides a valuable framework, characterizing moral commentary into five core components: care, fairness, loyalty, authority, and sanctity (Graham et al., 2018). This theoretical lens has been applied in previous research to explore the moral dimensions of social media content across various subjects, including vaccine hesitancy (Beiró et al., 2023; Kalimeri et al., 2019) and the conceptualization of geopolitical conflicts, such as the war between Ukraine and the Russian Federation (Amjadi & John, 2024). This theory is a useful analytical tool for understanding user communication, as it helps dissect the moral elements within online discourse.

2.2 Platform Features and Content Dynamics

The structure, design, and moderation practices of platforms have an influence on how users engage with sensitive or polarizing topics. Yarchi et al. (2021) shows how specific algorithmic features of Facebook, Twitter, and WhatsApp contribute to different dynamics of political polarization within the context of the Israeli-Palestinian conflict. Their findings indicated that Twitter fostered ideological alignment and increased hostility, while Facebook led to aggressive but unproductive inter-group exchanges. In contrast, WhatsApp proved to be a more constructive space for discussions across different viewpoints. These results underscore how platform structure shape discussion dynamics. Further illustrating this, Guess et al. (2023) proved that even subtle changes in feed organization within the same platform, such as shifting from algorithmic to reverse-chronological feeds, significantly impact user exposure and engagement, leading to different consumption patterns for political and trustworthy content.

The role of moderation in shaping online toxicity, however, presents a more complex picture. Avalle et al. (2024) found that toxicity tends to increase as discussions progress, with remaining participants becoming more active, potentially reinforcing the formation of echo chambers. Their multi-platform study also suggested that unmoderated platforms generally exhibit higher toxicity levels than moderated ones. However, this finding is not universally consistent across all research. For instance, Noor et al. (2023) observed that Reddit and Twitter, both moderated platforms, showed higher overall toxicity levels than Parler, a platform with minimal moderation, in the context of COVID-19 discussions. Taken together, these findings indicate that the relationship between moderation and toxicity is not straightforward. Moderation may help limit toxic discourse in some contexts, however its effectiveness appears to depend on how it is implemented, the platform's design, the community's norms, and the specific topics being discussed.

Beyond platform design, the specific topics being discussed also significantly influence how conversations unfold across different digital environments. Alipour et al. (2023) investigated the evolution of discussions about COVID-19 and the launch of ChatGPT across various digital platforms, demonstrating how each platform's environment and user base foster distinct patterns of engagement. They found that even within the same platform, conversations on different topics evolve in unique ways. Moreover, Nanayakkara et al. (2024) illustrated how different platforms reflect distinct attitudes, public sentiment, and advocacy about the same social justice issue (the Black Lives Matter Movement of 2020). By comparing content across Twitter, YouTube, and online news media, they concluded that each platform emphasized different thematic and emotional tones. These findings collectively illustrate how platforms can significantly shape the thematic and emotional characteristics of public debates depending on the subject matter.

Despite growing interest in digital discourse, there is still a limited understanding of user interactions in less dominant languages, such as Portuguese. Although prior research has examined phenomena in social media content in Portuguese (Almeida et al., 2023; Carvalho et al., 2024; Guimarães et al., 2020; Leite et al., 2020; Miranda et al., 2024; Santos et al., 2022), much of these work has been limited to isolated platforms or topics, leaving a gap in cross-platform comparative research. Much of the existing literature is based on English-language content, which raises questions about the extent to which discourse patterns generalize across different linguistic and cultural contexts. As language both shapes and reflects social dynamics, overlooking linguistic diversity may constrain our understanding of how discourse unfolds in varied sociocultural contexts. Without broader linguistic diversity in research, our understanding of online discourse remains incomplete.

Building on such gap, the present study addresses this by exploring how conversations about politics and football unfold in Portuguese on Instagram, Reddit, X and TikTok. By analysing comments on posts from each of the platforms and respective sentiment conveyed, level of toxicity and moral framing (through the lens of MFT), this study seeks to provide insights into how discourse characteristics compare across different digital environments within the same linguistic context, thereby contributing to a more globally representative understanding of online communication.

3. DATA AND METHODS

3.1 DATA

The data for this study were gathered from four of the most popular social media platforms in Portugal: Instagram, Reddit, TikTok and X. These social media platforms allow users to share content through text, images or videos, as well as to interact with others via comments and likes, thereby encouraging users to express their opinions and initiate discussions. Despite of those shared functionalities, the platforms differ in design, norms, algorithms and user demographics. This study aims to determine whether these differences result in distinct digital ecosystems with different discourses.

We collect data concerning the Portuguese 2025 Legislative Elections and the Football League. These topics are broad enough to be present in discussions across the social media platforms selected and regard two polarizing themes liable to controversy, thus generating debate.

Political discussions spark a lot of debate and are prone to phenomena like echo chambers, hence being a very polarising issue on social media (Bail et al., 2018; Etta et al., 2024). The data collected precedes the 2025 legislative elections in Portugal, a period when political content on social media surges, driven by political parties sharing their ideas and media coverage of the electoral campaign, which amplifies public discourse around political themes.

Football is the most popular sport in Portugal, with millions of fans following each match and subsequently sharing their opinions on social media. Like their interactions in real life, football fans engage passionately with their favourite teams and players and often interact with each other or with rival supporters. As the national championship progresses towards the final stages, fans become increasingly tense, eager to express their support or disappointment with the game results, their team, and even the perceived unfairness of referee decisions. The intensity of those interactions reflects a wide range of sentiments, from enthusiastic support to heated disagreements, and often touches on moral and ethical issues.

These two distinct yet socially polarizing and emotionally loaded topics, allow us to compare how discourse quality might vary, not only across platforms, but also across discussions regarding different domains.

The collected sample spanned the course of six weeks, between January 4, 2025, and May 18, 2025. The only exception is the data collected from Instagram regarding football discussions that start from 07/04/2025. This period was selected to capture the social media discourse surrounding two major national events: the Portuguese legislative elections, including the official electoral campaign period from May 4 to May 16, culminating on election day, May 18; and the final stages of the Portuguese Premier Football League, with the last matches taking place on May 17. This timeframe ensures the collection of data during peak moments of public engagement across both events.

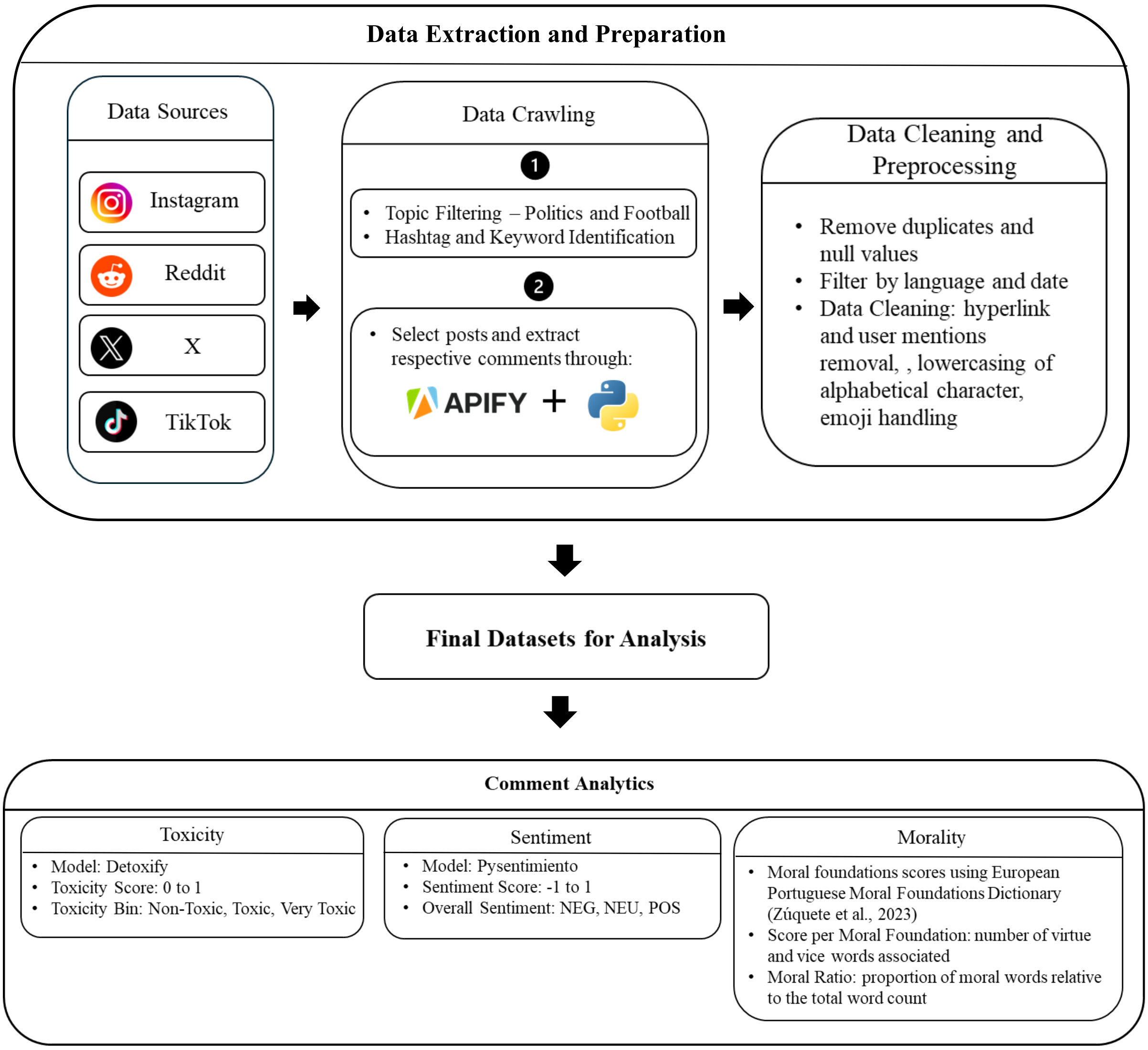


Figure 3.1 - Illustration of pipeline used for the data extraction and analysis

As data were gathered from multiple sources, the collection process varied according to the specific characteristics of each platform. The pipeline used in this study is represented in Figure 3.1. For Instagram, TikTok, and X, data was collected through Apify[1], an online platform that provides no-code web scraping tools. We utilized the available pre-built automated tools that can extract publicly accessible data from websites to gather information, including comment content, timestamps, engagement metrics (the number of likes per comment), and language. Instagram required a preliminary step to crawl and compile URLs of posts relevant to the topics of interest. This was done using Python scripts, after which Apify's automated tool retrieved the associated comments. In contrast, for X and TikTok, these tools enabled direct filtering and collection of comments on posts identified through those hashtags. For Reddit,

[1] https://apify.com/

Python was also used, through the PRAW (Python Reddit API Wrapper) library, to access posts and comments within popular Portugal-focused subreddits[2]. All subsequent data processing and analysis were conducted using Python.

To collect comments related to the topics of interest, it was first necessary to find posts about politics and football across the platforms. To do that, we utilised a set of relevant hashtags and keywords concerning each of the themes. Hashtags are labels preceded by the "#" symbol that are used on social media platforms to categorise content and facilitate the discovery of posts related to specific topics. Keywords refer to terms or phrases commonly associated with a given subject, which can also be used to identify relevant content. Posts related to the Portuguese football league were identified by the presence of hashtags or keywords such as “primeira liga,” “futebol portugal,” “liga portugal,” or names of major Portuguese football teams including “fcporto,” “slbenfica,” and “sportingcp.” The same approach was applied to the elections topic, in which case posts were selected using hashtags like “legislativas2025” and “eleições2025”, alongside the name of the parties with most representatives in the parliament - “Partido Socialista (PS)”, “Partido Social Democrata (PSD)” and “CHEGA” - as well as the names of the main candidates of those parties “Pedro Nuno Santos”, “Luís Montenegro” and “André Ventura”. Following the selection of the relevant posts, the respective user comments were collected. This process resulted in the collection of two datasets per platform, totalling eight, corresponding to the two selected topics.

In order to ensure the reliability and quality of the data for subsequent analyses, we applied a series of preprocessing steps designed to filter out irrelevant or low-quality content that could otherwise affect the accuracy of our results. First, any duplicate and null entries were removed. Given the nature of our analysis, we utilized the language field to filter the data, retaining only comments written in Portuguese and within the specified timeframe. Since the goal is to prepare the comments for toxicity detection, sentiment analysis and moral analysis, other steps include the removal of hyperlinks and user mentions (e.g., "@username"), normalisation of accented characters to their unaccented equivalents, conversion of all text to lowercase, and trimming of whitespace.

In total, 58 881 comments associated with posts about Football and 94 167 comments related to discussions about Politics were collected. Table 3.1 details the number of comments sourced from each platform and the number of comments collected per week per platform and topic.

[2] A subreddit is a community in Reddit dedicated to a specific topic or theme.

Table 3.1 - Number of comments collected by platform and topic

	Politics	Football	Total
Instagram	11548	22990	34538
Reddit	43025	24602	67627
X	32144	5668	37812
TikTok	7450	5621	13071
Total	94167	58881	153048

3.2 Toxicity Analysis

Toxicity analysis was conducted using Detoxify[3], a multilingual pre-trained model that supports Portuguese text created by Laura Hanu at Unitary. Even though Detoxify returns seven categories of toxicity scores in terms of level and type (toxicity, severe toxicity, obscene, threat insult, identity attack, and sexually explicit), in our analysis, we consider only the overall level of toxicity. The model evaluates text by providing a probability score ranging from 0 to 1, with higher scores indicating a greater likelihood that the text will be labelled as toxic. Since these scores range from 0 to 1, using a cut-off threshold of 0.5 is in line with other studies (Geller et al., 2023; Noor et al., 2023); a comment with a score greater than 0.5 is considered toxic. We further classify each comment into three categories according to their toxicity level:

- Non-toxic: Comments with a score of 0.5 or less.
- Toxic: Comments with a score greater than 0.5 but no more than 0.8.
- Very Toxic: Comments with a score greater than 0.8.

The thresholds were set to differentiate between non-toxic, moderately toxic, and highly toxic comments, facilitating interpretability in the analysis. This allows us to distinguish not only between civil and hostile discourse, but also to differentiate between more moderate and extreme forms of toxicity. In Table 3.2, we present examples of comments with their corresponding toxicity scores and levels to illustrate this distinction.

Table 3.2 – Examples of Comments with Corresponding Toxicity Scores and Levels

Comment	Toxicity Score	Toxicity Level
"Vai ter que ser um deles, mas há uma alternativa que é obrigá-los a partilhar o poder com alguém menos mau."	-0.09	Non-Toxic
"@user_mention tu é que começaste a falar de coisas que nada têm a ver com o Campeonato 24/25 por isso não vou discutir mais contigo, porque pessoas como tu são o Cancro do Futebol em Portugal."	0.72	Toxic
"@user_mention precisas de tomar a medicação, tens um grave atraso no cérebro provocado pela m**** que ouves nas televisões e jornais"	0.96	Very Toxic

[3] https://github.com/unitaryai/detoxify

Automated toxicity detection models, particularly when applied to social media comments, have constraints. As Sheth et al. (2022) noted, toxicity detection ideally should account for broader context and domain-specific analysis that state-of-the-art algorithms often overlook. Although offensive terms or coarse language can be identified through lexicons, the meaning of many words and expressions depends on how they are used and can change over time or across cultures. This means that words that seem harmless in one context might be offensive in another, and the same word may carry different meanings depending on the cultural background. Furthermore, online interactions frequently involve cultural references, slang, humour and elements such as images or emojis. Without accounting for conversational context, user relationships, or speaker intent, detection efforts risk both false positives, where harmless comments are misclassified as toxic, and false negatives, where subtle toxicity goes unnoticed.

3.3 Sentiment Analysis

In order to understand the general opinion of people on each platform we conduct sentiment analysis using Pysentimiento, a state-of-the-art toolkit to extract opinions from social media text that supports Portuguese text (Pérez et al., 2024). The model assigns a probability to each class label – positive, negative or neutral. From there, a sentiment score was calculated as the difference between the probability of the positive class and that of the negative class that reflects the overall sentiment of the text. This score ranges from -1 to 1, where:

- Values closer to 1 indicate a stronger positive sentiment
- Vaues around 0 indicate neutral sentiment
- Values near -1 indicate a stronger negative sentiment

The sentiment label corresponds to the class with the highest probability, representing the sentiment the comment is most likely to express. In Table 3.3, examples of comments are presented along with their corresponding sentiment scores and labels.

When analysing social media text, it is important to handle emojis as they have become an integral part of social media syntax. Specially for sentiment analysis, several studies have highlighted that including emojis in sentiment classifications tasks often leads to improved accuracy compared to relying only on text (Xu et al., 2024). Emojis were converted into their Portuguese textual description using the emoji library.

Table 3.3 - Examples of Comments with Corresponding Sentiment Scores and Labels

Comment	Sentiment Score	Sentiment Label
"Excelente escolha. Fiquei curioso em saber mais!"	0.98	Positive
"Honestamente não duvido, apenas espero que tenha boa cabeça para acompanhar o talento."	0.12	Neutral
"Ou as pessoas são ignorantes ou este país está entregue ao fatalismo."	-0.96	Negative

3.4 MORAL LOADING

The comments were analysed using the Moral Foundations Dictionary (MFD) in European Portuguese developed by Zúquete et al. (2023) available online. The MFD consists of a list of words and word stems, each linked with one or more moral foundations and labelled with their associated vice or virtue. Virtue content carries a positive association with the moral element, whereas vice content has a negative implication. The Moral Foundations Theory assumes the existence of five innate pillars that capture human morality (Graham et al., 2013):

- Harm/Care: Ability to feel the pain of others and be sensitive to their suffering, associated with compassion and generosity, and condemning cruelty and callousness.
- Fairness/Cheating: Concern with justice, rights, and equality, associated with altruism, honesty, and punishing cheaters.
- Loyalty/Betrayal (Ingroup): Focus on standing in one's group or community, associated with solidarity, self-sacrifice for the group, and patriotism.
- Authority/Subversion: Respect for tradition and legitimate authority, associated with leadership and social order.
- Purity/Degradation: Sense that certain things are sacred or pure and must be protected, associated with disgust and contamination avoidance.

We apply the MFT framework to characterize each comment into five elements when applicable. For each instance, there are corresponding columns indicating the number of virtue and vice words associated with each moral foundation on the MFD. From this data, we also derive the moral ratio, which reflects the proportion of moral words relative to the total word count in the comment. In Table 3.4, examples of comments with moral language are displayed along with their moral ratio and the corresponding moral foundation detected.

Table 3.4 - Examples of Comments with Corresponding moral ratio and Moral Foundations detected

Comment	Moral Ratio	Moral Foundation Detected
"Tenho muita pena que este partido se aproveite da dor das pessoas mais vulneráveis para os enganar com promessas vazias."	0.1	Harm
"Acho que nunca vi uma equipa com 61% de posse perder 1-4 justamente."	0.07	Fairness
"Juntos somos mais fortes. Precisamos de mudar este país."	0.22	Ingroup
"Na minha opinião centram a sua atuação em causas muito distantes dos problemas reais do país e quando falam de problemas reais não têm soluções."	0.04	Authority
"Depois perguntam-me porque tenho nojo do teu clube. É por causa de adeptos como tu!"	0.63	Purity

By examining online discourse through the lens of MFT, we can gain a deeper understanding of how the content and digital ecosystems of different platforms influence users' moral judgments. Our analysis enables us to understand how moral discourse is expressed on each of the online platforms and how it varies across different topics. Our focus is on identifying which of the five elements of MFT are most present when users engage in discussions about politics and football.

4. RESULTS AND DISCUSSION

In this chapter, we present a comparative analysis of Portuguese discourse on social media platforms, namely Instagram, Reddit, X, and TikTok. We focus on how discussions about politics and football across these platforms differ using the metrics extracted from user comments as a result of the previously explained methodology. Our analysis aims to compare the conversation dynamics of diverse social media platforms, taking into account the sentiment expressed in user comments and the prevalence of toxic language. These measures allow us to investigate the existence of different ambiences and whether specific online environments are more conducive to negative or inflammatory discourse than others.

The sentiment expressed by users in comments on Instagram, Reddit, and TikTok tends to be predominantly neutral. X, however, stands out as the only platform where users primarily express negative sentiment in conversations. Analysing Figure 4.1, it is clear that the distribution of the sentiment scores is different across digital platforms. Instagram and TikTok both show more symmetric distributions, indicating a balanced spread of sentiment across comments. Reddit's distribution, by contrast, is skewed with a longer tail toward negative sentiment. X shows a similarly balanced distribution overall, but a broader spread into negative sentiment, reinforcing the platform's tendency toward more negative conversations.

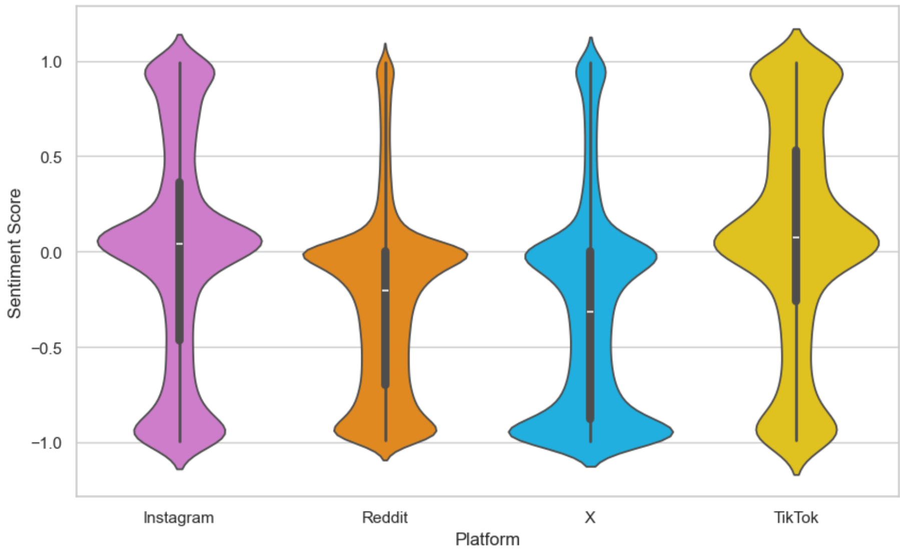


Figure 4.1 - Sentiment score distribution per platform

As shown in Table 4.1, X has the lowest mean sentiment score, reflecting the overall negativity in its user comments. In contrast, Instagram and TikTok had higher mean sentiment scores, indicating more neutral or slightly positive discourse. Reddit is in between but leaned more negative compared to the others. To determine whether the observed differences in mean sentiment scores across platforms are statistically significant, we then conduct a one-way ANOVA. The results, shown on Table 4.2, revealed a statistically significant effect of platform on sentiment score, indicating that the mean sentiment scores differed significantly between at least two of the platforms.

Table 4.1 - Mean and standard deviation of sentiment scores across platforms

	Mean Sentiment Score	Standard Deviation
Instagram	-0.01	0.59
Reddit	-0.28	0.47
X	-0.3	0.55
TikTok	0.08	0.59

Table 4.2 - One-way ANOVA results for sentiment scores across platforms

	Sum of Squares	df	Mean Square	F	p-value
Between Groups	3297.88	3	1099.29	3923.04	<.001
Within Groups	42885.13	153044	0.28	-	-
Total	46183.01	153047	-	-	-

In order to identify which specific platform pairs differed significantly in their mean sentiment scores, a Games-Howell post hoc test was performed. This test was chosen due to the violation of the homogeneity of variances assumption ($p < .001$; Levene's test – see Table B. 1) and unequal group sizes across platforms, which violate the assumptions required for Tukey's HSD test. The results, in Table 4.3 , indicate that sentiment scores differ significantly across all platform pairs ($p < .001$). The most pronounced differences were between TikTok and X and Reddit and TikTok, indicating considerably more negative sentiment on X and Reddit. The smallest significant difference was found between Reddit and X, suggesting a modest effect despite statistical significance.

Table 4.3 - Games–Howell pairwise comparisons of sentiment scores between platforms

Comparison	Mean Difference	SE	95% CI	p-value	Hedges' g
Instagram vs Reddit	-0.2681	0.0037	[0.2609, 0.2753]	<.001	0.52
Instagram vs TikTok	-0.0880	0.0061	[-0.0998, -0.0761]	<.001	0.15
Instagram vs X	0.3189	0.0043	[0.3106, 0.3273]	<.001	0.56
Reddit vs TikTok	-0.3561	0.0055	[-0.3668, -0.3454]	<.001	0.73
Reddit vs X	0.0508	0.0033	[0.0443, 0.0574]	<.001	0.10
TikTok vs X	0.4069	0.0059	[0.3954, 0.4184]	<.001	0.73

When it comes to the toxicity in the language used online, results suggest that the prevalence of toxic speech in social media discourse is overall very low. X stands out once again as the platform that contains a higher percentage of comments labelled as toxic (Toxicity score > 0.5) with 18% of total comments. Thus, X not only displays the most negative tone in discourse, but also the highest overall proportion of toxic comments in our data. The proportion of toxic comments on the remaining online platforms considered is less than 10% of the total comments (8% for Reddit, 6% for Instagram, and 4% for TikTok), as displayed in Figure 4.2. These proportions are consistent with the mean toxicity scores presented in Table 4.4, which show that X has the highest average toxicity score, while TikTok has the lowest. This further supports the conclusion that X fosters more toxic discourse on average, both in frequency and intensity.

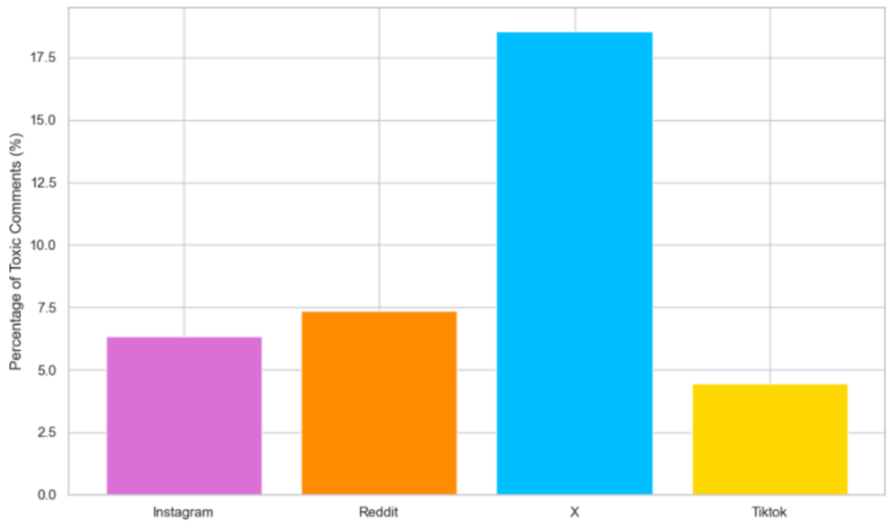


Figure 4.2 - Proportion of toxic comments per platform

Table 4.4 - Mean and standard deviation of toxicity scores across platforms

	Mean Toxicity Score	**Standard Deviation**
Instagram	0.08	0.20
Reddit	0.09	0.22
X	0.2	0.32
TikTok	0.05	0.15

These findings are further supported by the distribution of toxicity scores, as shown in Figure 4.3. While all platforms display distributions skewed toward low toxicity, indicating that most comments are non-toxic, X stands out with a broader spread and a heavier tail extending toward higher toxicity values.

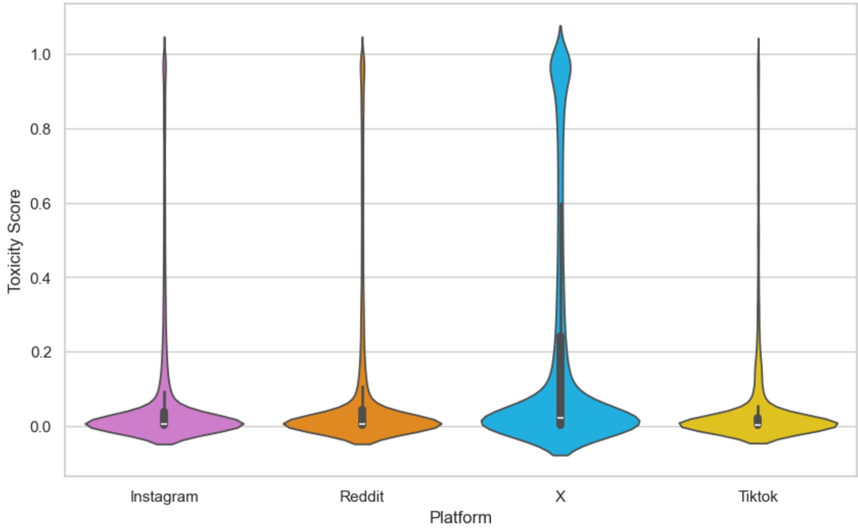


Figure 4.3 - Distribution of Toxicity Scores across Platforms

This suggests not only a higher proportion of toxic comments, but also greater variation in how toxicity manifests on the platform. In contrast, Instagram, Reddit, and TikTok all show tight concentrations near zero, with relatively narrow tails, reinforcing the overall low prevalence and intensity of toxic speech on these platforms.

A one-way ANOVA was conducted to compare the mean toxicity scores of user comments across Instagram, Reddit, X, and TikTok (Table 4.5). The analysis revealed a statistically significant effect of platform on toxicity scores, indicating that the mean toxicity scores differ significantly across at least two of the platforms.

Table 4.5 - One-way ANOVA summary for toxicity scores across platforms

	Sum of Squares	df	Mean Square	F	p-value
Between Groups	389.018	3	129.672	2216.413	<.001
Within Groups	8953.926	153044	0.059	-	-
Total	9342.944	153047	-	-	-

As with the sentiment scores, Levene's test indicates a violation of the assumption of homogeneity of variances ($p < .001$, Levene's test – see Table B. 1**Error! Reference source not found.**), as well as considerable differences in group sample sizes, therefore a Games-Howell post hoc test was employed to identify which specific platform pairs differed significantly in their mean toxicity scores.. The results, displayed in Table 4.6, show that all pairwise comparisons between platforms are statistically significant ($p < .001$). The largest differences in toxicity scores were observed between TikTok and X, and between Instagram and X, while the smallest but still significant difference was found between Instagram and Reddit. Although all pairwise differences were statistically significant, the actual differences in mean toxicity scores were relatively small in magnitude, given that toxicity scores range from 0 to 1. This is reflected in the effect sizes, most of which fall in the small to moderate range, indicating that while the differences between platforms are consistent and meaningful, they are not necessarily large in absolute terms.

Table 4.6 - Games-Howell post hoc test results for differences in toxicity scores between platforms

Comparison	Mean Difference	SE	95% CI	p-value	Hedges' g
Instagram vs Reddit	-0.0163	0.0014	[-0.0189, -0.0136]	< .001	0.08
Instagram vs TikTok	0.0252	0.0017	[0.0218, 0.0285]	< .001	0.14
Instagram vs X	-0.1204	0.0020	[-0.1242, -0.1165]	< .001	0.45
Reddit vs TikTok	0.0414	0.0016	[0.0383, 0.0446]	< .001	0.19
Reddit vs X	-0.1041	0.0019	[-0.1078, -0.1005]	< .001	0.40
TikTok vs X	-0.1455	0.0021	[-0.1497, -0.1414]	< .001	0.50

To better understand the intensity of toxic discourse on each platform, we further classify all comments into three toxicity categories based on their toxicity score: non-toxic (≤ 0.5), toxic (> 0.5 and ≤ 0.8), and very toxic (> 0.8). The results show that across all platforms, except TikTok, the number of very toxic comments exceeds the number of toxic comments, which indicates that when toxicity is present, it tends to occur in more severe forms. This suggests that on Instagram, Reddit, and X, the intensity of toxic discourse is more hostile. On TikTok, we observe the opposite, with fewer very toxic comments relative to toxic ones.

In order to examine the relationship between sentiment and toxicity, we analysed the mean sentiment scores across different toxicity categories. As illustrated in Figure 4.4 there is a consistent trend across platforms: as toxicity levels rise, the average sentiment scores tend to become more negative. Non-toxic comments express on average a neutral sentiment, while toxic and very toxic comments show progressively lower mean sentiment scores, confirming that a more emotionally negative tone accompanies toxicity.

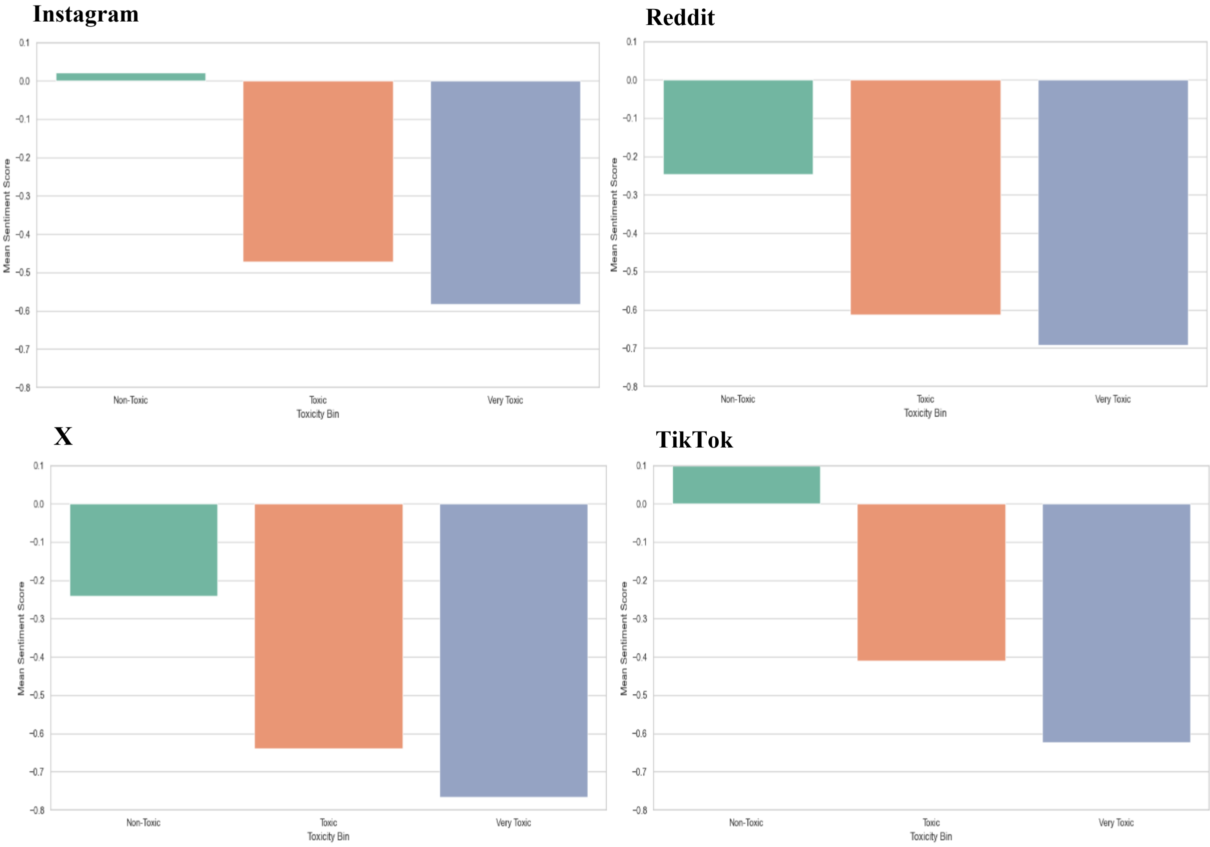


Figure 4.4 - Average sentiment score per toxicity level per latform

To visualise the relationship between sentiment and toxicity scores across platforms, Figure 4.5 presents two-dimensional histograms showing the joint distribution of these variables for each platform. Rather than individual points, the plots display the density of comments across different combinations of sentiment and toxicity scores, where lighter areas indicate a higher concentration of comments. While no clear linear relationship is evident, the plots reveal that higher toxicity scores generally coincide with more negative sentiment scores, particularly on platforms like X and Reddit. This suggests that emotionally negative language is more likely to be associated with toxicity, though the strength of this association appear to vary across platforms. This pattern is further supported by Spearman correlation coefficients, which are negative across all platforms, with the strongest association observed on X (ρ = –0.48), followed by Reddit (ρ = -0.37), and TikTok (ρ = -0.27).

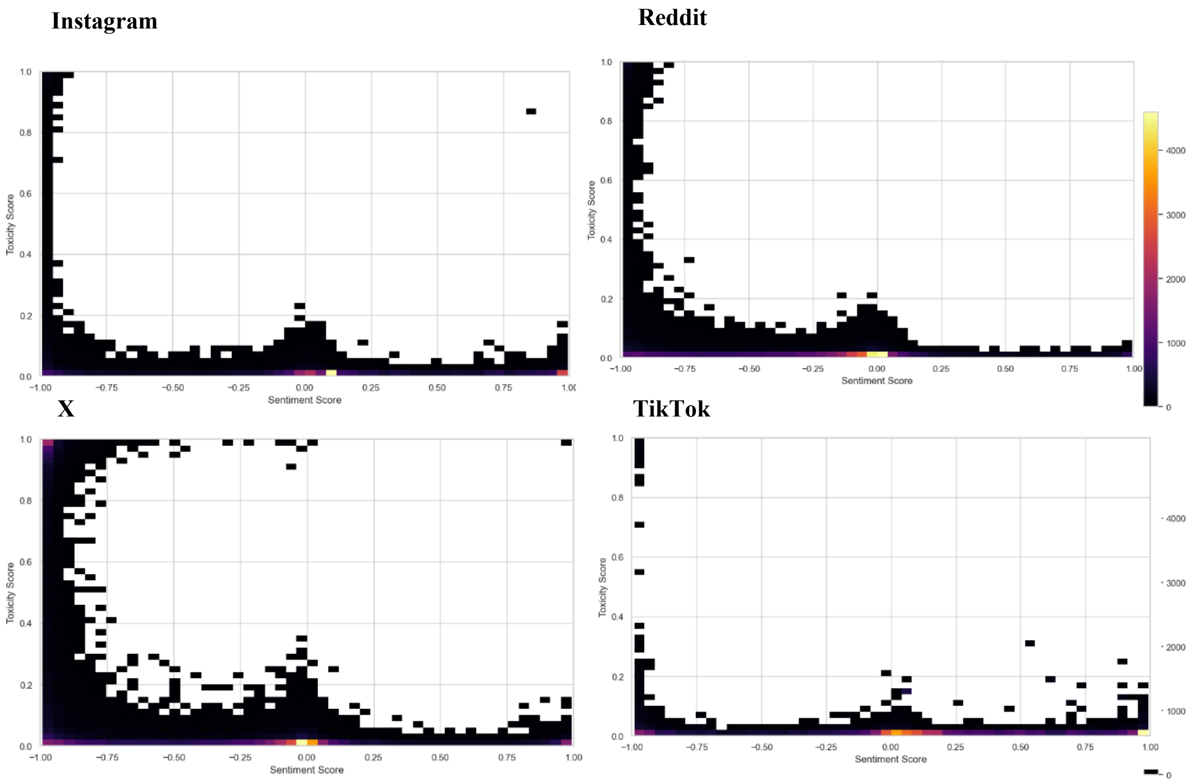


Figure 4.5 – 2D Histograms of sentiment vs. toxicity scores of comments across platforms

Prior research suggests that user engagement can reinforce toxic dynamics by helping to spread and normalise incivility (Avalle et al., 2024). If toxic or emotionally harmful comments consistently receive more likes, this could indicate a feedback loop, where users adopt more hostile tones as a way to gain more attention and approval. Likes are a measure of engagement and can indicate which types of comments are valued or encouraged by users. The content that receives engagement on social media depends on multiple factors, such as the number of people who see the post and its comments, and who those people are. People on different platforms may react differently to the same content, and users may adjust their tone or style of communication to drive engagement. This raises the question whether the platforms with the most toxic and negative discourse are also those where this kind of content is most rewarded. To address this, we employ the z-score of comment 'likes' as a proxy for user endorsement, allowing us to examine how engagement varies across sentiment and toxicity levels on each platform. This normalization accounts for differences in overall engagement rates and sample sizes across platforms. A z-score measures how much a comment like count deviates from the average number of likes on that platform, expressed in standard deviations. A positive z-score indicates that the comment received more likes than the platform average, while a negative z-score means it received fewer likes. Table 4.7 shows the average Z-score of comment likes by sentiment label (positive, neutral, negative) across platforms, alongside the mean and standard deviation of likes per comment. Table 4.8 presents the same metrics, organized by toxicity level (non-toxic, toxic, very toxic).

Table 4.7 - Z-Scores of Likes, Average Likes, and Comment Counts per Sentiment Label Across Platforms

Platforms	Avg. Likes	Std. Deviation	Sentiment Label	# Comments	Avg. Z-Score (Likes)
Instagram	7.1	48.1	Positive	8080	0.03
			Neutral	17419	-0.02
			Negative	9039	0.02
Reddit	6.3	22.6	Positive	5089	0.02
			Neutral	36818	-0.02
			Negative	25720	0.05
X	27	533.7	Positive	3897	0.04
			Neutral	16587	0.01
			Negative	17328	-0.01
TikTok	31.3	277.5	Positive	3801	0.01
			Neutral	6457	-0.01
			Negative	2813	0.01

Table 4.8 - Z-Scores of Likes, Average Likes, and Comment Counts per Toxicity Level Across Platforms

Platforms	Avg. Likes	Std. Deviation	Toxicity Level	# Comments	Avg. Z-Score (Likes)
Instagram	7.1	48.1	Non-Toxic	32534	<0.01
			Toxic	943	<-0.01
			Very Toxic	1061	-0.09
Reddit	6.3	22.6	Non-Toxic	62319	<-0.01
			Toxic	2394	0.04
			Very Toxic	2914	0.06
X	27	533.7	Non-Toxic	4591	-0.01
			Toxic	2203	<0.01
			Very Toxic	31018	0.04
TikTok	31.3	277.5	Non-Toxic	12613	-0.01
			Toxic	238	0.05
			Very Toxic	210	0.02

Surprisingly, our results indicate that the assumption of a feedback loop contributing to the perpetuation of toxicity and negativity in discourse does not hold, as measured by comment likes. We found no meaningful difference in how user' likes are distributed across comments based on their overall sentiment or toxicity level. The average z-score for likes across sentiment and toxicity levels, as well as across all platforms, consistently falls within an extremely narrow range, between -0.01 and 0.01. These findings suggest that likes alone may not capture how users respond to the sentiment or toxicity expressed in comments. Other factors, such as comment visibility and the degree to which users agree or support what is being said likely play a stronger role in shaping engagement.

To further investigate the types of discussions on each platform, we can examine the moral framing used in these discussions. Our goal is to understand whether platforms where toxic and emotionally negative discourse is more common are also spaces where users rely more on moral language to express or justify their views. In other words, does morality influence how hostility is communicated, and does this influence vary depending on the platform?

We use MFT, which breaks down morality into five foundations: Harm, Authority, Ingroup, Fairness, and Purity to investigate such a link. Words and expressions related to these values are identified in user comments to examine how frequently moral language is used and how it correlates with toxicity and sentiment. It is important to clarify that this analysis focuses solely on textual expressions of moral language and does not represent the moral values of the people engaging in these discussions.

Firstly, it is worth noting that most users across all platforms do not employ moral language in their comments. However, Reddit leads with the highest percentage of comments containing moral language (33%), followed by X (25%), Instagram (14%), and TikTok (12%) (Figure 4.6).

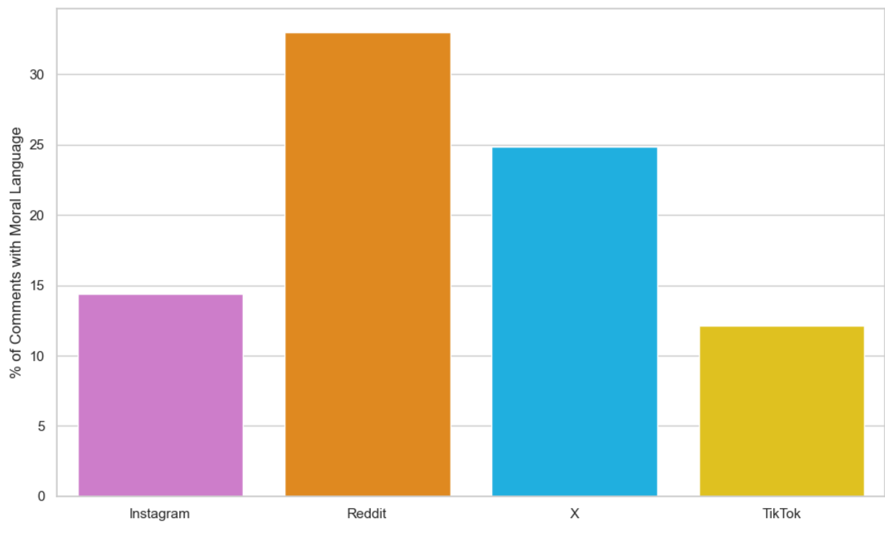


Figure 4.6 - Proportion of comments containing moral language per platform

Harm, Authority, and Ingroup are the moral foundations most detected to be frequently detected in the comments, suggesting that users often frame their comments around protection, respect for hierarchy, and loyalty to groups. The distributions of moral foundations among comments that contain moral language per platform can be found in Figure A. 1.

In the context of the discussions analysed, about football and politics, this makes sense: political conversations often involve themes of social justice, institutional respect, or national identity. In contrast, football discussions can reflect group loyalty and rivalry. It is also worth noting that comments regarding political discussions tend to use more moral language than those concerning football.

Interestingly, the platforms with the lowest overall proportion of comments containing moral language appear to show the highest average moral ratio when moral language is used. As shown in Table 4.9, TikTok has the highest mean moral ratio, followed by Instagram, while Reddit and X display lower averages, when only comments containing moral language are considered. These differences suggest variation not only in the presence of moral language but also in how extensively it is used within individual comments.

Table 4.9 - Mean and standard deviation of moral ratio, among comments with moral language, across platforms

	Mean	**Standard Deviation**
Instagram	0.11	0.13
Reddit	0.05	0.08
X	0.09	0.09
TikTok	0.12	0.12

To assess whether these observed differences are statistically significant, a one-way ANOVA was conducted on moral ratio scores among moral comments (Table 4.10). The results in show there is a statistically significant difference in the proportion of moral language among comments that contained moral content across platforms. This indicates that when users do employ moral language extent to which they do so differs significantly across at least two of the platforms. To identify these specific differences, a Games-Howell post hoc test was performed. The results, presented in Table 4.11, show that all pairwise comparisons between platforms were statistically significant ($p<.001$). It is worth noting however that the average word count per comment differed across platforms, which may have influenced the observed differences in moral ratio.

Table 4.10 - One-way ANOVA summary for moral ratio, among comments with moral language, across platforms

	Sum of Squares	df	Mean Square	F	p-value
Between Groups	14.94	3	4.981	575.80	<.001
Within Groups	331.49	38317	0.00865	-	-
Total	346.432	38320	-	-	-

Table 4.11 - Games-Howell post hoc test results for differences in toxicity scores between platforms

Comparison	Mean Difference	SE	95% CI	p-value	Hedges' g
Instagram vs Reddit	0.04787	0.00192	[0.0441, 0.0516]	< .001	0.52
Instagram vs TikTok	-0.01256	0.00359	[–0.0196, –0.0055]	< .001	-0.10
Instagram vs X	0.02124	0.00206	[0.0172, 0.0253]	< .001	0.20
Reddit vs TikTok	-0.06043	0.00313	[–0.0666, –0.0543]	< .001	-0.71
Reddit vs X	-0.02663	0.00107	[–0.0287, –0.0245]	< .001	-0.32
TikTok vs X	0.03379	0.00321	[0.0275, 0.0401]	< .001	-0.36

Our analysis reveals a consistent trend across all platforms studied: comments containing moral language are more negative in sentiment and more toxic on average than those without, as illustrated by Figure 4.7 and Figure 4.8. We found statistically significant differences in both sentiment and toxicity scores across all platforms using the Mann-Whitney U test, given the non-normal distribution of scores (all p<0.001). The detailed results of the test are in Table B. 2 and Table B. 3.

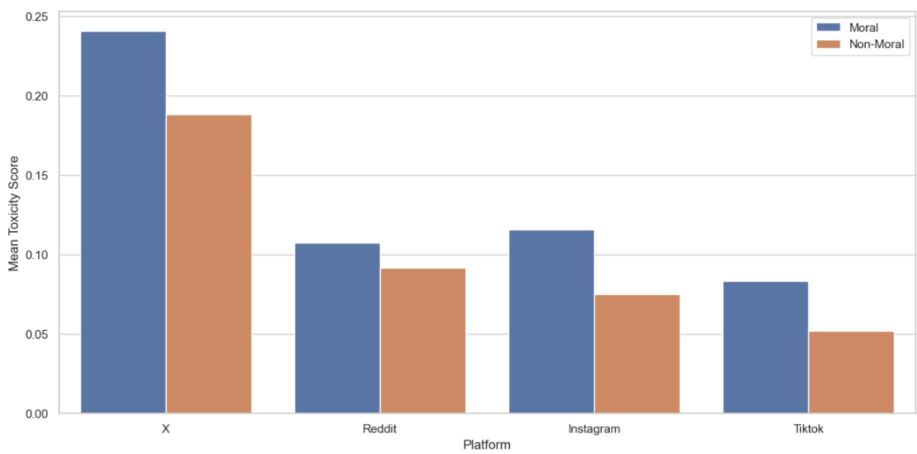


Figure 4.7 - Average toxicity score of Moral vs Non-Moral comments per platform

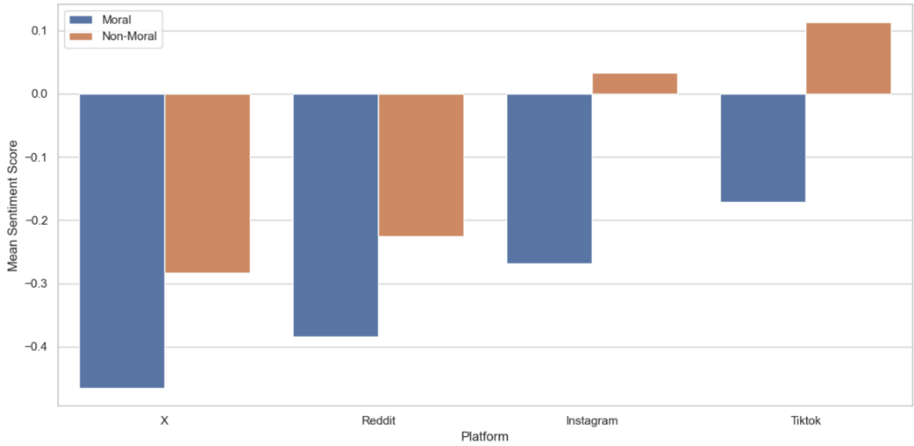


Figure 4.8 - Average sentiment score of Moral vs Non-Moral comments per platform

To understand the scale of these differences, we used the Common Language Effect Size (CLES), which represents the probability that a randomly selected score from comments with moral language is greater than a randomly selected score from comments without moral language. For sentiment, the probability that a comment containing moral language would express a more negative tone than one without was higher across all platforms: 64.8% on Instagram, 61.1% on Reddit, 59.9% on X, and 63.9% on TikTok. Similarly, for toxicity, the probability that a moral comment is more toxic than a non-moral one is 60.6% on Instagram, 51.2% on Reddit, 54.3% on X, and 60.4% on TikTok. The results suggest that moral framing often appears in more emotionally charged and confrontational contexts. Morality seems to intensify dialogue on social media, providing a way of users to justify stronger emotional responses or more aggressive language.

A particularly notable pattern emerges in relation to the Purity foundation. Among moral comments, we verify that Purity increases with toxicity. Figure 4.9 shows the distribution of moral foundations across toxicity levels.

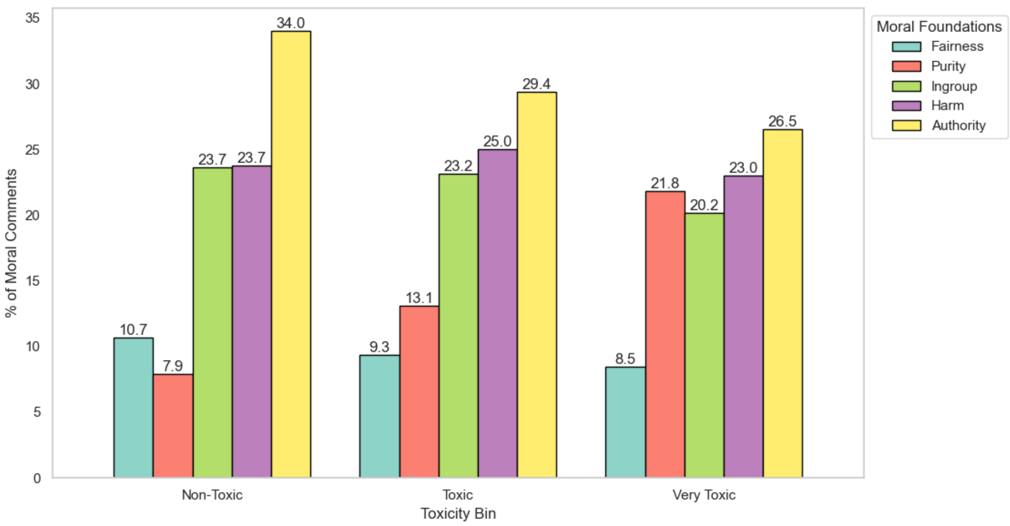


Figure 4.9 - Distribution of moral foundations across toxicity levels

While only 8% of non-toxic moral comments reference Purity, this rises to 22% among very toxic moral comments. This finding is consistent with psychological research suggesting that disgust (the emotion most closely related with Purity) tends to increase moral judgement and is linked to harsher responses (Krebs, 2022). Studies have also shown that moral disgust activates the same expressions and cognitive responses as physical disgust (Chapman et al., 2009). In this context, references to Purity may reflect an emotional response of disgust that is moralized and expressed through hostile language, particularly in toxic conversations.

We conclude that while moral language is present across all platforms, we do not find that it is necessarily more common on platforms with higher levels of toxicity or negativity. For example, Reddit, which has relatively high toxicity, also exhibits the highest frequency of moral language; however, X, which is more toxic overall, displays less moral framing. This suggests that the use of moral language may depend more on platform culture and user behaviour than on toxicity levels alone. It is also important to consider that differences in comment length across platforms, resulting from character limitations, could influence the frequency and expression of moral language, representing a limitation in direct comparisons of moral framing. The results show that when morality is present in a comment, it often coincides with more negative sentiment and higher toxicity, regardless of platform. This means that while moral discourse is not always hostile, it tends to appear more frequently in emotionally intense or confrontational discussions. This supports the idea that morality can play a role in how hostility is communicated, serving not only to frame arguments but also to justify stronger or more aggressive expression.

5. CONCLUSION

This study aimed to gain a deeper understanding of the discourse characteristics of user communication across Instagram, Reddit, X, and TikTok, with a focus on discussions related to politics and football within the Portuguese community. Through the analysis of comments on posts related to these themes, we were able to depict the sentiments expressed by users, the toxicity level of the language used, and the morality conveyed in each comment. This allowed for the comparison of discussions about the same themes, on the same timeframe across different social media platforms and draw conclusions about whether different digital environments are associated with distinct discourse characteristics.

Our analysis reveals variations in the characteristics of online discourse across the platforms considered, while also identifying some consistent patterns. We found the overall prevalence of toxic speech in online environments in Portuguese to be low. However, X consistently stands out as the platform with the most toxic and negative discourse, as it not only presents the highest proportion of toxic comments but also a predominantly negative sentiment in its overall discourse. For Instagram, Reddit, and X, when toxicity is present, it often manifests in more severe forms, with very toxic comments outnumbering toxic ones. Conversely, TikTok showed a different pattern, with a higher proportion of moderately toxic comments. We also observed a consistent negative correlation between sentiment and toxicity across all platforms, suggesting that more toxic comments tend to express greater negative sentiment.

Contrary to expectations, we found no evidence of a feedback loop whereby toxic or emotionally negative comments received more user endorsement in the form of likes. Engagement does not meaningfully vary across toxicity or sentiment levels. This suggests that other factors may play a larger role in driving engagement than the tone or civility of discourse alone. However, a limitation of this analysis is that it only considers likes as a proxy for engagement, excluding other relevant forms such as comment replies. This is due to data availability constraints, as such metrics were not accessible through the data collection process used in this study.

When it comes to users comments moral framing, our analysis reveals that most comments do not contain moral language, although Reddit featured the highest proportion of morally framed discourse, followed by X, Instagram, and TikTok. Another finding is that when moral language is employed, it tends to coincide with more negative sentiment and higher toxicity, supporting the idea that morality is often invoked in emotionally charged or confrontational contexts. Interestingly, even though moral language is linked to stronger sentiment and toxicity, it is not necessarily more prevalent on platforms with higher overall prevalence of toxic discourse. This suggests that moral framing may reflect specific platform discussion dynamics rather than simply being a marker of toxicity. We also observed that references to the Purity foundation were more common in very toxic comments, aligning with psychological theories that link moral disgust to harsher judgments and more hostile expressions.

Notably, our study reveals statistically significant differences in sentiment, toxicity, and moral language use across Instagram, Reddit, X, and TikTok. X consistently stands out, presenting the most negative sentiment and highest prevalence of toxic language in user comments, with statistically significant differences from all other platforms for all metrics. This suggests a unique environment on X that appears more conducive to hostile discourse. In terms of sentiment expressed in comments, Instagram and TikTok show the most similar mean scores, both leaning towards neutral or slightly positive discourse, and their distributions of sentiment are more symmetric. Similarly, for toxicity, Instagram, Reddit, and TikTok generally present lower mean toxicity scores and more concentrated distributions compared to X, indicating less intense and frequent toxic exchanges. Even though all pairwise comparisons for both sentiment and toxicity were statistically significant, the smallest differences in mean sentiment were found between Reddit and X, and for toxicity, between Instagram and Reddit, indicating these pairs are relatively more similar in their respective characteristics despite the statistical significance. Conversely, TikTok and X consistently demonstrate the largest and most pronounced differences in both sentiment and toxicity, highlighting them as the most distant platforms in terms of discourse characteristics. Our findings also indicate that while moral language is present across platforms, its use is often associated with more negative sentiment and higher toxicity, regardless of the platform, suggesting that moral framing frequently occurs within emotionally charged or confrontational contexts.

This study is not without limitations. The sample we collected focuses on discussions regarding solely two topics which, although distinct, leave out many other subjects that could display different patterns of discourse. In addition, the analysis is limited to a single language, Portuguese, which means the findings may not be generalisable to discussions in other linguistic or cultural contexts. Furthermore, it is important to recognize the limitations associated to the automated language models used for toxicity and sentiment analysis. These models enable large-scale insights, but they struggle to capture nuance, irony, and humour in text, which is especially relevant in the context of social media data where language tends to be more informal and idiomatic. Variations in comment length and formats across platforms, due to character limits, may also influence how often and how explicitly moral language is expressed, which should be considered when comparing moral framing across digital environments. It this study we did not control for the number of words per comment, which may affect the likelihood of detecting sentiment, toxicity, or moral language, particularly when comparing shorter comments (e.g., on TikTok) with longer ones (e.g., on Reddit). This limitation is especially relevant for the interpretation of the moral ratio, as shorter comments containing one or two moral words can yield disproportionately high ratios, while longer comments with more moral content may show lower ratios due to dilution across higher word counts. Moreover, analysing comments in isolation overlooks broader conversational context, such as replies, threads or post structure, which can shape how meaning, sentiment, or morality are communicated.

By analysing discussions in Portuguese across multiple platforms, this study contributes to a more inclusive and representative understanding of online communication. In doing so, the study extends existing research by incorporating a less commonly studied linguistic context and emphasizing the role of the digital environment in shaping how users interact, offering a broader understanding of the characteristics of user communication across social media platforms.

BIBLIOGRAPHICAL REFERENCES

Alhabash, S., & Ma, M. (2017). A Tale of Four Platforms: Motivations and Uses of Facebook, Twitter, Instagram, and Snapchat Among College Students? *Social Media + Society*, *3*(1), 2056305117691544. https://doi.org/10.1177/2056305117691544

Alipour, S., Galeazzi, A., Sangiorgio, E., Avalle, M., Bojic, L., Cinelli, M., & Quattrociocchi, W. (2023). *Cross-Platform Social Dynamics: An Analysis of ChatGPT and COVID-19 Vaccine Conversations* (No. arXiv:2310.11116). arXiv. https://doi.org/10.48550/arXiv.2310.11116

Almeida, P., Pereira, J., & Candido, D. (2023). Online hate speech on social media in Portugal: Extremism or structural racism? *Social Identities*, *29*(5), 419–435. https://doi.org/10.1080/13504630.2024.2324277

Amjadi, E., & John, R. S. (2024). Count on Me: Moral Language in Social Media and Policy Discourse during the Ukraine-Russia Conflict. *Data & Policy*, *7*, e22. https://doi.org/10.1017/dap.2024.98

Anjum, & Katarya, R. (2024). Hate speech, toxicity detection in online social media: A recent survey of state of the art and opportunities. *International Journal of Information Security*, *23*(1), 577–608. https://doi.org/10.1007/s10207-023-00755-2

Avalle, M., Di Marco, N., Etta, G., Sangiorgio, E., Alipour, S., Bonetti, A., Alvisi, L., Scala, A., Baronchelli, A., Cinelli, M., & Quattrociocchi, W. (2024). Persistent interaction patterns across social media platforms and over time. *Nature*, *628*(8008), 582–589. https://doi.org/10.1038/s41586-024-07229-y

Bail, C. A., Argyle, L. P., Brown, T. W., Bumpus, J. P., Chen, H., Hunzaker, M. B. F., Lee, J., Mann, M., Merhout, F., & Volfovsky, A. (2018). Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, *115*(37), 9216–9221. https://doi.org/10.1073/pnas.1804840115

Barberá, P. (2014). *How Social Media Reduces Mass Political Polarization. Evidence from Germany, Spain, and the U.S.*

Barberá, P. (2020). Social Media, Echo Chambers, and Political Polarization. In N. Persily & J. A. Tucker (Eds.), *Social Media and Democracy* (1st ed., pp. 34–55). Cambridge University Press. https://doi.org/10.1017/9781108890960.004

Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A., & Bonneau, R. (2015). Tweeting From Left to Right: Is Online Political Communication More Than an Echo Chamber? *Psychological Science*, *26*(10), 1531–1542. https://doi.org/10.1177/0956797615594620

Beiró, M. G., D'Ignazi, J., Prado, M. F., Bustos, V. P., & Kalimeri, K. (2023). Moral Narratives Around the Vaccination Debate on Facebook. *Proceedings of the ACM Web Conference 2023*, 4134–4141. https://doi.org/10.1145/3543507.3583865

Bossetta, M. (2018). The Digital Architectures of Social Media: Comparing Political Campaigning on Facebook, Twitter, Instagram, and Snapchat in the 2016 U.S. Election. *Journalism & Mass Communication Quarterly*, *95*(2), 471–496. https://doi.org/10.1177/1077699018763307

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(28), 7313–7318. https://doi.org/10.1073/pnas.1618923114

Carvalho, P., Caled, D., Silva, C., Batista, F., & Ribeiro, R. (2024). The expression of hate speech against Afro-descendant, Roma, and LGBTQ+ communities in YouTube comments. *Journal of Language Aggression and Conflict*, *12*(2), 171–206. https://doi.org/10.1075/jlac.00085.car

Castaño-Pulgarín, S. A., Suárez-Betancur, N., Vega, L. M. T., & López, H. M. H. (2021). Internet, social media and online hate speech. Systematic review. *Aggression and Violent Behavior*, *58*, 101608. https://doi.org/10.1016/j.avb.2021.101608

Chapman, H. A., Kim, D. A., Susskind, J. M., & Anderson, A. K. (2009). In Bad Taste: Evidence for the Oral Origins of Moral Disgust. *Science*, *323*(5918), 1222–1226. https://doi.org/10.1126/science.1165565

Chatzakou, D., Kourtellis, N., Blackburn, J., De Cristofaro, E., Stringhini, G., & Vakali, A. (2017, February 22). *Mean Birds: Detecting Aggression and Bullying on Twitter*. arXiv.Org. https://arxiv.org/abs/1702.06877v3

Cinelli, M., De Francisci Morales, G., Galeazzi, A., Quattrociocchi, W., & Starnini, M. (2021). The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, *118*(9), e2023301118. https://doi.org/10.1073/pnas.2023301118

Crockett, M. J. (2017). Moral outrage in the digital age. *Nature Human Behaviour*, *1*(11), 769–771. https://doi.org/10.1038/s41562-017-0213-3

Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, H. E., & Quattrociocchi, W. (2016). The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, *113*(3), 554–559. https://doi.org/10.1073/pnas.1517441113

Del Vicario, M., Vivaldo, G., Bessi, A., Zollo, F., Scala, A., Caldarelli, G., & Quattrociocchi, W. (2016). Echo Chambers: Emotional Contagion and Group Polarization on Facebook. *Scientific Reports*, *6*, 37825. https://doi.org/10.1038/srep37825

Dunbar, R. I. M., Arnaboldi, V., Conti, M., & Passarella, A. (2015). The structure of online social networks mirrors those in the offline world. *Social Networks*, *43*, 39–47. https://doi.org/10.1016/j.socnet.2015.04.005

Edelmann, A., Wolff, T., Montagne, D., & Bail, C. A. (2020). Computational Social Science and Sociology. *Annual Review of Sociology*, *46*(1), 61–81. https://doi.org/10.1146/annurev-soc-121919-054621

ElSherief, M., Kulkarni, V., Nguyen, D., Wang, W. Y., & Belding, E. (2018). Hate Lingo: A Target-Based Linguistic Analysis of Hate Speech in Social Media. *Proceedings of the International AAAI Conference on Web and Social Media*, *12*(1), Article 1. https://doi.org/10.1609/icwsm.v12i1.15041

Etta, G., Cinelli, M., Marco, N. D., Avalle, M., Panconesi, A., & Quattrociocchi, W. (2024). A Topology-Based Approach for Predicting Toxic Outcomes on Twitter and YouTube. *IEEE Transactions on Network Science and Engineering*, *11*(5), 4875–4885. https://doi.org/10.1109/TNSE.2024.3398219

Fan, Y., Lehmann, S., & Blok, A. (2022). Extracting the interdisciplinary specialty structures in social media data-based research: A clustering-based network approach. *Journal of Informetrics*, *16*(3), 101310. https://doi.org/10.1016/j.joi.2022.101310

Fortuna, P., & Nunes, S. (2019). A Survey on Automatic Detection of Hate Speech in Text. *ACM Computing Surveys*, *51*(4), 1–30. https://doi.org/10.1145/3232676

Garimella, K., De Francisci Morales, G., Gionis, A., & Mathioudakis, M. (2018). Political Discourse on Social Media: Echo Chambers, Gatekeepers, and the Price of Bipartisanship. *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18*, 913–922. https://doi.org/10.1145/3178876.3186139

Geller, M., Vasconcelos, V. V., & Pinheiro, F. L. (2023). Toxicity in Evolving Twitter Topics. In J. Mikyška, C. de Mulatier, M. Paszynski, V. V. Krzhizhanovskaya, J. J. Dongarra, & P. M. A. Sloot (Eds.), *Computational Science – ICCS 2023* (pp. 40–54). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-36027-5_4

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. (2013). Moral Foundations Theory. In *Advances in Experimental Social Psychology* (Vol. 47, pp. 55–130). Elsevier. https://doi.org/10.1016/B978-0-12-407236-7.00002-4

Graham, J., Haidt, J., Motyl, M., Meindl, P., Iskiwitch, C., & Mooijman, M. (2018). Moral foundations theory: On the advantages of moral pluralism over moral monism. In *Atlas of moral psychology* (pp. 211–222). The Guilford Press.

Guess, A. M., Malhotra, N., Pan, J., Barberá, P., Allcott, H., Brown, T., Crespo-Tenorio, A., Dimmery, D., Freelon, D., Gentzkow, M., González-Bailón, S., Kennedy, E., Kim, Y. M., Lazer, D., Moehler, D., Nyhan, B., Rivera, C. V., Settle, J., Thomas, D. R., … Tucker, J. A. (2023). How do social media feed algorithms affect attitudes and behavior in an election campaign? *Science*, *381*(6656), 398–404. https://doi.org/10.1126/science.abp9364

Guimarães, S. S., Reis, J. C. S., Ribeiro, F. N., & Benevenuto, F. (2020). Characterizing Toxicity on Facebook Comments in Brazil. *Proceedings of the Brazilian Symposium on Multimedia and the Web*, 253–260. https://doi.org/10.1145/3428658.3430974

Kalimeri, K., Beiro, M., Urbinati, A., Bonanomi, A., Rosino, A., & Cattuto, C. (2019). Human Values and Attitudes towards Vaccination in Social Media. *Companion Proceedings of The 2019 World Wide Web Conference*, 248–254. https://doi.org/10.1145/3308560.3316489

Kent, M. L., & Taylor, M. (2021). Fostering Dialogic Engagement: Toward an Architecture of Social Media for Social Change. *Social Media + Society*, *7*(1), 2056305120984462. https://doi.org/10.1177/2056305120984462

Krebs, D. L. (2022). 168Purity. In D. L. Krebs (Ed.), *Survival of the Virtuous: How We Became a Moral Animal* (p. 0). Oxford University Press. https://doi.org/10.1093/oso/9780197629482.003.0012

Leite, J. A., Silva, D., Bontcheva, K., & Scarton, C. (2020). Toxic Language Detection in Social Media for Brazilian Portuguese: New Dataset and Multilingual Analysis. In K.-F. Wong, K. Knight, & H. Wu (Eds.), *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing* (pp. 914–924). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.aacl-main.91

Matamoros-Fernández, A., & Farkas, J. (2021). Racism, Hate Speech, and Social Media: A Systematic Review and Critique. *Television & New Media*, *22*(2), 205–224. https://doi.org/10.1177/1527476420982230

McLoughlin, K. L., Brady, W. J., & Crockett, M. J. (2021). The role of moral outrage in the spread of misinformation. *TMS Proceedings 2021*. Technology, Mind, and Society, 2021. https://doi.org/10.1037/tms0000136

Miranda, S., Gouveia, C., Di Fátima, B., & Antunes, A. C. (2024). Hate speech on social media: Behaviour of Portuguese football fans on Facebook. *Soccer & Society*, *25*(1), 76–91. https://doi.org/10.1080/14660970.2023.2230452

Nanayakkara, A. C., Kumara, B. T. G. S., & Kapila Tharanga Rathnayaka, R. M. (2024). Cross-Platform Topic Analysis: Identifying Trends and Divergences in Social Media Discourse. *2024 8th SLAAI International Conference on Artificial Intelligence (SLAAI-ICAI)*, 1–6. https://doi.org/10.1109/SLAAI-ICAI63667.2024.10844977

Noor, N. B., Yousefi, N., Spann, B., & Agarwal, N. (2023). *Comparing Toxicity Across Social Media Platforms for COVID-19 Discourse* (No. arXiv:2302.14270). arXiv. https://doi.org/10.48550/arXiv.2302.14270

Papacharissi, Z. (2009). The virtual geographies of social networks: A comparative analysis of Facebook, LinkedIn and ASmallWorld. *New Media & Society*, *11*(1–2), 199–220. https://doi.org/10.1177/1461444808099577

Pérez, J. M., Rajngewerc, M., Giudici, J. C., Furman, D. A., Luque, F., Alemany, L. A., & Martínez, M. V. (2024). *pysentimiento: A Python Toolkit for Opinion Mining and Social NLP tasks* (No. arXiv:2106.09462). arXiv. https://doi.org/10.48550/arXiv.2106.09462

Recuero, R. (2024). The Platformization of Violence: Toward a Concept of Discursive Toxicity on Social Media. *Social Media + Society*, *10*(1), https://doi.org/10.1177/20563051231224264

Salawu, S., He, Y., & Lumsden, J. (2020). Approaches to Automated Detection of Cyberbullying: A Survey. *IEEE Transactions on Affective Computing*, *11*(1), 3–24. https://doi.org/10.1109/TAFFC.2017.2761757

Salehabadi, N., Groggel, A., Singhal, M., Roy, S. S., & Nilizadeh, S. (2022). *User Engagement and the Toxicity of Tweets* (No. arXiv:2211.03856). arXiv. https://doi.org/10.48550/arXiv.2211.03856

Santos, R. B., Matos, B. C., Carvalho, P., Batista, F., & Ribeiro, R. (2022). *Semi-supervised annotation of Portuguese hate speech across social media domains*. Schloss Dagstuhl-Leibniz-Zentrum fur Informatik GmbH, Dagstuhl Publishing. https://doi.org/10.4230/OASIcs.SLATE.2022.11

Sheth, A., Shalin, V. L., & Kursuncu, U. (2022). Defining and detecting toxicity on social media: Context and knowledge are key. *Neurocomputing*, *490*, 312–318. https://doi.org/10.1016/j.neucom.2021.11.095

Siegel, A. A. (2020). Online Hate Speech. In *Persily & J. A. Tucker (Eds.), Social Media and Democracy* (pp. 56–88). Cambridge: Cambridge University Press.

Valenzuela, S., Piña, M., & Ramírez, J. (2017). Behavioral Effects of Framing on Social Media Users: How Conflict, Economic, Human Interest, and Morality Frames Drive News Sharing. *Journal of Communication*, *67*(5), 803–826. https://doi.org/10.1111/jcom.12325

Van Bavel, J. J., Robertson, C. E., Del Rosario, K., Rasmussen, J., & Rathje, S. (2024). Social Media and Morality. *Annual Review of Psychology*, *75*(1), 311–340. https://doi.org/10.1146/annurev-psych-022123-110258

Vicario, M. D., Quattrociocchi, W., Scala, A., & Zollo, F. (2018). *Polarization and Fake News: Early Warning of Potential Misinformation Targets* (No. arXiv:1802.01400). arXiv. https://doi.org/10.48550/arXiv.1802.01400

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151. https://doi.org/10.1126/science.aap9559

Waterloo, S. F., Baumgartner, S. E., Peter, J., & Valkenburg, P. M. (2018). Norms of online expressions of emotion: Comparing Facebook, Twitter, Instagram, and WhatsApp. *New Media & Society*, *20*(5), 1813–1831. https://doi.org/10.1177/1461444817707349

Xu, Q. A., Jayne, C., & Chang, V. (2024). An emoji feature-incorporated multi-view deep learning for explainable sentiment classification of social media reviews. *Technological Forecasting and Social Change*, *202*, 123326. https://doi.org/10.1016/j.techfore.2024.123326

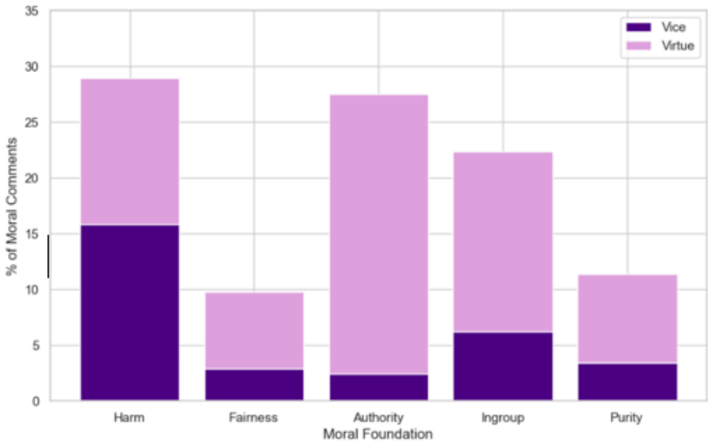
Yarchi, M., Baden, C., & Kligler-Vilenchik, N. (2021). Political Polarization on the Digital Sphere: A Cross-platform, Over-time Analysis of Interactional, Positional, and Affective Polarization on Social Media. *Political Communication*, *38*(1–2), 98–139. https://doi.org/10.1080/10584609.2020.1785067

Zúquete, M., Orghian, D., & Pinheiro, F. L. (2023). A Moral Foundations Dictionary for the European Portuguese Language: The Case of Portuguese Parliamentary Debates. In J.

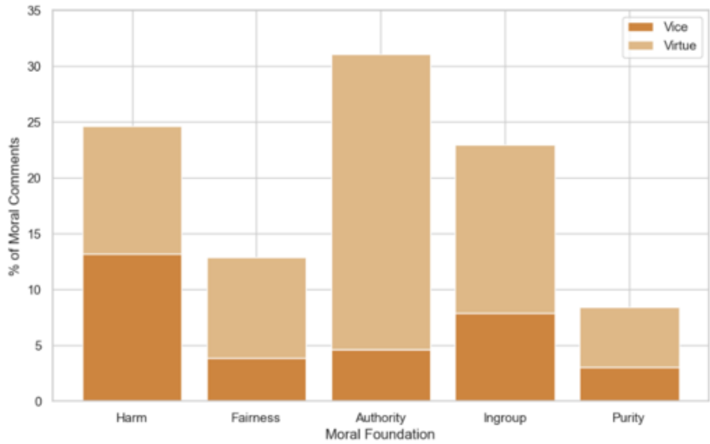
Mikyška, C. De Mulatier, M. Paszynski, V. V. Krzhizhanovskaya, J. J. Dongarra, & P. M. A. Sloot (Eds.), *Computational Science – ICCS 2023* (Vol. 14073, pp. 421–434). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-35995-8_30

APPENDIX A

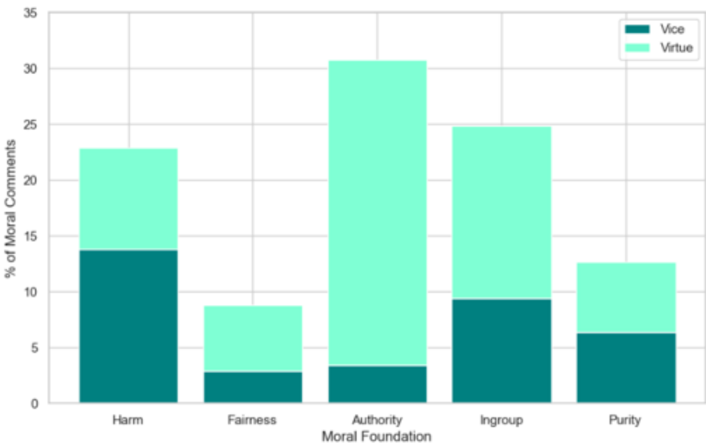
Instagram



Reddit



X



TikTok

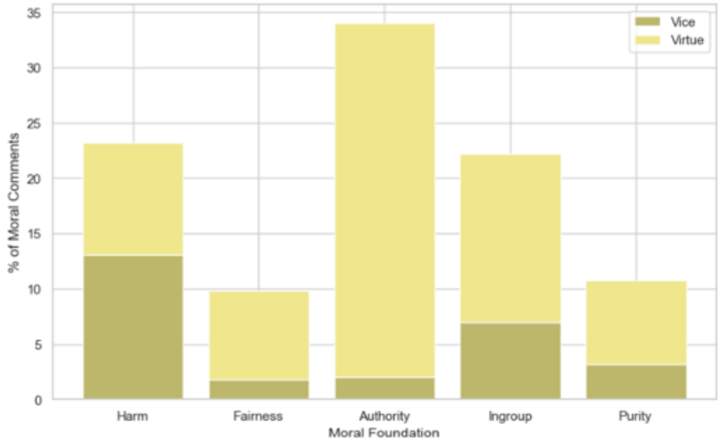


Figure A. 1 - Distribution of moral foundations among moral comments per platform

APPENDIX B

Table B. 1 - Levene's tests for homogeneity of variances across platforms for toxicity and sentiment scores

Variable	Levene's Statistic	df1	df2	p-value
Toxicity Score	2148.26	3	153044	<.001
Sentiment Score	924.08	3	153044	<.001
Moral Ratio (among comments with moral language)	250.37	3	38317	

Table B. 2 - Results of Mann–Whitney U Tests comparing sentiment scores between Moral and Non-Moral

Platform	U Statistic	p-value
Instagram	51727713	3.13e-246
Reddit	393,807,738	< .0001
X	107273726	3.66e-183
TikTok	6572762	3.67e-72

Table B. 3 - Results of Mann–Whitney U Tests comparing toxicity scores between Moral and Non-Moral comments per platform

Platform	U-Statistic	p-value
Instagram	89055133	< .0001
Reddit	518360358	< .0001
X	145269313	< .0001
TikTok	10995176	< .0001

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa