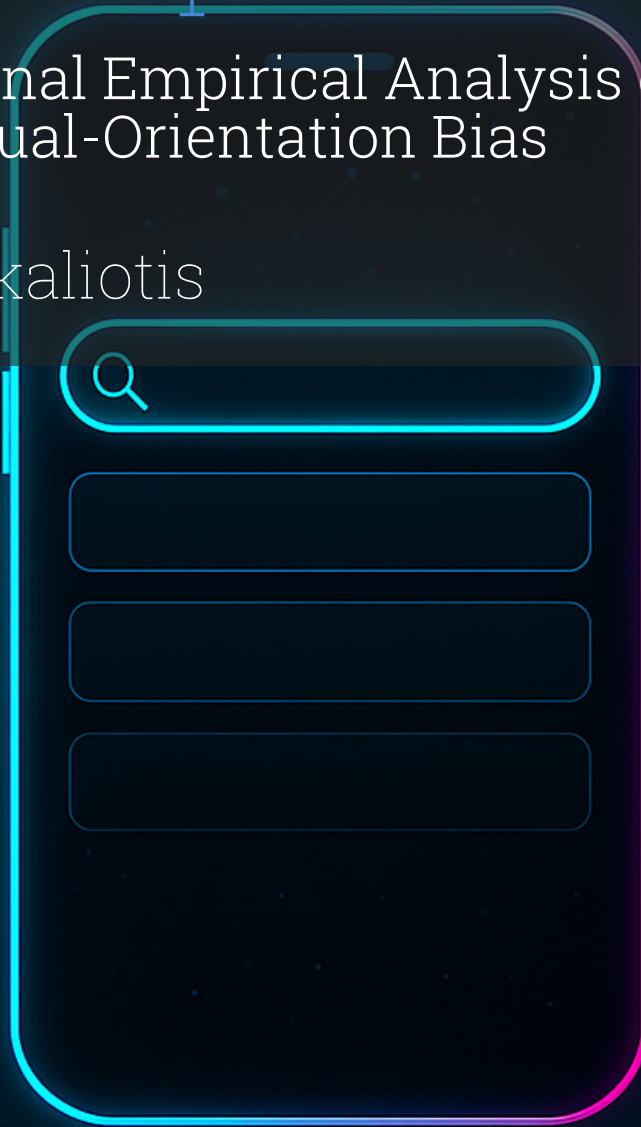


Investigating Toxic Language in TikTok's Autocomplete

A Cross-Regional Empirical Analysis of Gender, Race, and Sexual-Orientation Bias

Dimitrios Trikaliotis



Investigating Toxic Language in TikTok's Autocomplete

A Cross-Regional Empirical Analysis of Gender,
Race, and Sexual-Orientation Bias

By

Dimitrios Trikaliotis

to obtain the degree of **Master of Science**
in **Management of Technology**
at the Delft University of Technology,
to be defended publicly on Wednesday September 17, 2025 at 14:15 PM.

Student number:	6028268
Project duration:	February 3, 2025 – September 17, 2025
Thesis committee:	Prof. Dr. MJG (Michel) van Eeten, OG, TU Delft, Chairperson Dr. S. (Savvas) Zannettou, OG, TU Delft, 1st Supervisor Dr. SNR (Stefan) Buijsman, EPT, TU Delft, 2st Supervisor

Faculty of Technology, Policy and Management

An electronic version of this thesis is available at <https://repository.tudelft.nl/>



Acknowledgements

This thesis marks the conclusion of not just a research project, but also a meaningful chapter of my life, my journey through the MSc in Management of Technology at TU Delft. Looking back, it's clear that the completion of this work was only possible thanks to the support, guidance, and presence of many people along the way.

I am especially grateful to Dr. Savvas Zannettou, my daily supervisor, for his continuous support throughout this thesis. His availability, clear feedback, and thoughtful direction were indispensable, particularly during moments of uncertainty and doubt. I also extend my sincere thanks to Prof. Dr. Michel van Eeten, chair of my thesis committee, for pushing me to clarify my thinking and deepen the analytical rigor of this work. My deep appreciation also goes to Dr. Stefan Buijsman, whose broader philosophical and ethical reflections helped me consider the implications of algorithmic systems in a more nuanced way. Furthermore, I would like to thank the annotators who contributed their time and effort to evaluating the data, your input was critical to the success of this research.

I would like to thank TU Delft and the Management of Technology program for creating an academic environment where technical understanding, societal impact, and personal growth intersect. Over the past two years, I have been challenged to think critically about the role of technology in shaping our world and more importantly, about my role in that process. I am grateful for the opportunity to study in such a dynamic, interdisciplinary setting.

To my fellow MOT classmates and the friends I made throughout this journey, thank you for the laughs, the long hours in the library, the honest feedback, and the shared struggles that turned into some of my best memories. This experience would not have been the same without you.

Finally, I want to thank my family and close friends for their unconditional support throughout this master's. Your belief in me, even during the late nights and moments of stress, gave me the resilience to keep going. Your encouragement was often invisible, but always essential.

This thesis symbolizes the end of a demanding but deeply fulfilling period in my life. To everyone who shaped it in ways big or small: thank you!

Dimitrios Trikaliotis
Delft, September 2025

Executive Summary

Autocomplete feature may appear trivial, but on platforms like TikTok, where millions of users turn to search as a gateway to information, it plays a significant role in shaping what people ask and how they think. As the platform becomes increasingly central to digital culture and identity formation, the predictive suggestions it offers can reflect, reinforce, or challenge social biases. This thesis examines TikTok's autocomplete system through the lens of toxic language, analysing over 225,000 suggestions generated from more than 3,000 identity-related prompts across ten countries. The aim is to understand whether the system amplifies toxic stereotypes around gender, race, and sexual orientation, and whether these patterns vary by geography or ranking.

To approach this, the study conceptualises algorithmic bias as a form of representational distortion, when identities are disproportionately linked to hostile, stereotypical, or derogatory language. Toxicity is used as a measurable proxy for this phenomenon, with scores derived from Google's Perspective API and validated against a human-annotated subset. This framework allows the study to scale while maintaining conceptual clarity, offering a bridge between statistical insight and social meaning. The use of a cross-national dataset further adds depth, enabling the analysis to explore both universal and context-specific patterns of bias.

The findings show clear evidence that identity matters. Prompts referencing homosexual identities and Black identity terms consistently produced higher toxicity scores than their heterosexual and White counterparts. Gender-based disparities were also observed, though they were less pronounced. Interestingly, these patterns held steady across all ten countries studied. Despite cultural, regulatory, and linguistic differences, the toxicity distributions remained largely similar, suggesting that TikTok's autocomplete system likely runs on a globally standardised model that does not meaningfully adapt to regional contexts. In contrast, the ranking of suggestions, where a toxic output appears within the top eight, had only a marginal impact on overall exposure. While some toxic completions did surface in mid-list positions, their distribution lacked a clear or consistent pattern.

Taken together, the research offers an empirical audit of TikTok's search interface from a bias and fairness perspective. It shows that autocomplete, though often overlooked, can act as a subtle mechanism through which social hierarchies are reproduced. This insight carries implications for platform accountability and algorithmic design, especially in the context of ongoing policy efforts such as the EU AI Act, an emerging regulatory landscape that is likely to include systems like TikTok's autocomplete. By demonstrating that toxic associations are not isolated glitches but predictable patterns, the study highlights the limitations of moderation strategies that focus solely on removals or post hoc filtering. Ultimately, the thesis argues that mitigating algorithmic bias requires more than adjusting output rankings, it demands deeper attention to how predictive models are trained, evaluated, and governed. As TikTok continues to shape how a new generation accesses information, improving the social impact of features like autocomplete is not only a technical challenge but a public responsibility.

Contents

Acknowledgements	2
Executive Summary	3
1 Introduction	6
1.1 Framing Toxic Language and Study Rationale	7
1.2 Scope and Objectives	8
1.3 MOT relevance	10
2 Background and Related Work	11
2.1 Literature Review Methodology and Research Gap	11
2.2 Algorithmic Systems, TikTok Autocomplete and Stereotypes	12
2.4 YouTube's Autocomplete	14
2.5 Google's Autocomplete	14
3 Methodology	16
3.1 Case study design	16
3.2 Toxicity Assessment Models: Perspective API, ToxicBERT, and UnRoBERTa	17
3.3 Data Collection	20
3.4 Data Analysis	22
3.5 Pilot Validation Using Severity Ratings	23
3.6 Main Validation Using Probability Estimates	25
3.7 Practical Comparison of Toxicity Models	26
4 Results	29
4.1 Results overview	29
4.2 General Toxicity Trends	29
4.3 SRQ1: Toxicity by Identity Category	33
4.3.1 Distributional Patterns by Identity Category	34
4.3.2 Contextual Analysis by Declarative Themes	38
4.4 SRQ2: Cross-National Patterns	43
4.4.1 Distributional Patterns by Country	43
4.4.2 Robustness Checks Across Countries	45
4.5 SRQ3: Role of Ranking	46
4.5.1 Descriptive statistics	47
4.6 Summary of Key Findings	50
5 Discussion	52

5.1 Main Findings	52
5.2 Comparison with Existing Literature	53
5.3 Theoretical contribution.....	54
5.4 Policy implications	55
5.5 Limitations and Directions for Future Research	56
6 Conclusion	59
6.1 Addressing the research question	59
6.2 Academic relevance	60
6.3 Social relevance	60
Disclosure of AI Support	62
References.....	63
A Supplementary Data and Analyses	66
A1. Construction of Prefixes	66
A2. Prompts with 0 suggestions in all countries	67
A3. Frequency of Incomplete Autocomplete Results by Country	68
A4. Examples of Prompts with Incomplete Autocomplete Results Across Countries	68
A5. Top 5 Most Toxic Suggestions per Country	69
A6. Prompt–Opposite Toxicity Comparison Matrix	70
A7. Mean Toxicity by Prompt Form and Identity Subcategory	72
A8. Pairwise KS Test Results	72
A9. Boxplot: toxicity distribution by two-slot bins.....	73
B Evaluation of Toxicity Classification Models	74
B1. Human Annotation Guidelines – Severity Ratings (Used in Section 3.5).....	74
B2 Human Annotation Guidelines – Probability Estimates (Used in Section 3.6).....	75

1

Introduction

In recent years, social media platforms have become a central part of every day life, changing the way individuals access information, connect with others, and take part in public conversations (Harchekar, 2017). Among these platforms, the rapid rise of short-form video content has further accelerated this shift by offering users quick, easily consumable, and trend-driven media (Chen et al., 2023). This change has altered how users pay attention online, with short, attention-grabbing content now taking priority over more detailed or time-ordered formats.

At the forefront of this transformation stands TikTok, a platform that grew rapidly during the COVID-19 pandemic as users turned to online platforms to stay socially connected (Feldkamp, 2021). TikTok is unique in how it works: rather than centering user-to-user connections, it recommends content mainly through algorithms, especially via the “For You” feed. Scholars describe this as creating “algorithmized self,” where users see content that reflects and reinforces their preferences, habits, and identities (Bhandari & Bimo, 2022). This structure is different from traditional models of social media, such as Facebook or Instagram which focus more on personal connections. As such, TikTok blurs the boundaries between search engine, entertainment platform, and identity mirror, making it hard to fit into standard categories of digital media.

Building on this hybrid role, TikTok is not only a recommendation engine and entertainment platform, but is also increasingly used as a search tool. As younger users start using TikTok instead of Google to look up things like “where to eat,” “how to vote,” or “what is happening in Gaza” the platform’s search interface is becoming a vital entry point for knowledge access. Recent surveys indicate that up to 40% of Gen Z users now turn to TikTok or Instagram rather than Google for search-related tasks (Delouya, 2022). But unlike traditional search engines, TikTok’s search relies on a recommendation-driven system that does not have clear rules, editorial review, or a fixed way of organizing information. This evolving role makes it critical to understand how features like autocomplete influence what users see and how they explore information.

While existing research has explored how TikTok’s recommendation algorithms influence content exposure and user engagement (Chen et al., 2023; Zhang & Liu, 2021), the platform’s search interface, particularly its autocomplete function has received much less attention. Autocomplete refers to the predictive feature that generates suggested queries as users begin typing. These small suggestions can have a big impact on what users search, see, and think about. Research on other platforms, such as Google and YouTube has shown that not only helps users search faster but also influences how topics are framed by highlighting certain wordings or perspectives (Epstein et al., 2024b; Hiemstra, 2020). In this way, autocomplete suggestions act as discursive cues that both reflect and reinforce cultural biases, potentially influencing public opinion and making stereotypes seem more normal or acceptable.

These concerns are especially pressing given past controversies over autocomplete suggestions that have appeared as toxic, offensive, or discriminatory. Google, for instance, faced legal and regulatory pressure to modify or filter out harmful completions, such as racist or misogynistic queries, leading to policy changes and platform interventions (Hiemstra, 2020). In contrast, TikTok's response to this issue is still unclear. As TikTok becomes more of a tool for discovering information, understanding how autocomplete may contribute to toxic or biased narratives is not just important, it's a social and ethical responsibility.

1.1 Framing Toxic Language and Study Rationale

As a first step toward unpacking those risks, this chapter frames toxic language as the main lens for analysing autocomplete's social and ethical impact. Drawing on Gevers et al. (2022), toxic language is an inclusive term that covers abusive or offensive speech, hate speech, cyberbullying, threats, defamation, and even language that may not be illegal but is still insulting or socially harmful. In this thesis, toxicity refers to how much or how strongly such language appears in a given text. However, toxicity is not always easy to define, it depends heavily on context. What sounds offensive to one person might seem harmless or even normal to another, depending on personal experiences, cultural background, exposure to discrimination, and individual sensitivity.

This subjectivity becomes important when toxic language shows up in algorithmic systems like autocomplete. These systems don't just reflect language, they learn from patterns in data, which can include biased or harmful associations. A suggested query might seem neutral on its own but take a toxic meaning when associated with sensitive identity categories such as race, gender, or sexuality. For instance, autocomplete completions to prefixes like "why are women..." or "why are Muslims..." have, in various studies, returned suggestions that reinforce sexist or Islamophobic stereotypes (Houli et al., 2021). Moreover, repeated exposure to such prompts can gradually make biased ideas feel normal and influence social attitudes over time, making their overall impact more serious than a single encounter. It's important to remember that the idea of toxic speech is still evolving and that there is often disagreement about what counts as toxic speech. Definitions of harm are historically dynamic, they change over time in response to legal frameworks, cultural values, and collective social awareness. Language that once seemed normal or widely accepted may later be reclassified as harmful or exclusionary, as evidenced by the gradual delegitimization of gendered insults and ethnic slurs once tolerated in mainstream media (MacIntyre, 2003; Sellars, 2016). These changes highlight that toxicity is not just a linguistic issue but also a shifting concept shaped by moral, legal, and political change.

To address these complexities, this thesis uses a grounded and socially informed definition of toxicity. Rather than focusing only on explicit hate speech, we define toxic language on autocomplete suggestions as this that trigger stereotypes, undermine certain identities, or subtly support discriminatory ideas. This broader definition aligns with Noble (2018) concept of algorithmic oppression, which highlights how algorithmic systems can amplify existing inequalities through biased data and interface design.

Within this framework, toxicity is not treated as synonymous with bias. Instead, it acts as a sign that certain identities may often be linked with hostility, ridicule, or moral judgment in algorithmic outputs. When autocomplete suggestions repeatedly return toxic results for certain identity groups, they give us a way to study larger patterns of bias. To measure toxicity in a way that is both scalable and easy to understand, we use automated models that estimate how offensive a piece of text might seem. We are aware of the limits of these tools, especially when it comes to subjectivity, cultural context, and more subtle forms of harm (Goyal et al., 2022). Rather than treating these scores as exact or final, we use them as clues within a broader system, one shaped

by the data used to train the algorithms, how users interact with the platform, and the platform's design choices. This helps us study how these systems might treat different identities unequally. Seeing these patterns also raises the bigger question of how such technologies should be managed or regulated.

These concerns about bias and unequal representation also tie into wider conversations about platform responsibility and AI governance. Policies like the European Union's upcoming AI Act (Council, 2023) and ethical frameworks such as the OECD Principles on AI (OECD, 2019) emphasize transparency, fairness, and harm mitigation, principles especially relevant to recommender systems and search functionalities. TikTok's autocomplete feature, however, operates as a black box. Users don't know how suggestions are created, ranked, or filtered. This lack of openness raises concerns about how users might be influenced without realizing it, how biases might be repeated, and how people could lose the ability to make informed choices.

In response to these issues, our study examines how often TikTok's autocomplete returns toxic suggestions for identity-related prompts, and what those results reveal about bias. Toxicity serves as a workable lens because abusive or offensive language often makes it clear when stereotypes are involved, and earlier studies of search and recommendation systems have found that rises in toxicity often appear alongside the marginalization of certain groups (Davidson et al., 2017; Sap et al., 2019). While bias can show up in different ways, toxicity is a useful and measurable part of it. It gives us a consistent way to detect and compare harmful patterns across a large set of autocomplete results. By tracing these patterns, the study also contributes to wider debates about the societal impact of AI systems. Autocomplete is not just a functional tool, it subtly shapes how people search, frames which ideas feel "normal," and can reproduce existing social hierarchies. Understanding how this system works helps inform both academic conversations in digital media and ongoing policy discussions around ethical AI and responsible platform design. Even small design choices in search interfaces can affect how people think about identity, harm, and truth, making this type of research vital for both scholarship and public accountability.

1.2 Scope and Objectives

To study these patterns, our research takes a comparative, identity-based approach. We focus on three key dimensions gender, race, and sexuality, not only because they are historically linked to discrimination, but also because they cover a broad range of social and cultural sensitivities. By collecting autocomplete suggestions from multiple countries, we aim to identify both global trends and culturally specific patterns in how toxicity shows up.

We also look at the role of ranking, that is, whether more toxic suggestions tend to appear near the top of the list, making them more likely to influence user behavior. This combined focus on identity, geography, and position within the interface gives us a structured way to evaluate how widespread toxic autocomplete suggestions are on TikTok. By analyzing how toxic language varies across identity groups, how its frequency changes between countries, and how its position affects visibility, the study breaks down the central research question into measurable parts. This helps us better understand where toxic suggestions are most common and how likely users are to encounter them.

Based on these concerns, the main research (RQ) question guiding this study is:

RQ: What is the prevalence of toxic autocomplete suggestions on TikTok?

To delve deeper, the following sub- research questions (SRQ) are proposed:

- SRQ1: To what extent do autocomplete suggestions in TikTok reproduce toxic content

related to gender, race, or sexual orientation?

- SRQ2: How do TikTok’s autocomplete suggestions differ when accessed by users in different geographic locations?
- SRQ3: What role does ranking play in toxic autocomplete suggestions?

This study focuses on English-language content to improve consistency and ensure that results can be compared more reliably. Multilingual analysis often introduces added complexity due to cultural nuance and differences in how language is used, which can make interpretation less precise. By narrowing the scope to English, we are able to use a well-established set of tools and resources, such as validated toxicity and sentiment lexicons, annotation guidelines, and benchmark datasets that are not yet widely available or consistent in other languages (Davidson et al., 2017). To still reflect a range of social and cultural perspectives within the English-speaking world, we include representative countries from different global regions:

- Asia: Pakistan, Singapore
- Africa: Nigeria, Ghana
- Europe: UK, Ireland
- Americas: US, Trinidad and Tobago
- Oceania: Australia, New Zealand

These countries were selected not only because English is an official or primary language, but also to reflect a strategic diversity across geopolitical regions, colonial histories, religious backgrounds, levels of economic development, and norms around online communication. For example, Singapore and Ireland represent high-income but culturally distinct settings; Nigeria and Pakistan bring in postcolonial, multilingual societies; and Trinidad and Tobago and Ghana allow us to include Afro-Caribbean and West African perspectives, which are often missing in research on algorithmic bias. Meanwhile, the US, UK, and Australia serve as established reference points, particularly in discussions of tech policy and content moderation. Together, these countries provide a wide range of digital environments where English is used as a shared language, but with meaningful differences in how identity, harm, and discourse are understood.

In autocomplete systems, user attention typically focuses on the first few suggestions, making ranking a key factor in how toxic content spreads and how users experience the platform (Wang et al., 2018). By analyzing toxicity scores from the top-ranked suggestion down to lower-ranked ones, we can see whether the system tends to push harmful completions to the top, or whether it keeps them lower in the list, where they are less likely to be seen. This ranking analysis helps us understand real-world exposure: if toxic suggestions appear early, even a small number can have a large impact. But if they are only shown further down, their effect depends more on whether users scroll. Studying this ranking dynamic also gives insight into how well the system filters content. A well-designed platform would rank toxic suggestions lower, limiting exposure. By mapping toxicity across all ranked positions, we can assess whether TikTok’s autocomplete design helps reduce harmful language, or unintentionally boosts it by ranking it too highly.

This investigation contributes to ongoing debates about platform accountability, algorithmic transparency, and the social impact of AI-driven content systems. By systematically measuring toxicity across different identities, countries, and interface positions, this study provides a detailed picture of how predictive algorithms influence both user experience and public discourse.

The chapters that follow present the dataset, methods, and findings in depth, building a clearer understanding of where, how, and for whom toxic autocomplete suggestions appear on TikTok.

Chapter 2 reviews existing research on algorithmic systems, autocomplete functionality, and toxic language, including cross-platform comparisons. Chapter 3 explains the methodological approach, covering the case study design, data collection process, analysis techniques, and how classification tools were validated. Chapter 4 presents the results for each sub-research question. Chapter 5 discusses these findings in relation to earlier studies, theoretical frameworks, and platform policies, while also acknowledging the study's limitations and suggesting areas for future research. Chapter 6 concludes by summarizing the main insights and reflecting on their academic and societal significance.

1.3 MOT relevance

This thesis aligns closely with the interdisciplinary goals of the Master of Science in Management of Technology by bridging technical inquiry with societal impact. The project applies computational tools, such as machine learning-based toxicity scoring and cross-national data analysis, not simply for technical benchmarking, but to explore broader questions of equity, accountability, and digital ethics. In doing so, it demonstrates that technological systems must be studied not only for their performance or efficiency, but also in relation to the social contexts they influence and are influenced by.

At its core, the study reflects the MOT philosophy: combining technological innovation with careful consideration of its social impact. By examining TikTok's search infrastructure as a sociotechnical system, the research illustrates how platform design decisions can shape social discussion and identity representation in powerful ways. This aligns with the program's emphasis on the governance of emerging technologies, responsible AI, and platform accountability. Furthermore, the methodological combination of data analytics with qualitative validation mirrors the MOT focus on evidence-based decision-making in complex technological environments. In sum, the thesis embodies the MOT approach by treating technology not as a neutral tool, but as a dynamic force that must be understood, evaluated, and shaped in light of societal needs and ethical responsibility.

Background and Related Work

2.1 Literature Review Methodology and Research Gap

To conduct the literature review, Google Scholar was used as the primary search engine. Preliminary experiments were conducted to determine the most effective keywords. The goal was to locate existing literature on autocomplete functionality in well-established platforms such as Google and YouTube, as well as studies addressing social-media autocomplete toxicity, and specifically to target research examining TikTok’s autocomplete feature as a distinct phenomenon. It was found that combining the keywords “autocomplete,” the platform name, and “toxicity” produced the most relevant results, except for TikTok, where adding the word “toxicity” returned too few results to analyze, the final review was based on four structured searches: “autocomplete toxicity Google,” “autocomplete toxicity YouTube,” “autocomplete toxicity social media,” and “autocomplete TikTok.” To avoid missing broader but relevant literature, a general search using “autocomplete bias” was also conducted. This allowed for the inclusion of studies that may not frame their findings through the lens of toxicity specifically, but still address algorithmic harm, discrimination, or unfair suggestion patterns. To complement these platform-focused searches, an additional query using terms like “autocomplete toxicity ethics” and “autocomplete algorithm regulation” was also conducted. This was intended to capture literature that addressed broader ethical or policy discussions around autocomplete and algorithmic suggestion systems. We initially reviewed the first five pages (50 results) of each search. If two full consecutive pages (20 results) yielded no articles that included terms such as “bias,” “toxicity,” “harm,” or “algorithmic fairness” in the title, abstract, or conclusion, we ended the search. This cutoff was based on pilot tests, which showed that results became increasingly off-topic after this point. Each remaining paper was retained if it (1) focused on autocomplete or predictive text systems, (2) discussed bias, toxicity regulatory landscape or ethical concerns in algorithmic outputs, and (3) referred to empirical or technical methods. This process resulted in an initial pool of 150 papers. From these, 32 were selected for in-depth analysis based on their alignment with the research objectives. Selected studies offered substantive engagement with identity-related harms (e.g., race, gender, sexual orientation), employed empirical or model-driven approaches to measuring toxicity, or analyzed autocomplete outputs as a user-facing feature. Studies were excluded if they referred to autocomplete only in passing, lacked methodological detail, or focused broadly on moderation without addressing algorithmic suggestion systems.

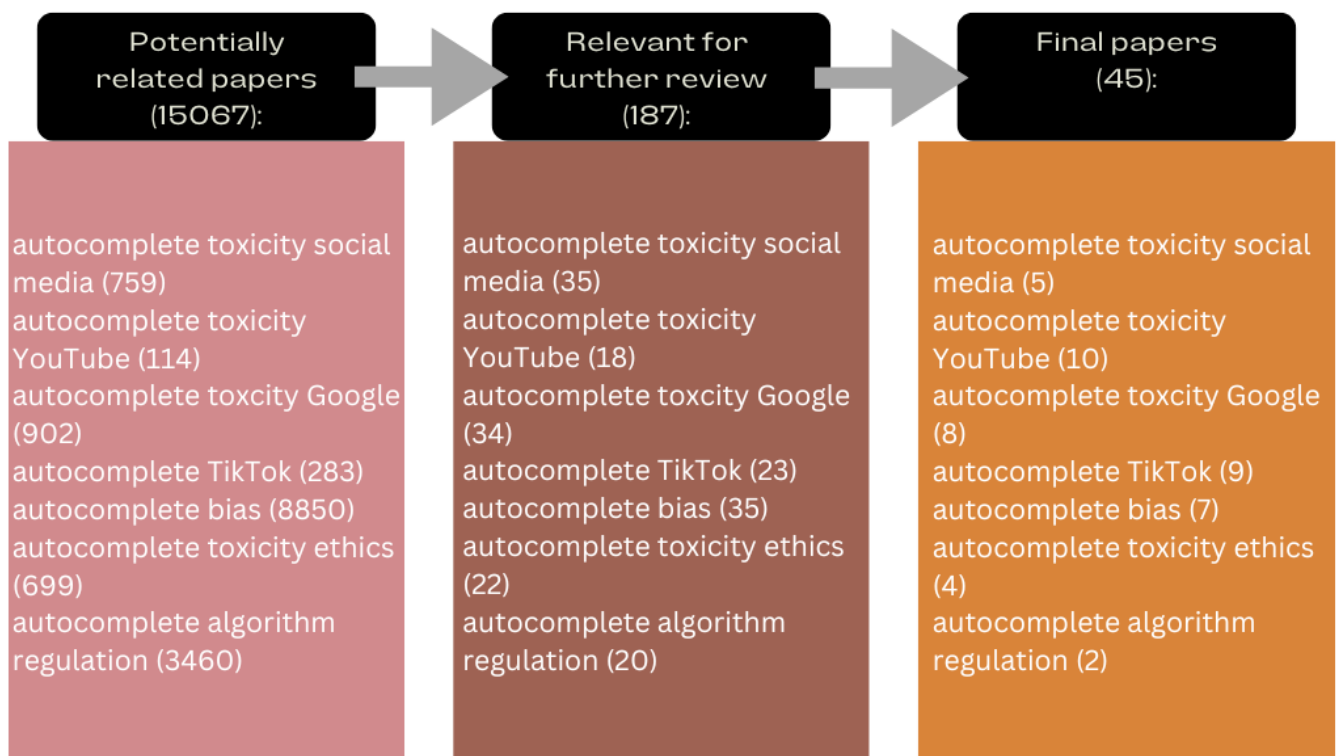


Figure 2.1: The selection process for the final papers

Prior research has focused extensively on autocomplete features in traditional search platforms. Studies on Google and YouTube have examined how autocomplete can reproduce stereotypes, amplify misinformation, and influence public opinion through biased or toxic suggestions (Baker & Potts, 2013; Epstein et al., 2024a; Hiemstra, 2020). Autocomplete has been found to reflect racial, gendered, and political bias, leading some platforms to implement moderation mechanisms under public or regulatory pressure (Graham, 2023; Leidinger & Rogers, 2023). Other work has framed autocomplete as a sociotechnical system that not only reflects but also shapes discourse norms (Azzopardi, 2021; Hazen et al., 2020). Research on YouTube, for instance, has shown that autocomplete can steer users toward stereotype-reinforcing or extremist content depending on ranking and prompt structure (Ha et al., 2024; Hussein et al., 2020). However, our review reveals a clear gap in the literature regarding TikTok. While prior research has focused on major search engines and video platforms, no empirical study has examined how TikTok’s autocomplete operates, or how it might surface toxic language during the search process.

This study contributes in two key ways. First, it addresses this empirical gap by offering the first systematic analysis of TikTok’s autocomplete toxicity. Second, it expands the existing focus on Western platforms by introducing a cross-national comparison across ten culturally and politically distinct English-speaking countries. These additions help build a broader understanding of algorithmic toxicity in rapidly evolving, non-traditional search contexts.

2.2 Algorithmic Systems, TikTok Autocomplete and Stereotypes

Algorithmic systems now shape the digital information environment, influencing not only what users see but how they make sense of the world (Hazen et al., 2020). Trained on historical data, search logs, demographics, engagement metrics, machine-learning models predict what users will want to do next (Buolamwini & Gebru, 2018). While such personalization streamlines discovery, it can also surface toxic language or amplify entrenched social biases when those biases are

encoded in the underlying data (Noble, 2018). Building on the notion of the algorithmic self (Bhandari & Bimo, 2022), TikTok continuously reflects users' identities and behaviours back to them, creating a feedback loop that narrows the range of content encountered and risks reinforcing stereotypes. Researchers warn that these echo-chamber dynamics allow discriminatory ideas to flourish whenever models learn from racist, sexist, or otherwise prejudiced patterns (Houli et al., 2021).

One of the most visible examples of algorithmic prediction is autocomplete. The feature is designed to speed up query entry (Ward et al., 2012) but can redirect users toward offensive, stereotypical, or misleading content when its suggestions reflect toxic framing (Hazen et al., 2020). Figure 2.2 shows a typical example: typing “why are women...” in TikTok surfaced completions such as “why are women always cold” and “why are women crazy,” both of which perpetuate gender stereotypes.

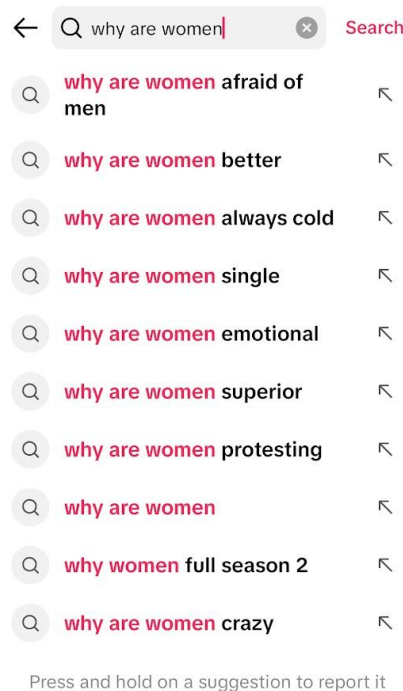


Figure 2.2: Example of TikTok’s Autocomplete Interface

Although autocomplete function has been studied on search engines like Google, TikTok’s implementation is comparatively under-researched. As noted in Chapter 1, younger users increasingly treat TikTok as a search interface, yet little is known about how its autocomplete system filters and ranks suggestions. TikTok’s “For You Page” (FYP) already prioritises recommended videos over social-network ties (Simpson & Semaan, 2021); its search module similarly relies on engagement signals and historical interaction data to predict query completions (Chen & Shi, 2022). If that data encodes harmful stereotypes, autocomplete may embed toxic framings before a user even finishes typing.

Determining whether TikTok’s autocomplete reproduces toxic content is therefore essential for assessing the platform’s broader influence on user perceptions, identity representation, and public discourse. The next section outlines the study design used to measure the prevalence, ranking, and geographic variation of toxic autocomplete suggestions, and to examine whether the feature contributes to algorithmic oppression (Noble, 2018) by consistently linking identity categories with hostile or offensive language.

This connection between prediction and social bias becomes clearer when viewed through the lens of stereotypes. Stereotypes are “beliefs about the characteristics of groups of individuals” (Stangor

& Schaller, 2000) and do not disappear in digital systems; instead they are re-encoded when identity-linked suggestions repeatedly mirror dominant narratives. TikTok's autocomplete, trained on historical engagement data, projects those prior patterns into the present. In this research we treat bias not as a random statistical error but as a representational distortion: certain identities (e.g., gender, race, sexuality) are disproportionately paired with reductive or hostile completions, reproducing structural inequalities embedded in training data, user interactions, or moderation policies (Blodgett et al., 2020).

To detect these patterns, we use toxicity as a proxy for stereotype activation. When prompts like "Why are [group] so ..." reliably return hostile or mocking suggestions, the pattern is unlikely to be accidental; rather, it exposes discriminatory residue in the underlying corpus (Houli et al., 2021). In this sense, autocomplete doesn't just reflect user behavior, it anticipates intent along socially encoded defaults and nudges attention toward stereotype-confirming framings. Measuring toxicity scores across suggestions therefore provides a scalable indicator of linguistic harm and helps identify where representational bias is most active.

2.4 YouTube's Autocomplete

YouTube's autocomplete feature has attracted growing attention due to concerns that it reinforces toxic suggestions and spreads misinformation. One audit found that query suggestions can reinforce entrenched racial stereotypes: for example, by associating Black individuals with negative traits such as poor hygiene or by repeating the "model minority" trope for Asian communities (Ha et al., 2024). These examples show how algorithms, trained on large-scale user data, can unintentionally reproduce societal biases and steer users toward toxic content, even before a search is completed.

Moreover, personalization appears to reinforce these issues. Research indicates that while initial autocomplete suggestions are only mildly shaped by demographic variables like age, gender, or location, a user's watch history quickly becomes a major influence (Hussein et al., 2020). This dynamic can steer viewers into narrow information bubbles, especially when reinforced by YouTube's broader recommendation system (Brown et al., 2022). The resulting feedback cycle may amplify ideologically charged or toxic-laden content, reducing exposure to diverse perspectives. Even as YouTube implements policy updates and content moderation guidelines (YouTube, n.d.) to curb hateful or misleading material, autocomplete remains a critical entry point through which hateful ideas can gain traction. These issues raise broader concerns about the societal impact of automated suggestions: if the initial layer of search guidance is skewed, the entire discovery pathway may be compromised from the start.

2.5 Google's Autocomplete

Compared to YouTube, Google's autocomplete has been studied more extensively, offering a wealth of insights into how search engines can reinforce stereotypes. Investigations reveal that even subtle variations in suggested queries, like excluding negative terms for certain political candidates or emphasizing them for others, may shape public perception in ways users barely notice (Epstein & Robertson, 2015). Although Google employs a combination of algorithmic updates and moderated "patches" to filter openly offensive terms, studies continue to document inconsistencies across regions and categories (Leidinger & Rogers, 2023). For instance, search prompts related to immigration in the United States often default to "illegal immigrant," potentially promoting a stigmatizing perspective, while other countries have more nuanced or varied suggestions (Muñoz-Bates, 2015). Such examples reflect deeper structural biases in autocomplete systems. Robertson et al. (2019) demonstrated this by repeatedly analyzing the algorithm's

outputs, showing how Google and Bing suggestion networks encode and reinforce misleading associations across politically sensitive topics. In addition, spikes in negative query completions following polarizing news events may unintentionally amplify hostility or resentment (Hazen et al., 2020).

Stereotypical associations are not limited to race and nationality. Research indicates that gender toxicity persists, with women's names frequently linked to marital status while men's queries more often foreground professional achievements (Graham, 2023). Similarly, ageism can surface through search predictions that frame older adults in predominantly negative or patronizing terms (Roy & Ayalon, 2020). These patterns illustrate how autocomplete can act as a mechanism of "incidental learning," reinforcing harmful beliefs by repeatedly surfacing skewed or pejorative suggestions (Baker & Potts, 2013). Despite the sophistication of Google's moderation strategies, scholars argue that the vast amount of user-driven data and the complexity of algorithmic models means that toxic content often manifests faster than they can be mitigated (Azzopardi, 2021).

Significantly, both YouTube and Google's documented challenges shed light on a critical gap surrounding TikTok's emerging role as a de facto search platform, particularly among younger demographics (Delouya, 2022). While research on TikTok's recommendation feeds is gaining momentum, its autocomplete mechanism remains underexamined, despite the potential for similarly toxic suggestions to influence users at the earliest stage of inquiry. Given the patterns seen across other major platforms, it is plausible that TikTok's autocomplete might also reflect societal prejudices rooted in its training data, yet the academic conversation has not yet fully turned to this possibility. The potential consequences are significant, especially as TikTok becomes a go-to source for news and information, any toxic autocomplete could shape perceptions among a highly impressionable audience. Understanding how these effects unfold and whether TikTok's developers address them remains an open question, one that becomes even more pressing considering the platform's rapid adoption and global reach.

3

Methodology

3.1 Case study design

TikTok has experienced rapid global adoption, emerging as a major platform for both entertainment and finding information (Huang, 2024). Platform’s autocomplete is our window into the platform’s hidden search logic. The feature is central to how millions now begin a query, yet the rules guiding its suggestions are opaque. By analyzing the language those suggestions contain, especially signs of toxicity, we can see whether the system directs users toward harmful stereotypes and, in doing so, illustrate wider questions of algorithmic steering on social platforms. As introduced in Section 1.2, this study adopts a comparative, identity-based approach, focusing on gender, race, and sexual orientation. To support this selection with a methodologically grounded rationale, we reviewed prior studies on search engines and YouTube autocomplete (Table 3.1) to identify the most commonly analyzed identity-related categories. This review confirmed that gender, race, and political affiliation were the most frequently studied dimensions.

However, a closer inspection of political bias research showed that it predominantly revolves around specific political figures, elections, or regional events, such as studies analyzing Angela Merkel (Bonart et al., 2020) or U.S. elections (Epstein et al., 2024a), rather than broader identity-based biases. Since our study aims to examine generalizable patterns across multiple aspects of identity rather than single political events, we opted to exclude political affiliation from our selection.

Instead, the study incorporates sexual orientation, which emerged as the next most frequently studied dimension and remains a highly relevant and evolving topic, particularly among younger generations (Simpson & Semaan, 2021). This selection ensures that our research captures identity-related biases that are both well-established in prior work and socially significant in contemporary discourse. Thus, we focus on gender, race, and sexual orientation as the key dimensions in our study, aligning with both existing literature and the broader implications of algorithmic bias in autocomplete systems.

Table 3.1: Most common identity dimensions in autocomplete bias studies

References	Gender	Race	Sexual Orientation	Profession	Religion	Political	Age	Communities	Migrants
Lin et al., 2023	x	x	x						
Nadeem et al., 2020	x	x		x	x				
Ha et al., 2024		x							
Kulshrestha et al., 2019						x			
Al-Abbas et al., 2020	x								
Baker & Potts, 2013	x	x	x						
Roy & Ayalon, 2020	x						x		
Bonart et al., 2020						x			
Leidinger & Rogers, 2023	x		x		x	x	x	x	
Lin et al., 2024	x								x
Muñoz-Bates, 2015									x
Epstein et al., 2024						x			
Epstein & Robertson, 2015						x			
Frequency Count:	7	4	3	1	2	5	2	1	2

Furthermore, as noted in Section 1.2, this study also incorporates a geographic comparison to examine how autocomplete suggestions vary across English-speaking countries. This approach allows for a more nuanced view of how language, culture, and local usage patterns might influence algorithmic outputs. By selecting countries from multiple world regions, our design captures cultural and digital diversity within a consistent linguistic framework. In this sense, TikTok’s autocomplete feature serves as a pertinent case study for exploring the interplay between platform design, societal toxicity, and user engagement.

3.2 Toxicity Assessment Models: Perspective API, ToxicBERT, and UnRoBERTa

This study uses three machine learning models to assess the toxicity of autocomplete suggestions: the Perspective API, ToxicBERT, and UnRoBERTa. All three models used in this study are either developed by or trained on datasets created by Jigsaw, a technology incubator within Google that specializes in online safety and content moderation tools. Jigsaw created both the Perspective API and the widely used Toxic Comment Classification Challenge datasets (2018 and 2019), which have become standard benchmarks for training and evaluating toxicity detection models. While Perspective API represents Jigsaw’s production-ready model, tools like ToxicBERT and were developed independently by Unitary, a company focused on building AI tools for detecting harmful online content and were fine-tuned on Jigsaw’s labeled datasets to support academic and fairness-focused research. The key features of the three models used in this study are summarised in Table 3.2.

Table 3.2: Overview of Toxicity Models Used in This Study

Feature	Perspective API	ToxicBERT	UnRoBERTa
Developer	Jigsaw (Google)	Unitary	Unitary
Model Type	Custom machine learning model (not open-sourced)	Built on BERT , a popular language model that reads text in both directions	Built on RoBERTa, a faster, more robust version of BERT
Training Dataset	Jigsaw internal data (Wikipedia talk pages, news comments), annotated for perceived toxicity	Jigsaw 2018 Toxic Comment Classification (Wikipedia comments)	Jigsaw 2019 Unintended Bias in Toxicity (Civil Comments with identity labels)
Scoring	Probabilistic (0–1)	Probabilistic (0–1)	Probabilistic (0–1)

Use Case Strength	General-purpose content moderation and large-scale audits	Strong in detecting explicit toxicity (e.g., slurs, threats)	Bias-aware; distinguishes toxicity from neutral identity mentions
Accessible via API	Public API (free for research use)	via Detoxify library	via Detoxify library

The Perspective API is the most widely used of the three. It is a machine learning–based text analysis service designed to detect harmful language in online environments. It has been adopted in academic research, journalism, and platform governance to audit toxicity, harassment, and bias in user-generated content.

Technically, the Perspective API is a supervised learning model trained on large, crowd-annotated datasets, including Wikipedia talk pages and news comment sections. Human annotators label each comment for perceived toxicity, and the model learns to predict future toxicity judgments based on these inputs. It outputs a probability score from 0 (very unlikely) to 1 (very likely) that a typical person would perceive the input as toxic (Figure 3.1). These scores are used in this study to assess autocomplete suggestions, treating higher probabilities as stronger indicators of potential stereotype activation.

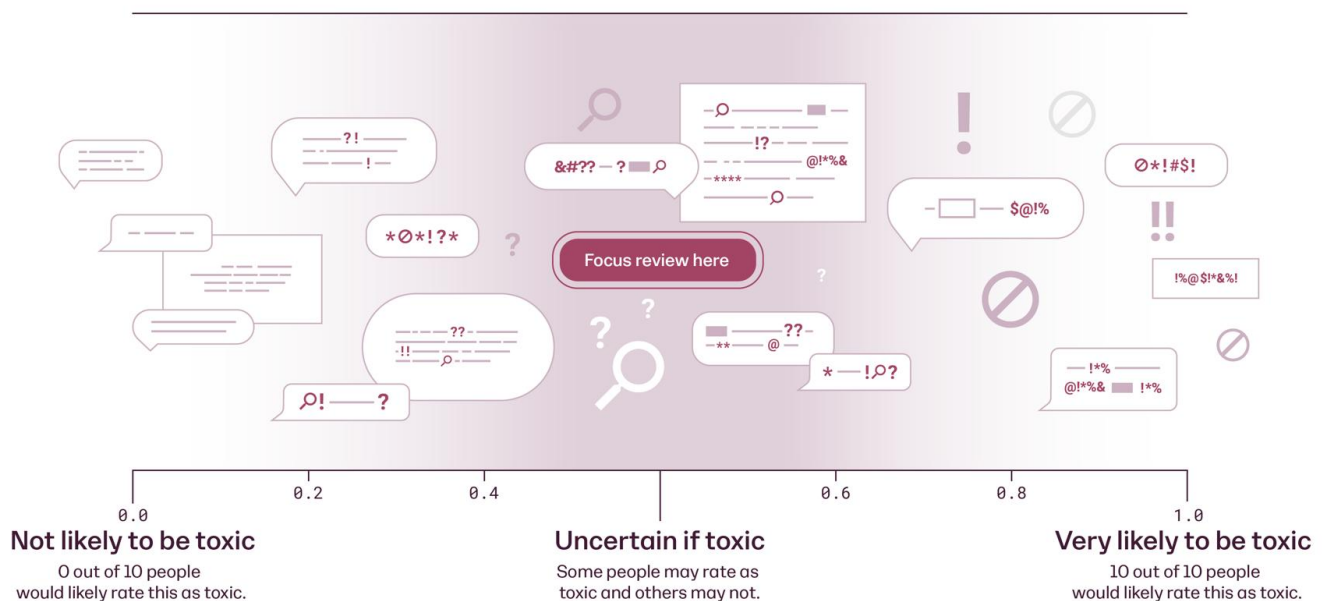


Figure 3.1: Toxicity scoring scale used by the Perspective API.

Furthermore, the Perspective API provides multiple toxicity dimensions (e.g., general toxicity, severe toxicity, identity attack, threat), but this study focuses on general toxicity, which has shown the most reliable performance across topics. In a validation step, general toxicity scores aligned more closely with human ratings than other dimensions and captured subtle harms such as sarcasm or coded language that may not involve direct threats or slurs.

Conceptually, general toxicity offers a broad, interpretable measure of perceived offensiveness that captures a range of negative framing strategies, sarcasm, ridicule, and coded language, that may not trigger narrower categories like threat or identity attack. It allows for scalable detection of problematic language without relying exclusively on overt slurs or aggressive phrasing. This generality is particularly important for studying stereotype reinforcement in autocomplete, where suggestions may encode bias without using explicitly abusive terms.

To complement Perspective, we also use two models from the open-source Detoxify library: ToxicBERT and UnRoBERTa (He et al., 2024). Both are built on BERT (Bidirectional Encoder Representations from Transformers) is a deep neural language model developed by Google that learns the meaning of words based on their context, using a transformer architecture that processes entire sequences bidirectionally (Devlin et al., 2019) but trained on different datasets and reflect distinct approaches to classifying toxicity.

ToxicBERT is based on the original BERT model and fine-tuned on the 2018 Jigsaw Toxic Comment Classification dataset, which includes online comments focused on overtly toxic content, such as profanity, slurs, insults, and threats. It produces a probability score between 0 and 1, estimating how likely a reader is to perceive the content as toxic. ToxicBERT is known for aggressive scoring, flagging a broad range of content as harmful, especially when surface-level abusive language is present. This makes it effective at detecting explicit toxicity but prone to false positives in ambiguous or context-sensitive cases.

UnRoBERTa, by contrast, uses a RoBERTa architecture and is fine-tuned on the 2019 Jigsaw Unintended Bias in Toxicity dataset, which includes identity-based annotations to reduce false positives on references to race, gender, sexuality, and religion. Like the other models, it also produces probability scores from 0 to 1. UnRoBERTa is designed to minimize bias in classification by distinguishing between genuinely harmful content and neutral identity references. As a result, it tends to be more conservative in its scoring, especially for identity-related prompts, making it well suited for fairness auditing.

Although all three models produce probability scores on the same 0–1 scale, their behavior differs due to their training focus. ToxicBERT captures explicit toxicity, Perspective API captures a broad spectrum including implicit harm, and UnRoBERTa filters for fairness around identity terms. This combination allows the study to assess autocomplete toxicity from multiple angles, revealing how different models might flag the same suggestion for different reasons or miss certain patterns altogether. The following Venn diagram (Figure 3.2) summarizes the key similarities and distinctions among the three models.

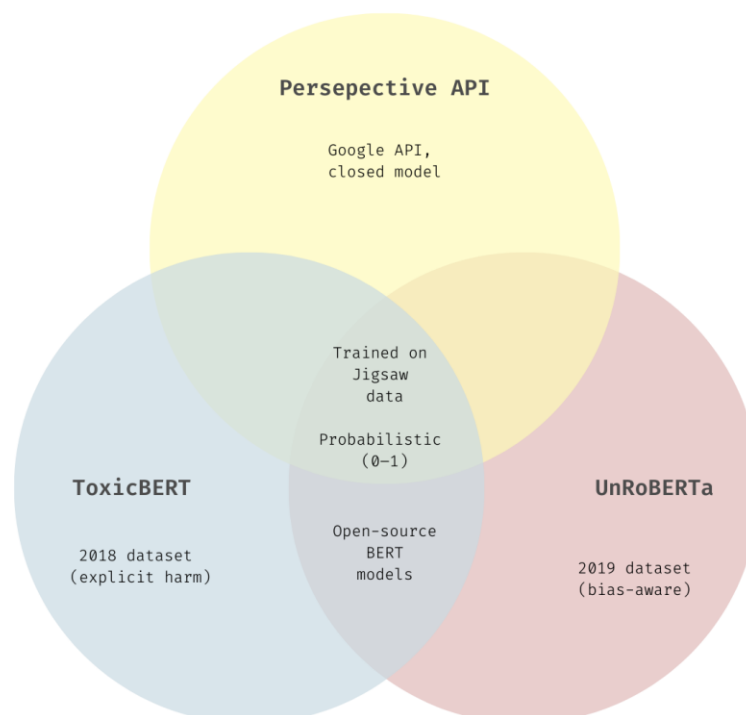


Figure 3.2: Comparative Overview of Toxicity Detection Models

While these models differ in scope and sensitivity, none are without limitations. They have been validated on large, heterogeneous datasets and offer useful baselines for toxicity research, but like all automated tools, they can diverge from human judgment, especially in culturally specific or ambiguous cases. We therefore do not treat any model’s output as objective truth. Consistent with the critical approach explained earlier, we treat the scores as rough estimates of how offensive a phrase might seem, based on how these models work within a larger system of algorithmic moderation. In this study, Perspective API, ToxicBERT, and UnRoBERTa are not seen as final judges of what counts as harmful, but as different tools that help us spot and compare possible signs of toxicity. The validation process described in Section 3.5 and 3.6 was conducted to compare the three tools and determine which model most closely aligns with human judgment, ultimately guiding our choice for the main analysis.

3.3 Data Collection

The central goal of this study is to assess whether TikTok’s autocomplete feature contributes to the spread of toxic suggestions, particularly in relation to identity-based queries. To do so, we adapt the methodological framework proposed by Lin et al. (2023), which provides a structured approach to constructing bias-sensitive prompts and analyzing autocomplete results. While Lin’s model was originally designed for search engines such as Google, it is well suited to the TikTok context and is modified here accordingly.

The process begins with the construction of a set of 3,136 prompts. Each prompt consists of two components: a selection condition (sel) and a semantic content trigger (con). The sel corresponds to specific identity-based values, including “men” or “women” for gender, “Black” or “White” for race, and “gay person” or “straight person” for sexual orientation. This structure reflects a methodological simplification aimed at enabling clear, controlled comparisons across prompts. By holding the grammatical form and thematic context constant while varying only the identity term, the study can better isolate differences in how autocomplete responds to each group.

This binary design helps surface imbalances in toxic predictions across identity pairs and supports scalable, systematic analysis. While it would be ideal to include a broader spectrum of identities, such as non-binary, Latino, Asian, transgender, or multiracial individuals, doing so would increase linguistic variation and reduce comparability, especially within a limited sample size. These exclusions are not conceptual but practical, chosen to maintain internal consistency and analytical clarity. The binary structure provides a constrained but rigorous framework to identify patterns of toxicity; it is not meant to suggest that other identities are less important, but rather to ensure that the patterns observed are not confounded by uncontrolled variability.

Synonyms and related terms are included to reflect common variations and ensure broader coverage. The con component is designed to be syntactically neutral but capable that trigger biased autocomplete suggestions. Following Lin et al., interrogative structures such as “why,” “should,” and “are” are used in combination with themes related to appearance, behavior, personality, career, and relationships. For example, prompts like “Why are Black people...” or “Should gay men...” are designed to test how TikTok’s autocomplete system completes phrases that may activate stereotypes. Details of query-prefix construction can be found in Appendix A1.

Building on Lin et al. (2023), we further categorize each prompt into one of eight declarative themes, Appearance & Physical, Identity, Behavior, Personal, Cognitive, Career, Domestic, and Relationship, because these domains have been shown to capture distinct spheres where social-group-based stereotypes frequently manifest. By adopting these categories, we ensure that our analysis aligns with an empirically validated framework for auditing algorithmic bias in autosuggestions and allows for direct comparison with prior work.

Each of the 3,136 prompts was automatically submitted through TikTok’s mobile search interface using a Python script executed in Visual Studio Code. The script ran on a Chromium-based headless browser, a type of automated browser that simulates user activity without displaying a visual interface, allowing programmatic control over input and page loading. To ensure consistent and region-specific interaction with the platform, ten TikTok accounts were created, one for each target country, using ProtonMail-generated email addresses. The script systematically entered each prompt and waited for the autocomplete suggestions to appear. Once the suggestions were fully rendered, the HTML content of the page was captured and parsed to extract the visible completions. This procedure was repeated for each of the ten selected countries, Pakistan, Singapore, Nigeria, Ghana, UK, Ireland, USA, Trinidad and Tobago, Australia and New Zealand, by routing internet traffic through a VPN to geolocate each session accordingly.

The full collection of data took two weeks to complete. Each suggestion was stored alongside its associated prompt, country, identity category, and rank position. After data collection, a second script was used to assign toxicity scores from all three models, Perspective API, ToxicBERT, and UnRoBERTa, resulting in a parallel dataset of 225,436 toxicity scores per model. Specifically, the Perspective API’s general toxicity model outputs a score ranging from 0 (very unlikely to be perceived as toxic) to 1 (very likely to be perceived as toxic by most readers). While all three models were included at this stage, the main analysis uses Perspective API, based on its comparative performance in human validation (see Section 3.5 and 3.6). This produced a structured dataset of autocomplete results, each annotated with metadata and a toxicity score. The automated and standardized nature of this pipeline ensured consistency across countries and minimized subjective bias throughout the data collection and evaluation process.

Although TikTok typically returns up to eight autocomplete suggestions per query, the actual number of completions varies across countries and prompts. This variation arises from two key factors. First, TikTok’s autocomplete is dynamic and selectively triggered: suggestions are only generated when a query has sufficient search volume, topical relevance, or is deemed safe under platform policy. Some prompts return a full list of eight, others only three or five, and in certain cases, none at all. Second, regional content policies and legal frameworks influence the visibility of specific prompts. TikTok tailors its autocomplete outputs according to the regulatory and cultural environment of each country, which means that a query may return suggestions in one location but be suppressed or filtered in another.

Across the ten participating countries, the total number of autocomplete suggestions collected is 225,436 entries, with per-country totals ranging from 21,514 (Pakistan) to 23,736 (UK). These discrepancies are expected given the platform’s adaptive filtering mechanisms and do not indicate systemic bias or data inconsistency. While most prompts returned the full set of eight suggestions, some returned fewer. If they were not systematic or large enough to meaningfully distort group-level comparisons. For instance, the proportion of prompts with zero results ranged from 0.34% (UK) to 0.56% (USA), while prompts returning fewer than eight suggestions made up 1.2% to 3.6% of total prompts depending on the country. This raises a valid question: if one side of a binary pair (e.g., “homosexual”) consistently receives fewer completions than its counterpart (e.g., “heterosexual”), it could bias the comparison of toxicity scores simply due to volume. To address this, exploratory checks were conducted to compare result counts across identity terms within each category. While minor discrepancies exist, they were not systematic or large enough to meaningfully distort group-level comparisons. To further illustrate country-level variability and data completeness, Table 3.3 provides a detailed breakdown of total autocomplete results, along with the number of prompts returning zero results and fewer than eight results.

Table 3.3: Summary of autocomplete suggestion counts by country

Country	Total Results	Prompts with Zero Results	Prompts with <8 Results
Pakistan	21,514	122	766
Singapore	22,359	94	591
United Kingdom	23,736	80	292
Ireland	21,942	101	674
Ghana	23,734	81	302
Nigeria	22,500	91	568
USA	21,791	123	689
Trinidad and Tobago	22,641	96	543
Australia	22,682	90	540
New Zealand	22,538	96	566

To ensure transparency and replicability, Appendix A2 provides detailed documentation of result variability across prompts. This includes a list of the 98 prompts (3.1% of the total) that returned zero autocomplete suggestions in all ten countries. These cases tend to involve sexual, identity or intimacy related topics and likely reflect platform-specific moderation thresholds or insufficient query volume. In addition to the zero-result cases, Appendix A3 includes a frequency matrix showing how many prompts returned fewer than eight suggestions, categorized by country and by completion count (0 through 8). A second table in Appendix A4 presents illustrative examples of such prompts across countries, offering insight into how localized platform behavior and content governance affect result visibility. These patterns, though relatively infrequent, highlight the dynamic nature of TikTok’s autocomplete system. Their presence does not compromise the validity of the dataset or study; rather, they underscore the importance of documenting platform contingencies in algorithm auditing research.

3.4 Data Analysis

The analytical approach taken in this study aims to detect meaningful patterns in how TikTok’s autocomplete system generates potentially toxic suggestions in response to identity-sensitive queries. The focus is on examining not just the presence of toxic content, but its distribution across identity groups, countries, and ranking positions within the autocomplete results. This comparative lens allows for the detection of systemic disparities that might not be visible through frequency counts alone.

Each suggestion in the dataset is annotated with metadata such as identity dimension (e.g., race, gender, sexual orientation), regional origin, and rank position. This structured format enables multidimensional disaggregation of results. Mean toxicity scores were compared across matched identity subgroups within thematic contexts (e.g., appearance, criminality, family roles), allowing for the detection of group-level disparities in specific discursive frames. While prompt phrasing was not analyzed in isolation, the interaction between identity terms and thematic framing offered insight into how certain contexts amplify or dampen toxic associations.

Data preprocessing and analysis were conducted in Python, using packages such as Pandas, NumPy, and Seaborn. Visualizations include kernel density plots, boxplots, heatmaps, and linear regression charts to illustrate how toxicity varies across subgroups and ranking positions. As noted

in the data collection section, prompts that yielded fewer than eight suggestions or none at all are also examined for patterns in absence. These anomalies are not treated as noise but rather as algorithmically informative outcomes, potentially reflecting content suppression, regional moderation, or platform-side filtering. Supporting matrices and samples of these anomalies are included in Appendix A.

All subsequent statistical testing and pattern exploration follow this foundation, with full results and interpretations presented in Chapter 4.

3.5 Pilot Validation Using Severity Ratings

As an initial step toward understanding how well automated models reflect human perceptions of toxicity, we conducted a preliminary validation study focused on the severity of perceived harm in autocomplete suggestions. This early-stage analysis aimed to evaluate whether model-generated toxicity scores align with how humans interpret different levels of toxic expression. In doing so, the study also surfaced challenges associated with subjective judgment and informed key design decisions for the more comprehensive validation procedure described in Section 3.6.

We began by drawing a simple random sample of 150 autocomplete suggestions from the full dataset of 225,436. These completions were selected to include a range of identity categories, toxicity levels, and structural forms, ensuring that the sample captured the diversity of expressions encountered in the broader corpus. Each suggestion was then independently rated on a five-point toxicity scale (1 = Not toxic; 5 = Extremely toxic) by three annotators: the author and two peers currently enrolled in the MSc in Management of Technology program. All three raters applied the same evaluation rubric (Appendix B1), which provided detailed definitions and boundary criteria for each score level. This rubric was designed to limit inconsistency in how content is understood and promote a consistent framework for subjective judgment.

To benchmark the automated classifiers, we ran the same 150 suggestions through three toxicity detection models: Perspective API, ToxicBERT, and UnRoBERTa. Each model returned a continuous toxicity score, which we converted into a discrete 1–5 scale by applying calibrated threshold cutoffs aligned with the human rubric. This allowed us to compute direct accuracy comparisons: for each model, we measured the proportion of suggestions for which its assigned score matched the rounded average of the three human ratings.

To evaluate the consistency of human judgments, we calculated Fleiss' κ , a statistical measure of inter-rater reliability that adjusts for chance agreement. The result, $\kappa = 0.0857$, is extremely low, indicating agreement only marginally above random chance. Conventional benchmarks interpret values between 0.00–0.20 as “slight,” 0.21–0.40 as “fair,” 0.41–0.60 as “moderate,” 0.61–0.80 as “substantial,” and 0.81–1.00 as “almost perfect” agreement (Landis & Koch, 1977). That our κ falls at the bottom of the scale strongly suggests a fundamental lack of shared understanding among annotators.

This finding is particularly striking given that all three raters shared similar educational backgrounds, geographic context, and were working from the same rubric. One would reasonably expect a group with such alignment to reach at least moderate consensus. Instead, the results expose just how subjective and interpretive toxicity judgments can be, even under seemingly controlled conditions. It also raises concerns about the viability of the five-point rating scale itself, which may force annotators to draw artificial distinctions between adjacent categories like “very toxic” and “extremely toxic.”

Given the extremely low inter-rater agreement ($\kappa = 0.0857$), the accuracy scores of the three models on the five-point scale should be interpreted with caution. Although Perspective API showed the

highest alignment with human ratings (53.3%), such comparisons are difficult to trust when the underlying human judgments lacked internal consistency. This further reinforced the possibility that the five-point scale introduced unnecessary ambiguity, particularly in the mid-range categories like “slightly” or “very toxic,” where both humans and models struggled to draw clear boundaries. Table 3.4 below summarizes the performance of each model on the five-point classification task, including overall accuracy, and precision and recall scores for the non-toxic class (Class 1), which dominates the dataset.

Table 3.4: Toxicity Model Performance (5-Point Scale)

Model	Accuracy	Precision (Class 1)	Recall (Class 1)
Perspective API	53.3%	0.76	0.68
UnRoBERTa	43.3%	0.56	0.84
ToxicBERT	12.7%	0.30	0.08

For this reason, we decided to try a simpler three-point scale to see if that would lead to more stable and useful results. To address these problems and improve interpretability, we reran the validation process using a simplified three-point scale:

1 = Not toxic

2 = Ambiguously toxic

3 = Toxic

After this change, Fleiss’ κ rose to 0.34. While this technically moves us into the “fair agreement” range, it’s still surprisingly low, especially considering that annotators were now choosing between only three options. This highlights just how subjective toxicity judgments can be, even with a reduced scale and shared guidelines. It suggests that the core issue isn’t only the number of options, but the nature of the task itself: people interpret “toxicity” in very different ways.

Once again, Perspective API showed the best match with human ratings, reaching 72% accuracy, followed by UnRoBERTa at 60%, and ToxicBERT at just 20.7%. Table 3.5 summarises the comparative performance of the three classifiers after collapsing scores to the simplified three-point scale.

Table 3.5: Toxicity Model Performance (3-Point Scale)

Model	Accuracy	Precision (Class 1)	Recall (Class 1)
Perspective API	72.0 %	0.81	0.68
UnRoBERTa	60.0 %	0.77	0.75
ToxicBERT	20.7 %	0.68	0.20

However, we must interpret these numbers carefully. Even after switching to a simpler three-point scale, human agreement was still weak, so model “accuracy” is being measured against a noisy benchmark. The takeaway here isn’t just which model scored highest, but how hard it is to define toxicity in a way that people interpret the same way.

That said, it’s important to note a key limitation in our setup: we are essentially comparing two different things. Human annotators rated the severity of toxicity (low, medium, high), while models produce probabilities, i.e., how likely a sentence is to be toxic. These are not the same, and shouldn’t be treated as fully interchangeable. Still, we decided to map probability to severity for practical reasons: it offers a rough way to evaluate how well models reflect human perception. While imperfect, this comparison helps identify broad patterns and gives early insight into whether a model is picking up on the same signals that humans associate with harmful content.

Overall, this pilot helped confirm two things: that judging toxicity is highly subjective, even in small groups and that rating scales alone can't fully solve the problem. With these insights, we move into Section 3.6, where we run a more systematic validation. This time, we focus on comparing model probability outputs with human likelihood estimates, using 10 annotators to better simulate how real users might perceive toxicity at scale.

3.6 Main Validation Using Probability Estimates

The pilot validation in Section 3.5 showed that asking annotators to rate toxicity on a fixed severity scale often led to inconsistent and subjective results. To improve reliability and better reflect how modern toxicity detection models operate, the main validation adopts a different approach. Rather than rating how toxic a suggestion feels, annotators were simply asked to make a binary decision: toxic or not toxic. This more straightforward task avoids forcing people into arbitrary categories and instead mirrors the way training data is often collected for models like the Perspective API, where each example is labeled as toxic or not by multiple raters.

For this phase, we collected judgments from 10 independent annotators for each of the same 150 autocomplete suggestions used in Section 3.5. Each annotator gave a binary label for each suggestion. Then, rather than applying majority voting or assigning discrete categories, we simply computed the proportion of annotators who marked each suggestion as toxic. For example, if 3 out of 10 annotators judged a suggestion as toxic, the resulting human score was 0.3. This score was then interpreted as the probability that a randomly chosen person would find the suggestion toxic, a logic that aligns directly with the output design of models like Perspective API, which are trained to predict the likelihood that content will be perceived as harmful.

This setup treats variation in human judgment not as noise, but as a meaningful signal, reflecting the same ambiguity and disagreement that models are meant to capture. By avoiding predefined buckets or consensus thresholds, we embrace the probabilistic nature of toxicity perception itself. This approach also allows for a direct, one-to-one comparison between human-derived scores and model-generated probabilities, without the need to force labels into strict categories.

As before, each suggestion was rated anonymously using a brief guideline and five calibration examples. The final dataset consists of continuous human scores ranging from 0 to 1 in 0.1 increments for each suggestion, which were then compared to the probability scores of the three models: Perspective API, UnRoBERTa, and ToxicBERT. Because both models and humans were now operating on the same 0–1 probability scale, this setup allowed for a more realistic and fair comparison of how well model predictions align with collective human perception.

To quantify this alignment, we evaluated each model's predictions against the human probability scores using several continuous comparison metrics. As outlined earlier, each human score represents the proportion of annotators (0 to 1) who judged a suggestion as toxic. Since the output of all models is also a probability between 0 and 1, we can evaluate alignment using continuous-value comparison metrics, rather than hard classification categories.

To capture this alignment from multiple angles, we employ the following three metrics:

- **Mean Absolute Error (MAE):** Measures the average difference between model scores and human scores, regardless of direction. Lower values are better, as they indicate closer overall alignment. For example, a model with MAE = 0.2 deviates by 20 percentage points on average from what humans perceive.
- **Close Match Rate (± 0.1):** Percentage of predictions that fall within ± 0.1 of the human score. Higher values are better, showing how often the model is nearly indistinguishable from human perception.

- **Close Match Rate (± 0.2):** Same as above, but with a more forgiving ± 0.2 window. Higher values here capture broader agreement while allowing for minor variation.

These metrics were chosen because they reflect different expectations: MAE penalizes consistent over- or under-scoring, while the close match rates indicate how frequently the model is functionally indistinguishable from human consensus. The results are summarized in Table 3.6:

Table 3.6 Model Comparison Based on MAE and Close Match with Human Scores

Model	MAE	Close Match ± 0.1	Close Match ± 0.2
Perspective API	0.192	34.00%	60.67%
UnRoBERTa	0.322	35.33%	45.33%
ToxicBERT	0.540	10.00%	19.33%

As shown in the table, Perspective API has the lowest mean absolute error (MAE), meaning its predictions are, on average, the closest to what human annotators reported. It also has the highest percentage of cases where its scores are within ± 0.2 of the human score, and it comes very close to UnRoBERTa in the stricter ± 0.1 range. While UnRoBERTa is more careful in how it handles identity-related prompts, as discussed in Section 3.2, this caution sometimes leads it to score potentially harmful suggestions too low. ToxicBERT, on the other hand, performs the worst across all metrics. It tends to flag many suggestions as toxic even when they appear neutral, resulting in poor overall alignment with human perception.

Given its overall performance across all metrics, Perspective API is selected as the primary model for the remainder of the analysis.

3.7 Practical Comparison of Toxicity Models

To build on the validation analysis and provide a more concrete understanding of how the three toxicity detection models behave in practice, we conducted a detailed comparison using the same random sample of 150 autocomplete suggestions. Rather than relying only on aggregate correlation scores or summary statistics, this section zooms in on specific cases to illustrate how each model responds to particular types of language. This approach enables a more detailed view of model behavior, helping clarify what it means when we say, for example, that ToxicBERT tends to over-flag or that UnRoBERTa is more conservative. The examples discussed were selected not only because they triggered unusually high disagreement or strong alignment among models, but also because they reveal edge cases that test the boundaries of each model's design philosophy as outlined in Section 3.2. These examples help connect theoretical expectations to real outputs and highlight the practical implications of model choice for bias detection in autocomplete systems.

The results are presented in Table 3.7 below. For each suggestion, we include the predicted toxicity score from each model and note the maximum disagreement between them. This format enables a compact yet interpretable comparison.

Table 3.7 Selected Examples of Model Behavior on Identity-Related Autocomplete Suggestions

Prompt	Suggestion	Perspective	ToxicBERT	UnRoBERTa	Max Diff
why lesbian	why lesbians are ugly	0.78	0.07	1.00	0.93
wife in relationship	husband and wife	0.13	0.999	0.00	0.999

	relationship in ramadan				
women cognition	cognition	0.02	0.999	0.00	0.999
white woman running	woman running around meme	0.23	0.85	0.00	0.85
how many straight people	why straight people exist	0.51	0.98	0.02	0.97
gay person looks	send this to someone gay person	0.59	0.22	0.99	0.77
straight man housework	housework men	0.04	0.998	0.03	0.97
straight woman housework	straight from work	0.02	0.999	0.00	0.999
straight man career	straight men being straight	0.24	0.999	0.00	0.999

Across multiple cases, ToxicBERT consistently assigns very high toxicity scores to syntactically neutral or vague suggestions, such as “straight from work” (0.9992) and “cognition” (0.9993, in the context of “women cognition”). These results reflect its tendency to react to surface-level token associations, like “straight,” “woman,” or “work” rather than contextual meaning. This pattern confirms earlier characterizations of ToxicBERT as overly aggressive, especially in identity-related contexts, where it tends to over-flag due to patterns learned from toxic co-occurrence in its training data.

By contrast, UnRoBERTa shows a more conservative approach, with suppressed scores in the same cases (e.g., 0.0001 for “straight from work” and “cognition”, and 0.0157 for “why straight people exist”). This aligns with its training on the Unintended Bias in Toxicity dataset, which aims to reduce false positives on identity references. Notably, UnRoBERTa still assigns a high score of 0.9989 to “why lesbians are ugly”, indicating that while it is cautious, it remains sensitive to clearly harmful content a balance that affirms its design goals.

The Perspective API typically falls between these two extremes. For instance, it assigns a score of 0.5079 to “why straight people exist”, a phrase that might carry subtle implications depending on context. While ToxicBERT flags it as highly toxic (0.9809) and UnRoBERTa assigns a near-zero score (0.0157), the Perspective API’s moderate rating may better reflect the ambiguity and potential for harm without overreacting. Similarly, it gives “husband and wife relationship in ramadan” a score of 0.1270, suggesting some sensitivity to cultural or religious framing. These mid-range scores on identity-related prompts indicate that the model is broad in scope and able to register nuance, though it can still vary in calibration across different edge cases.

Taken together, these examples illustrate the distinct behavioral tendencies of each model. ToxicBERT, though capable of identifying overt hate speech, may distort bias assessments if used alone due to its overflagging. UnRoBERTa fulfills its fairness-oriented role by minimizing false positives and only flagging unambiguous harm. Perspective API provides a contextual middle ground, showing broader sensitivity to subtle forms of toxicity, though it can occasionally misjudge

culturally or identity-related phrases. These differences reinforce the value of using the models in combination, since their different strengths make it possible to carry out a more balanced and complete analysis of stereotype reinforcement in TikTok's autocomplete predictions.

4

Results

4.1 Results overview

The present chapter presents the empirical findings derived from the dataset of 225,436 autocomplete suggestions collected across ten countries. The primary aim is to empirically examine the prevalence and distribution of toxic content within TikTok’s autocomplete system, with a specific focus on identity-sensitive queries. Following the validation results in Section 3.5 and 3.6, all toxicity scores reported in this chapter are based on the Perspective API model. The analysis is structured around the three core research questions introduced earlier, addressing (1) the extent to which toxic content appears across gender, race, and sexuality dimensions, (2) the variation in autocomplete toxicity across geographic regions, and (3) the relationship between the rank order of suggestions and their toxicity scores.

To ensure analytical clarity and alignment with the research design, the results are presented in a modular structure. First, we provide an aggregate overview of toxicity scores in the dataset to establish a general baseline. This includes descriptive statistics that quantify the presence of toxicity across all autocomplete suggestions. Next, we explore intra-group differences across identity dimensions by statistically comparing mean toxicity scores within matched identity pairs, including gender (e.g., “men” vs. “women”), race (e.g., “Black” vs. “White”), and sexual orientation (e.g., “gay person” vs. “straight person”). The comparison strategy couples summary measures with distribution-level evidence, Kolmogorov-Smirnov tests, violin-plot density profiles, and tallies of high-toxicity outliers, to capture both central tendencies and tail behaviour. These analyses offer insight into whether and how TikTok’s autocomplete system may reproduce or reflect societal biases. This is followed by a cross-regional analysis that evaluates whether and how toxicity levels vary across the ten national contexts under study. Finally, the chapter examines the internal structure of autocomplete results by analyzing whether toxicity is concentrated at higher-ranking suggestions, a crucial factor for understanding user exposure.

Each subsequent section of this chapter corresponds to one research question, combining descriptive summaries, statistical tests, and interpretive commentary to highlight key patterns and differences in the data. Where appropriate, tables and visualizations are included to support and clarify the findings. Appendix A material is referenced for secondary results and validation procedures.

4.2 General Toxicity Trends

This section summarizes the toxicity patterns found in TikTok autocomplete results drawn from ten English-speaking countries (225,436 suggestions). The goal is to establish a general baseline before addressing the study’s three core research questions in subsequent sections. Figure 4.1

visualises the full score distribution. By characterizing these overall patterns, we gain critical context for interpreting both within-group disparities and cross-regional trends explored later in the chapter.

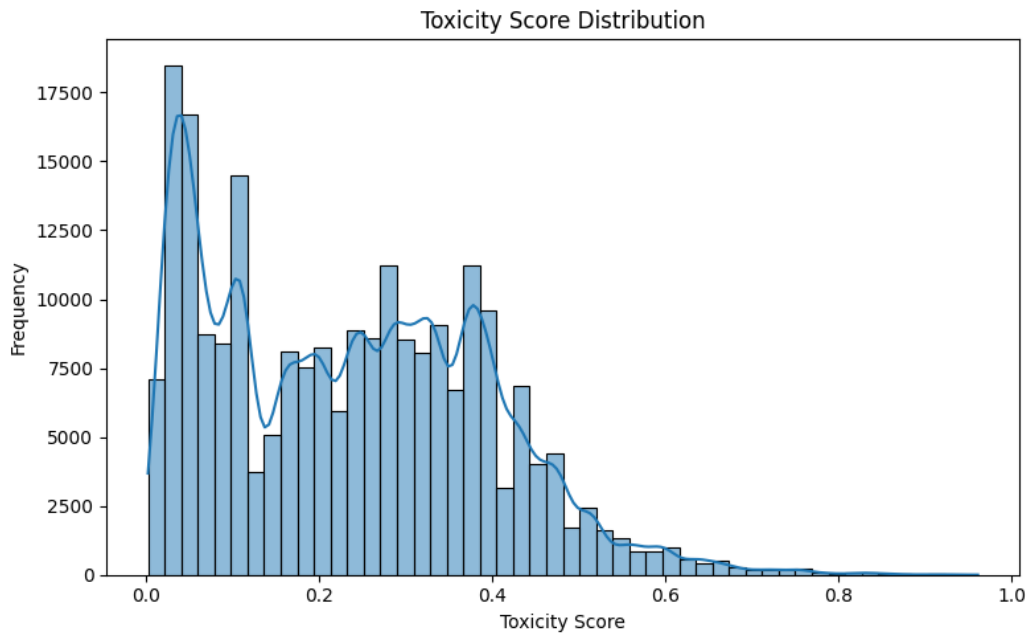


Figure 4.1: Histogram and KDE of toxicity scores

Each suggestion in the dataset is linked to one of 3,136 identity-based prompts, constructed across three identity dimensions, gender, race, and sexual orientation, and collected through VPN-simulated access to national TikTok interfaces.

The toxicity scores which represent the estimated probability that a human annotator would perceive a suggestion as toxic, range from 0 (very unlikely to be seen as toxic) to 1 (very likely to be seen as toxic), yield a mean of 0.234, a median of 0.232, and a standard deviation of 0.157, indicating a right-skewed distribution concentrated around low to moderate probabilities of perceived toxicity. While the majority of completions are unlikely to be viewed as toxic, the tail-end of the distribution contains high-probability predictions that raise concerns about potential linguistic harm.

To support interpretability, toxicity values were grouped into five bins:

- 0.0–0.2 (very low – unlikely to be perceived as toxic)
- 0.2–0.4 (low – mildly likely to be perceived as toxic)
- 0.4–0.6 (moderate – moderately likely to be perceived as toxic)
- 0.6–0.8 (high – very likely to be perceived as toxic)
- 0.8–1.0 (very high – extremely likely to be perceived as toxic)

While earlier sections applied a simpler three-level distinction for certain comparisons, this five-bin scale is used here to enhance readability and precision during descriptive analysis. Using five bands strikes a useful balance. Equal 0.2-point steps map neatly onto a familiar five-level verbal scale (none → extreme), so the results stay easy to explain to non-specialists. At the same time, five bins retain meaningful distinctions in the mid-range of toxicity likelihood without introducing the noise that appears when you slice the data into very thin deciles; fewer than five would blur important differences, while more than five would over-fit scarce extreme cases. The scheme also matches the binning approach adopted in earlier algorithm-auditing work (Wulczyn et al., 2017),

making our findings simpler to compare with the literature. It is important to note that these categories reflect probabilistic assessments rather than categorical judgments; for instance, a toxicity score of 0.65 suggests that the model assigns an estimated 65% probability to the presence of toxic language, not a definitive classification. Table 4.1 summarizes the breakdown across score intervals, and Figure 4.2 visualises the same counts as a bar chart.

Table 4.1: Distribution of Autocomplete Suggestion Toxicity by Score Range

Toxicity Range	Count	% of Dataset
0.0–0.2	101,048	43.5%
0.2–0.4	90,216	40.0%
0.4–0.6	30,460	13.5%
0.6–0.8	3,479	1.5%
0.8–1.0	233	0.1%

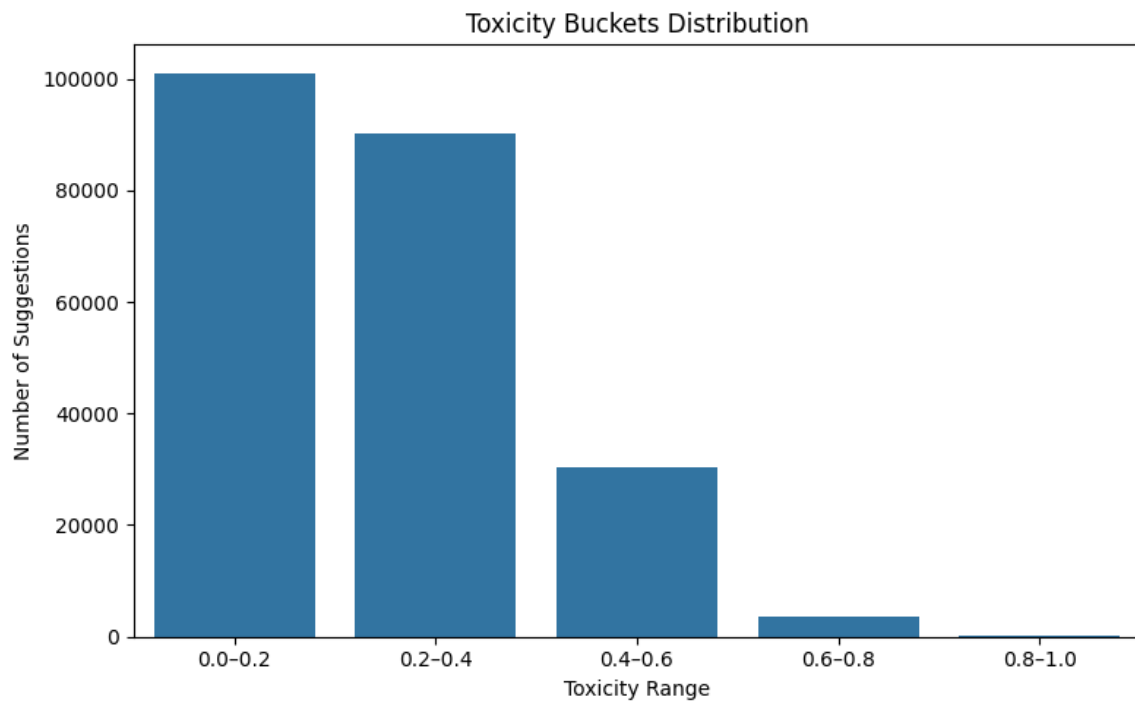


Figure 4.2: Five-band breakdown of toxicity percentages

Although overtly toxic completions (scores ≥ 0.80) make up only 0.1 % of the dataset, they raise disproportionate concern. Figure 4.1 makes the long right-hand tail visible, and Figure 4.2 quantifies its size. Appendix A5, which lists the five highest-scoring suggestions per country, shows a striking consistency: almost every extreme-toxicity phrase combines an identity term, most often “gay,” “black,” or “lesbian” with ridicule, bodily-function expressions. The repetition of nearly identical insults across ten national interfaces suggests that TikTok’s autocomplete model may be recycling a small but highly salient pool of hateful patterns rather than surface-level regional vernacular, pointing to an underlying training-data bias or feedback-loop amplification.

These extreme completions raise a deeper question: is the toxicity introduced by TikTok’s suggestions, or is it already present in the prompt? And what happens when we reverse the identity term, for example, changing “*gay man*” to “*straight man*”, or “*Black person*” to “*white person*”?

To explore this, we built a comparison table using 35 pairs of prompts, each with an identity-based term and its semantic opposite. For each, we compared:

- The toxicity of the original prompt
- The toxicity of the autocomplete suggestion
- The toxicity of the opposite prompt
- The opposite suggestion, if one was generated, and its toxicity

The full set of 30 pairs illustrating this pattern can be found in Appendix A6. Table 4.2 presents a selection of representative examples from this table.

Table 4.2: Prompt–Suggestion Asymmetries

Type	Prompt	Suggestion	Prompt Toxicity	Suggestion Toxicity
Original	<i>gay people smell</i>	<i>gay smelling fart</i>	0.731	0.902
Opposite	<i>straight people smell</i>	-	0.661	-
Original	<i>gay person smell</i>	<i>smell that smells like gay</i>	0.751	0.851
Opposite	<i>straight person smell</i>	-	0.621	-
Original	<i>lesbians knowledge</i>	<i>useless lesbians</i>	0.332	0.853
Opposite	<i>straight women knowledge</i>	-	0.202	-
Original	<i>lesbians employ</i>	<i>useless lesbians</i>	0.361	0.853
Opposite	<i>straight women employ</i>	-	0.249	-
Original	<i>gay people make decisions</i>	<i>gay people farting</i>	0.294	0.869
Opposite	<i>straight people make decisions</i>	-	0.252	-
Original	<i>gay people cognition</i>	<i>gay people farting</i>	0.399	0.869
Opposite	<i>straight people cognition</i>	-	0.282	-
Original	<i>gay people babysit</i>	<i>gay people farting</i>	0.557	0.869
Opposite	<i>straight people babysit</i>	-	0.401	-
Original	<i>black man looks</i>	<i>black man looks like a monkey</i>	0.279	0.854
Opposite	<i>white man looks</i>	-	0.164	-
Original	<i>black woman skin</i>	<i>black woman skit</i>	0.236	0.924
Opposite	<i>white woman skin</i>	<i>white woman skit</i>	0.153	0.869

One of the clearest patterns is that opposite prompts most of the times don't return any suggestion at all, even when the wording is nearly identical. In few cases, the prompt itself already carries a relatively high toxicity score but the suggestion is consistently more toxic, pushing the interaction further into harmful territory. More importantly, when we flip the identity term, for example, *gay* to *straight*, the prompt becomes less toxic, and autocomplete stops returning a suggestion altogether. This happens even though the phrase structure and meaning remain essentially the same. That difference alone, changing only the identity term, leads to different risk assessments and different outcomes.

These examples show that the same structure, just with a different identity term can result in very

different outcomes. Prompts linked to marginalized identities are not only more likely to be rated toxic, but also far more likely to be completed with something offensive. In most cases, the prompt itself is not especially toxic, and is significantly lower than the suggestion that follows it. Autocomplete often takes a low- or moderate-toxicity prompt and transforms it into a highly toxic suggestion. This suggests that the model is not just reflecting prompt content but is actively intensifying it.

Even when opposite suggestions do exist, their toxicity is generally lower. The prompt “*black woman skin*” triggers “*black woman skit*” with a score of 0.92, while “*white woman skin*” produces the same suggestion but with a lower score of 0.87 and a less toxic prompt to begin with. Another striking pattern is the repetition of crude completions, particularly “*gay people farting*”, across prompts that have nothing to do with the topic. Autocomplete attaches this insult across a range of unrelated contexts. This repetition suggests that autocomplete is not simply reacting to prompt content but may be recycling a small set of offensive suggestions disproportionately tied to certain identities

Summarizing the main toxicity trends, the descriptive evidence shows that most autocomplete content is neutral, but a persistent minority reproduces identity-targeted slurs. This pattern warrants the deeper causal and comparative tests presented in the sections that follow. It is also notable that even when suggestions are absent, prompts associated with marginalized identities often carry slightly higher toxicity scores than their dominant-group counterparts, despite having nearly identical structure and meaning. In most cases, the suggestions were notably more toxic than the prompts, indicating that autocomplete can intensify toxicity rather than merely reflect it.

4.3 SRQ1: Toxicity by Identity Category

This section tackles our first research question: Does TikTok’s autocomplete assign different toxicity levels to prompts that reference different identity groups, gender, race, or sexual orientation? Using the toxicity scores introduced in Section 4.2, we compare matched prompt pairs (e.g., “Black person ...” vs “White person ...”) to detect systematic disparities. Our working expectation is that references to minoritized identities (“female,” “Black,” “gay/lesbian”) will attract higher toxicity than references to dominant identities (“male,” “White,” “heterosexual”). To fully address whether identity-based toxicity is a systemic pattern or limited to isolated cases, we combine statistical comparisons, distributional visualizations, and contextual analysis. This multi-method approach allows us to assess not just if disparities exist, but how they manifest across linguistic form and thematic framing.

To guide the analysis, we tested the following null hypotheses:

- H_{01} (Gender): The distribution of toxicity scores for male-related autocomplete suggestions is identical to the distribution for female-related suggestions.
- H_{02} (Race): The distribution of toxicity scores for prompts referring to Black identities is identical to the distribution for prompts referring to White identities.
- H_{03} (Sexual Orientation): The distribution of toxicity scores for prompts involving homosexual identities is identical to the distribution for prompts involving heterosexual identities.

Each null hypothesis was evaluated using a Kolmogorov–Smirnov (KS) test, which compares the full cumulative distributions of toxicity scores without assuming normality. The KS statistic (D) measures the maximum vertical distance between two empirical cumulative distribution functions; a small p-value indicates that the two distributions differ more than would be expected by chance. Significance was assessed at $p < .001$. The following table summarizes subgroup means, standard deviations, KS statistics, and p-values for each identity pairing:

Table 4.3: Group means (with SD), KS statistics, and significance for each identity pairing. (* $p < .001$.) **

Variable	Subgroup	Mean (SD)	KS Stat (D)	p-value	Signif.
Gender	Male	0.099 (0.087)	0.118	< .001	***
	Female	0.119 (0.097)			
Race	Black	0.331 (0.101)	0.367	< .001	***
	White	0.253 (0.104)			
Orientation	Homosexual	0.394 (0.152)	0.610	< .001	***
	Heterosexual	0.206 (0.120)			

Although all three Kolmogorov–Smirnov tests return $p < .001$, formally rejecting H_{01} , H_{02} , and H_{03} , the magnitude of the observed differences varies considerably. For gender, $D = 0.118$ means that at most 11.8 percent of the male and female toxicity curves diverge at any point. With over 225 000 observations per subgroup, the test is exceptionally sensitive, so even this modest gap yields a vanishingly small p-value. In substantive terms, however, such an 11.8 percentage-point separation produces almost complete overlap between the two distributions, an effect too small to carry meaningful interpretive weight. By contrast, the larger statistics for race ($D = 0.367$) and sexual orientation ($D = 0.610$) correspond to clear, visually distinct separations in their distributions and thus represent both statistically and practically significant disparities.

4.3.1 Distributional Patterns by Identity Category

To more clearly illustrate the differences revealed by the KS tests, we examined the full distribution of toxicity scores across identity categories using violin plots (Figures 4.3-4.5). These visualizations allow us to capture not just central tendencies, but also the shape and extremity of the distributions, offering a more complete picture of how toxicity manifests across different social groups.

In the case of gender (Figure 4.3), female-associated suggestions exhibited both a higher median toxicity and a longer upper tail than male-associated suggestions. While the overall difference was modest, the distribution for female prompts showed more outliers in the high-toxicity range, indicating that although extreme completions are rare, they are more likely to appear for prompts involving women.

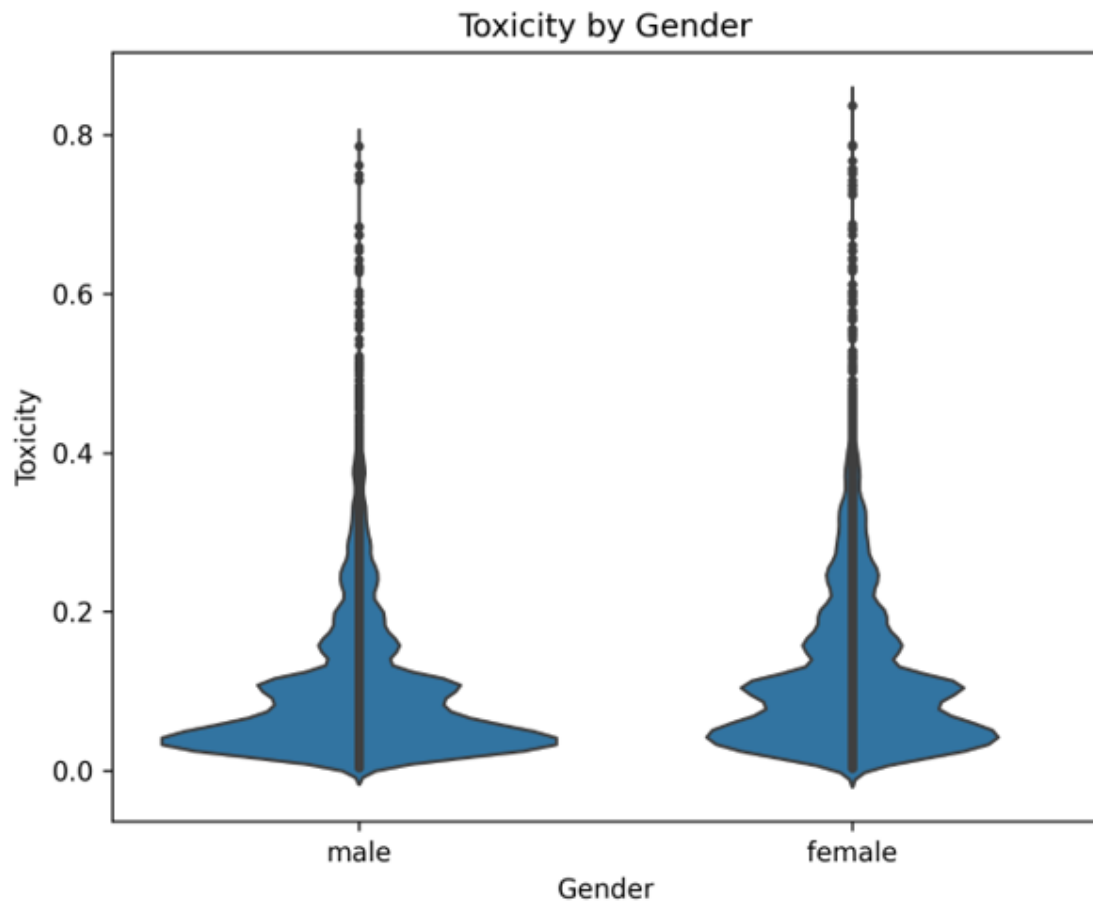


Figure 4.3: Toxicity Distribution by Gender (violin plot)

For race (Figure 4.4), the difference is more pronounced. Prompts involving Black identities cluster around a higher mean (approximately 0.33), with a wider spread toward the most toxic end of the scale. In contrast, prompts referencing White identities are more tightly concentrated around a lower average (~ 0.25) and show fewer cases exceeding a toxicity score of 0.6. This disparity suggests not only a general elevation in toxicity for Black-related prompts, but also a greater density of extreme (high-toxicity) outputs.

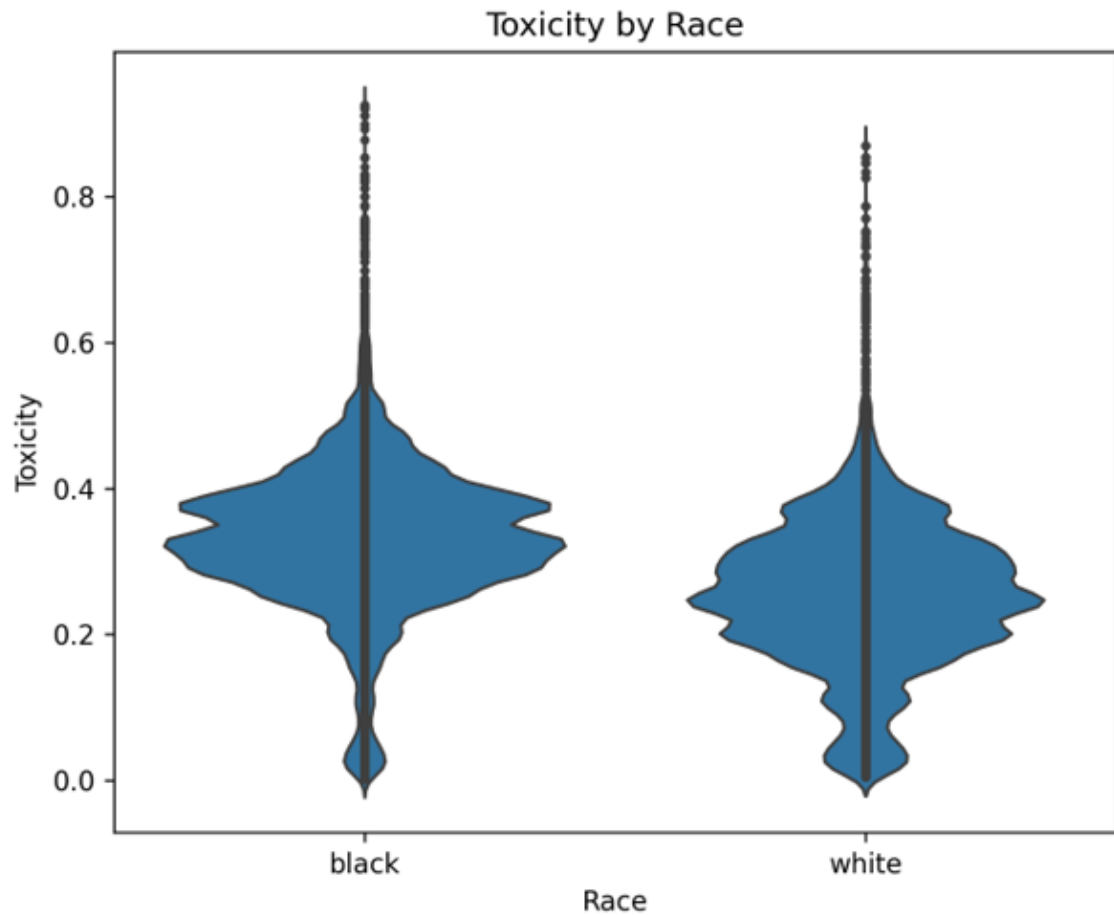


Figure 4.4: Toxicity Distribution by Race (violin plot)

The clearest difference occurred in the case of sexual orientation (Figure 4.5). Prompts related to homosexual identities showed a much higher density of completions in the 0.4 to 0.8 range, along with an extended right-hand tail. Some suggestions even approached maximum toxicity (above 0.9), a phenomenon nearly absent in heterosexual-associated results. This implies that, relative to all other identity dimensions, prompts involving sexual orientation, particularly non-heterosexual identities, are most prone to triggering extreme and hostile completions.

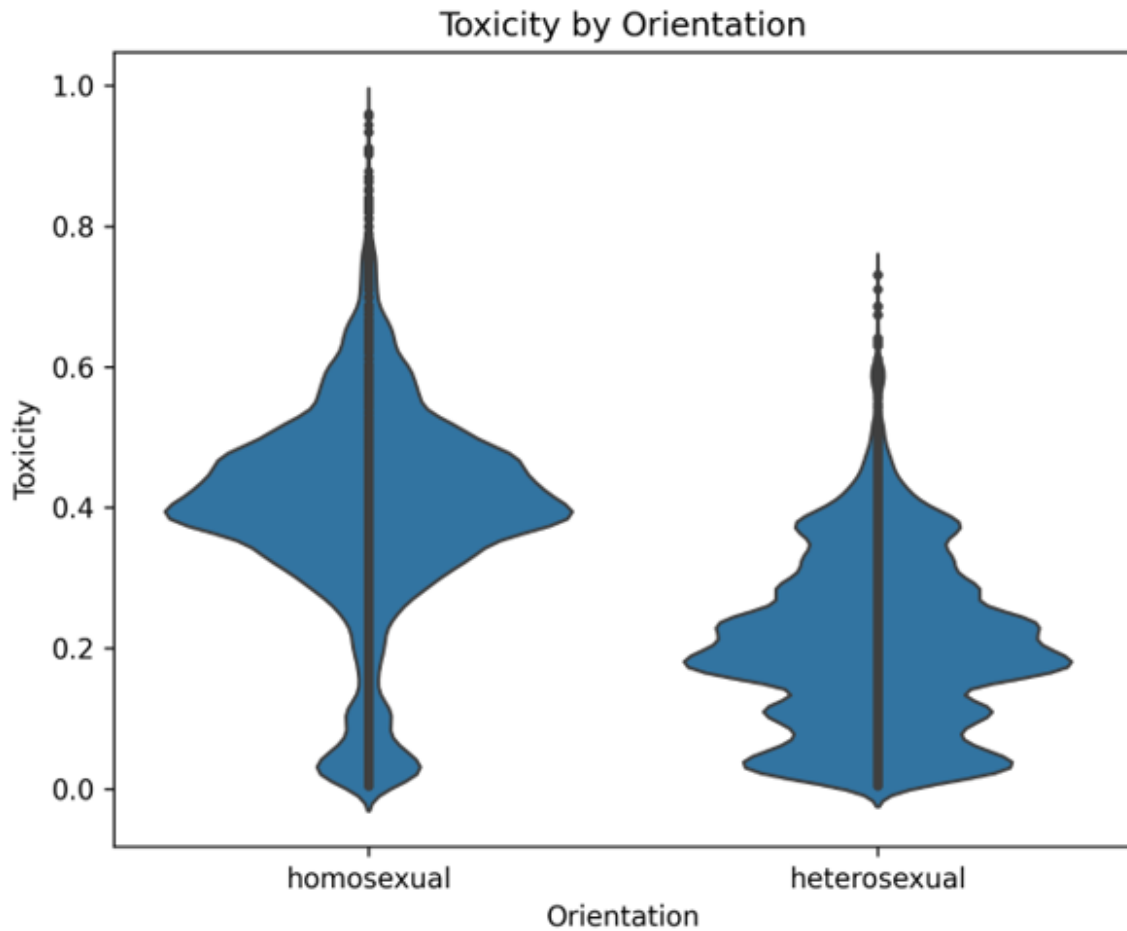


Figure 4.5: Toxicity Distribution by Sexual Orientation (violin plot)

To further quantify the presence of these extreme outliers, we calculated the proportion of suggestions exceeding a toxicity score of 0.7 for each subgroup. Figure 4.6 visualises these percentages. While completions at this level of toxicity are rare for most identity groups falling below 0.15% for male, female, white, and heterosexual prompts, the incidence is markedly higher for certain minoritized identities. Specifically, the proportion rises to 0.42% for Black-related prompts and spikes to 1.66% for homosexual-related prompts. This represents more than a tenfold increase compared to dominant-group baselines. Such stark disparities at the upper end of the toxicity scale are especially concerning, as these completions are not only harmful in content but also carry disproportionate algorithmic weight due to their extremity. The findings suggest that identity terms associated with marginalized groups are significantly more likely to trigger the most egregiously toxic outputs, underscoring a key axis of algorithmic bias in autocomplete behavior.

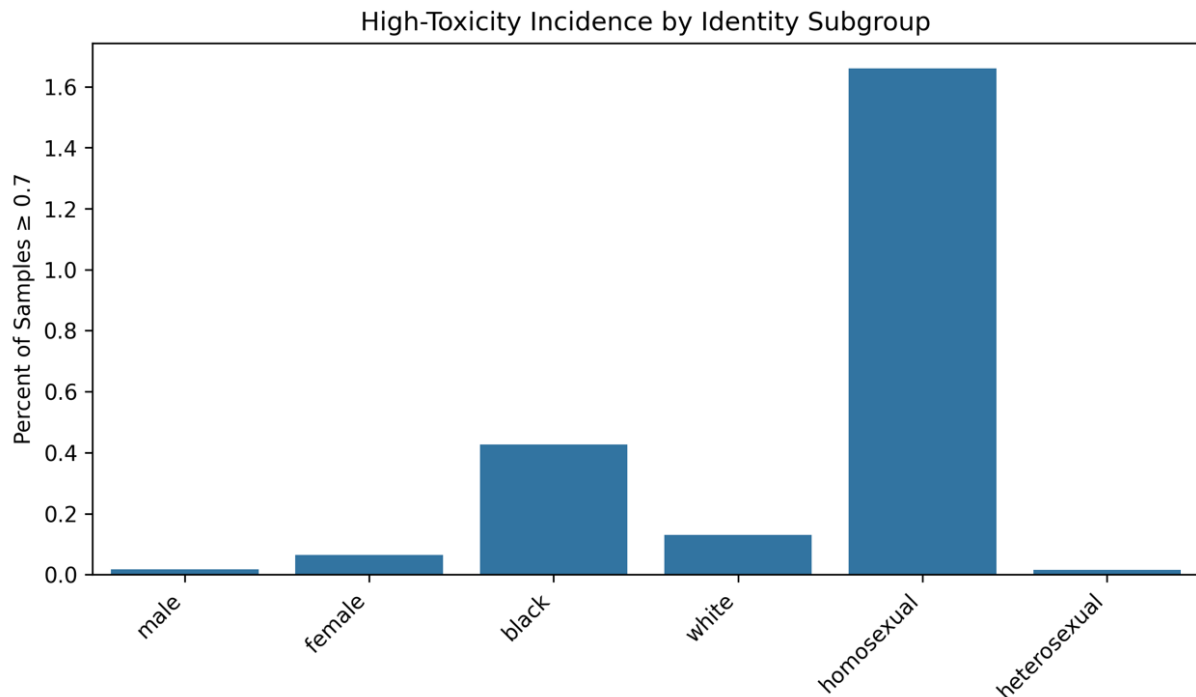


Figure 4.6: High-toxicity incidence (score ≥ 0.7) by identity subgroup.

Taken together, these distributional patterns reinforce the earlier statistical results while adding crucial texture. They show that toxicity disparities are not limited to small average differences but are deeply embedded in the upper tails of the distribution, where linguistic harms are most acute and algorithmic bias most evident.

4.3.2 Contextual Analysis by Declarative Themes

As noted earlier, toxicity must be read considering what aspect of a social group a query targets, so, as already explained in the data-collection section, every suggestion is mapped onto one of eight empirically validated themes. Table 4.4 below summarizes the eight themes, offering brief descriptions along with illustrative examples that reflect typical prompt structures associated with each category.

Table 4.4: Declarative Themes Used in Contextual Toxicity Analysis

Theme	Description	Illustrative Example
Appearance & Physical	Prompts about looks or body attributes	“Why do women have curly hair?”
Identity	Prompts about labels or self-descriptors	“What does it mean to be gay?”
Behavior	Prompts about actions or conduct	“Do Black people behave differently?”
Personal (Personality)	Prompts about character traits	“Are female leaders more empathetic?”
Cognitive	Prompts about intellect or mental capacity	“Are homosexuals less intelligent?”

Career	Prompts about jobs or professions	“Should women pursue engineering?”
Domestic	Prompts about home roles or chores	“Why do Black moms cook soul food?”
Relationship	Prompts about interpersonal or romantic ties	“Is gay marriage acceptable?”

In addition to visual comparisons, we conducted Kolmogorov–Smirnov tests within each declarative theme to assess whether the full distributions of toxicity scores differed by identity subgroup across contexts. For each theme, we compared the two subgroup cumulative distributions, recording the KS statistic (D) and associated p-value for each pair (e.g., male vs. female). Tables 4.5–4.7 summarize mean toxicity, standard deviation, KS D, and significance for gender, race, and sexual orientation themes, respectively. While every theme yields $p < .001$ (reflecting the large sample sizes), the KS statistic D varies substantially by theme, indicating that some contexts (e.g., “Identity” for orientation, $D \approx 0.623$) show pronounced distributional separation, whereas others (e.g., “Career” for gender, $D \approx 0.121$) represent more modest differences that overlap heavily in practice. These variations in D help distinguish between statistically significant but practically minor gaps and those that reflect meaningful distributional separation, reinforcing that identity-linked toxicity is especially pronounced in certain themes like Identity and Appearance, particularly for race and sexual orientation.

Table 4.5: Gender-based significance testing by theme

Theme	Male Mean (SD)	Female Mean (SD)	KS Stat	p-value	Signif.
Appearance & physical	0.126 (0.100)	0.136 (0.101)	0.079	< .001	***
Behavior	0.106 (0.089)	0.133 (0.101)	0.153	< .001	***
Personal	0.086 (0.074)	0.111 (0.085)	0.195	< .001	***
Cognitive	0.085 (0.071)	0.099 (0.084)	0.149	< .001	***
Career	0.061 (0.059)	0.077 (0.072)	0.121	< .001	***
Domestic	0.125 (0.094)	0.163 (0.100)	0.207	< .001	***
Relationship	0.114 (0.091)	0.130 (0.097)	0.136	< .001	***
Identity	0.103 (0.086)	0.141 (0.105)	0.166	< .001	***

Table 4.6: Race-based significance testing by theme

Theme	Black Mean (SD)	White Mean (SD)	KS Stat	p-value	Signif.
Appearance & physical	0.334(0.104)	0.255 (0.118)	0.388	< .001	***

Behavior	0.341 (0.095)	0.271 (0.093)	0.375	< .001	***
Personal	0.323 (0.112)	0.240 (0.097)	0.368	< .001	***
Cognitive	0.307 (0.108)	0.243 (0.102)	0.344	< .001	***
Career	0.305 (0.094)	0.228 (0.107)	0.396	< .001	***
Domestic	0.376 (0.112)	0.283 (0.115)	0.363	< .001	***
Relationship	0.311 (0.073)	0.243 (0.076)	0.458	< .001	***
Identity	0.357 (0.093)	0.270 (0.090)	0.425	< .001	***

Table 4.7: Orientation-based significance testing by theme

Theme	Homosexual Mean (SD)	Heterosexual Mean (SD)	KS Stat	p-value	Signif.
Appearance & physical	0.439 (0.148)	0.187 (0.114)	0.742	< .001	***
Behavior	0.380 (0.146)	0.215 (0.121)	0.584	< .001	***
Personal	0.405 (0.131)	0.206 (0.120)	0.650	< .001	***
Cognitive	0.354 (0.164)	0.227 (0.113)	0.517	< .001	***
Career	0.349 (0.174)	0.166 (0.123)	0.587	< .001	***
Domestic	0.388 (0.176)	0.174 (0.113)	0.606	< .001	***
Relationship	0.359 (0.093)	0.221 (0.103)	0.571	< .001	***
Identity	0.431 (0.115)	0.258 (0.106)	0.623	< .001	***

Within each theme, we compared the mean toxicity scores of matched identity subgroups (e.g., male vs. female, Black vs White, homosexual vs. heterosexual) to assess whether certain discursive frames amplify or suppress bias. Grouped bar charts (Figures 4.7-4.9) were generated to visualize these differences across gender, race, and sexual orientation.

In the case of gender (Figure 4.7), the largest disparities appear in the Career and Relationship themes. Prompts referencing women in these domains yield noticeably higher toxicity than those referencing men, suggesting algorithmic sensitivity to stereotypes involving professional roles and interpersonal dynamics. A smaller but still consistent gap is visible in the Appearance & Physical theme, where female prompts average higher toxicity than male ones.

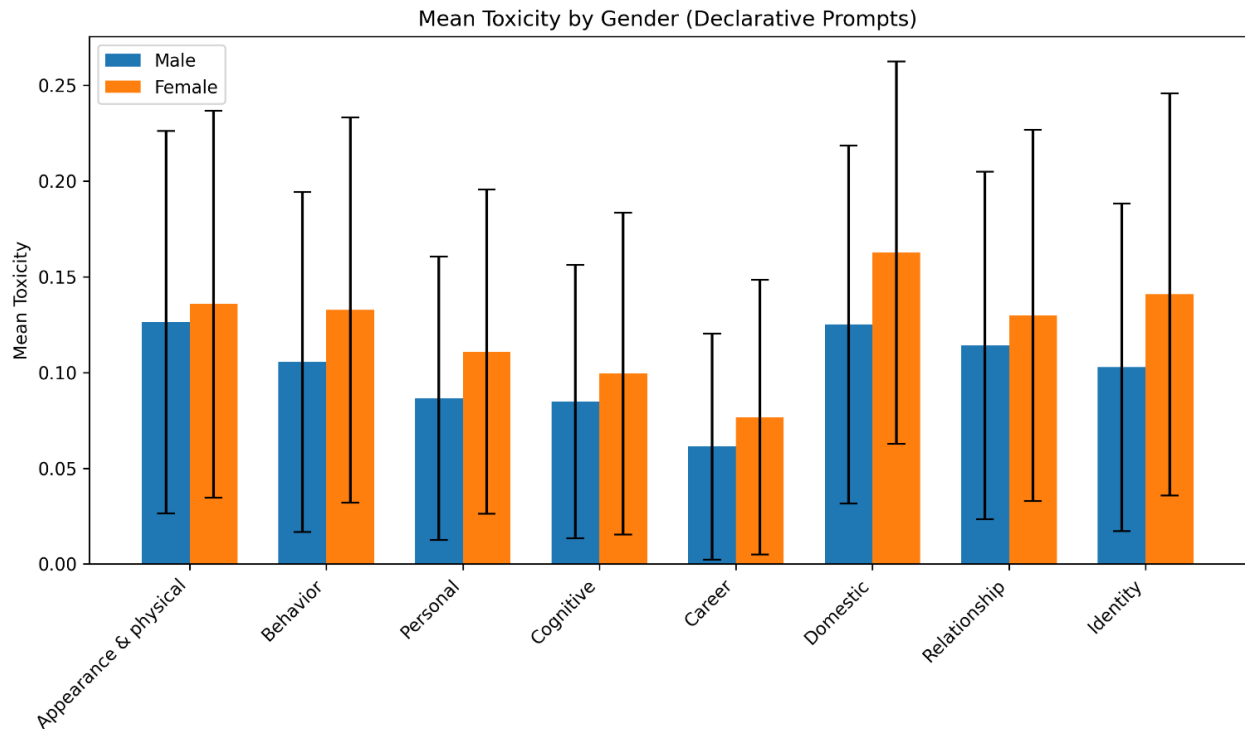


Figure 4.7: Mean Toxicity by Gender (Declarative Prompts)

For race (Figure 4.8), the most striking difference occurs in the Appearance & Physical theme. Prompts referencing Black individuals average a toxicity score of around 0.34, compared to approximately 0.25 for White-associated prompts. This suggests that racialized physical descriptors may trigger more harmful completions. Gaps are also evident in Behavior, Identity, and Cognitive themes, each showing a difference of at least 0.08, indicating that racial disparities extend across multiple evaluative dimensions, from conduct to intellect.

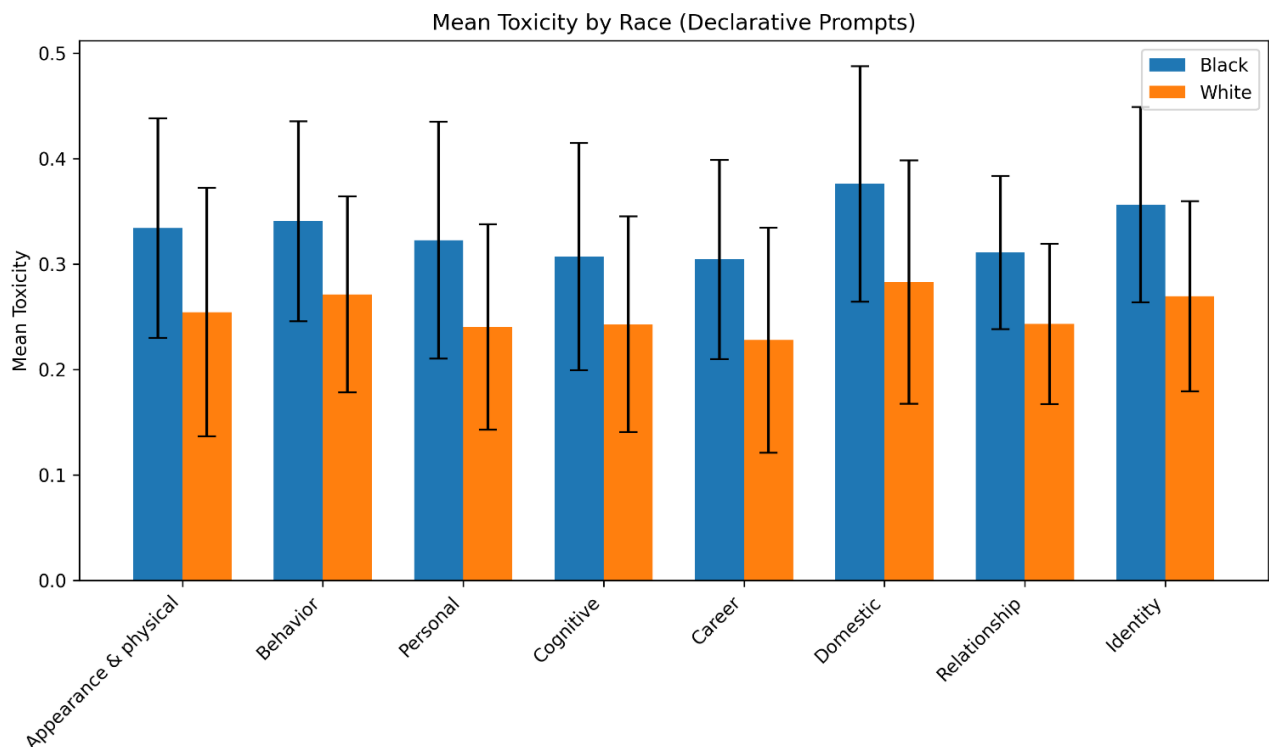


Figure 4.8: Mean Toxicity by Race (Declarative Prompts)

Sexual orientation shows the most dramatic difference (Figure 4.9). In the Identity theme, homosexual-related prompts yield average toxicity levels exceeding 0.40, while heterosexual-related prompts remain below 0.20. Disparities are also prominent in the Behavior and Appearance themes, where toxic completions more frequently target non-heterosexual identities. Together, these findings underscore that toxicity is not evenly distributed across contexts, it intensifies depending on how identity is framed.

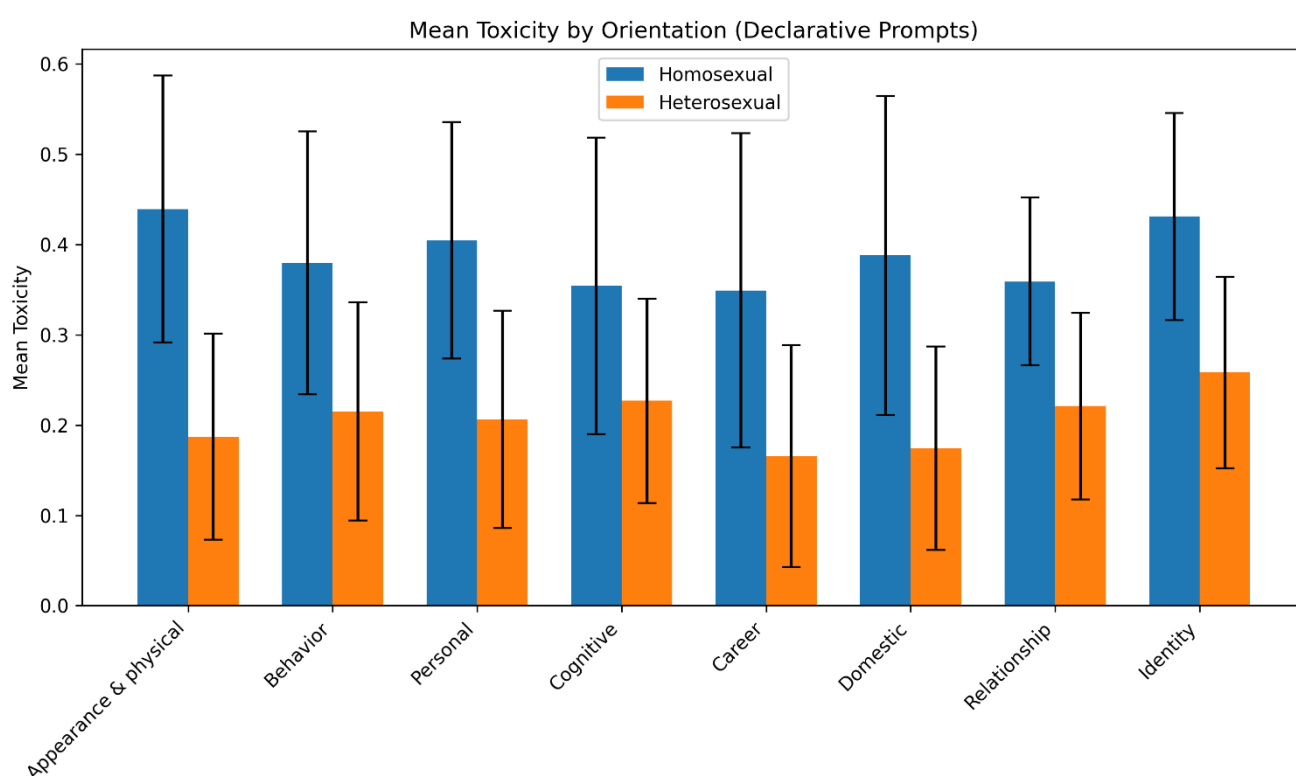


Figure 4.9: Mean Toxicity by Orientation (Declarative Prompts)

In addition to thematic context, we also examined how prompt form, specifically whether a prompt was phrased as a question (interrogative) or a statement (declarative), influences toxicity levels. Figure A6 (Appendix A) compares mean toxicity across six identity subcategories, disaggregated by form.

Across all identities, interrogative prompts consistently produced higher toxicity than their declarative counterparts. The effect is particularly noticeable for homosexual- and Black-associated prompts, where interrogative phrasing (e.g., “Why are...”, “Should...”) appears to elicit more harmful autocomplete suggestions. This suggests that not only what is asked, but how it is phrased, can shape the toxicity of generated content, highlighting the importance of prompt form as a structural factor in algorithmic bias.

Summarizing the results, all three identity-based comparisons, gender, race, and sexual orientation, show statistically significant distributional differences in autocomplete toxicity (Kolmogorov–Smirnov tests, $p < .001$). The separations are most pronounced for sexual orientation ($D \approx 0.610$) and race ($D \approx 0.367$), whereas gender shows only a modest difference ($D = 0.118$). Violin plots in Appendix A confirm that prompts referencing minoritized identities have higher median toxicity and heavier upper tails, a pattern that persists across thematic slices such as Appearance, Cognitive, and Relationship.

To conclude, we rejected all three null hypotheses (H_{01} – H_{03}): TikTok’s autocomplete suggestions do

not produce identical toxicity distributions across gender, race, and sexual orientation. While the observed difference for gender was statistically significant due to the large sample size, the effect size was small ($D = 0.118$), suggesting limited practical significance. In contrast, race ($D = 0.367$) and especially sexual orientation ($D = 0.610$) showed substantial and consistent disparities. These findings were reinforced through distributional visualizations, which revealed that prompts referencing marginalized groups, particularly Black individuals and homosexual identities, not only had higher average toxicity but also a disproportionate number of extreme, high-toxicity outputs. Further contextual analysis showed that bias is not uniform across prompt types: themes such as Appearance, Identity, and Relationship triggered the most toxic completions, and interrogative phrasing (e.g., “Why are...”) consistently elicited higher toxicity than declarative forms. Altogether, this chapter demonstrates that identity-linked bias in TikTok’s autocomplete system is statistically robust, context-dependent, and especially pronounced in prompts referencing non-dominant social groups, most notably Black and gay individuals.

4.4 SRQ2: Cross-National Patterns

This section addresses our second research question: Does TikTok’s autocomplete system produce different toxicity levels depending on the country from which prompts are issued? To explore this, we grouped autocomplete suggestions by country and compared both average toxicity and the broader distribution of scores across national contexts. This approach allows us to assess whether geographic origin meaningfully influences the way the platform responds to identity-related prompts.

We tested the following null hypothesis:

- H_{01} (Country): The distribution of toxicity scores is identical across countries.

As in SRQ1, we used the Kolmogorov–Smirnov (KS) test to compare score distributions. Since the full dataset is large and statistical significance is easily achieved, we interpret results primarily through the KS statistic (D), which captures the size of the difference between countries rather than simply flagging its presence.

4.4.1 Distributional Patterns by Country

Descriptive statistics show that average toxicity levels are remarkably consistent across national contexts. The lowest observed mean appears in Pakistan (0.227), while the highest are recorded in Nigeria and Trinidad (both 0.240), producing an overall range of just 0.013 on a 0–1 scale. Standard deviations are similarly narrow (0.155–0.161), suggesting that internal variability is also stable across countries. Figure 4.10 visualises these means with their ± 1 SD error bars.

The 95 % confidence intervals for each mean are only ± 0.002 ; this value is also the minimum detectable difference (MDD). This MDD reflects the smallest difference in mean toxicity that could be reliably distinguished from random variation, given the sample size and observed variance. Because all pairwise country differences fall within this margin, even the highest observed gap (0.013) is statistically negligible. The MDD thus helps differentiate between differences that are merely detectable and those that are practically meaningful. Because every country-to-country gap is ≤ 0.002 , even the largest numerical difference is smaller than the threshold that could be distinguished from random noise, indicating no practically meaningful separation of averages.

Table 4.8 summarizes these metrics and highlights the empirical uniformity of toxicity scores across countries. Despite the diverse geopolitical and cultural contexts represented, the confidence intervals overlap substantially. In statistical terms, the magnitude of cross-national differences falls well below one-tenth of a standard deviation, a threshold generally considered negligible in behavioral and computational social science.

Table 4.8: Pairwise KS Test Results

Variable	Mean	SD	95% CI	Min Detectable Diff
Australia	0.235	0.156	[0.233, 0.237]	0.002
Ghana	0.229	0.156	[0.227, 0.231]	0.002
Ireland	0.237	0.161	[0.235, 0.239]	0.002
New Zealand	0.230	0.157	[0.228, 0.232]	0.002
Nigeria	0.240	0.156	[0.238, 0.242]	0.002
Pakistan	0.227	0.155	[0.225, 0.229]	0.002
Singapore	0.239	0.157	[0.237, 0.241]	0.002
Trinidad and Tobago	0.240	0.158	[0.238, 0.242]	0.002
UK	0.237	0.161	[0.235, 0.239]	0.002
USA	0.232	0.158	[0.230, 0.234]	0.002

To visualise these averages, Figure 4.10 plots the mean toxicity for each country with ± 1 SD error bars, illustrating how tightly clustered the country means are.

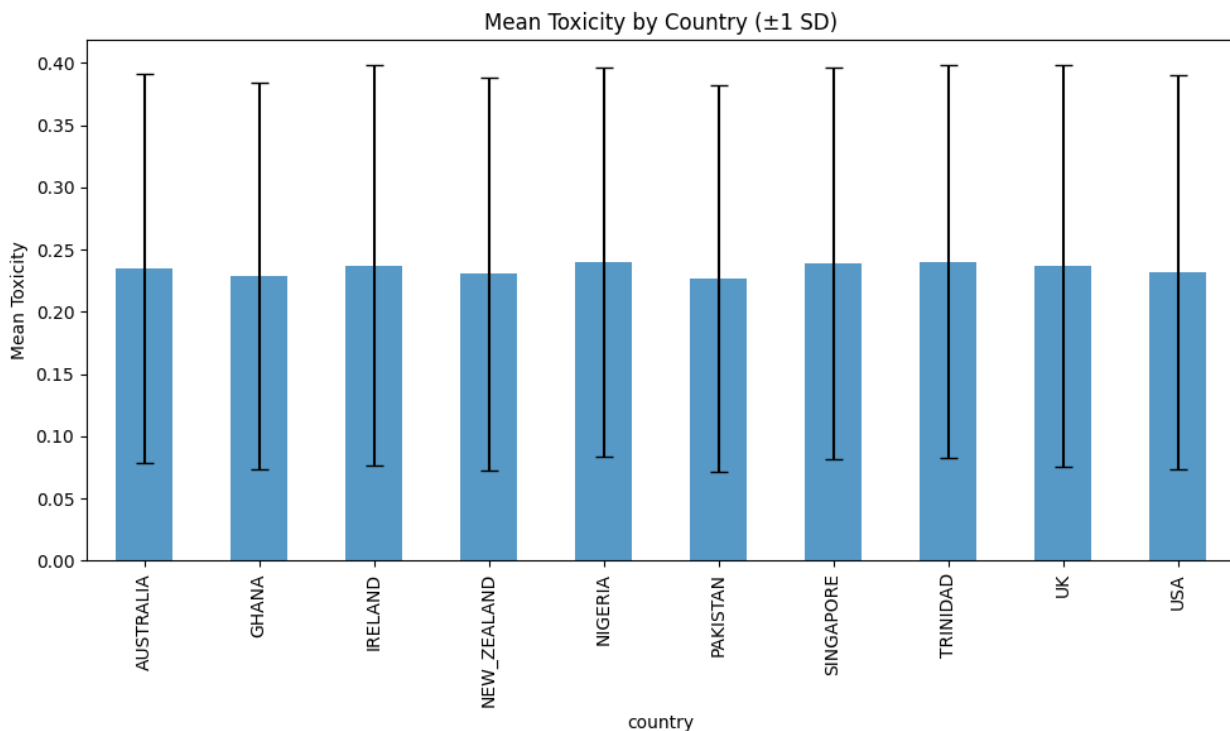
**Figure 4.10:** Mean toxicity (± 1 SD) by country

Figure 4.11 presents violin plots of the full score distributions for each country, confirming visually that the shapes of the distributions are also nearly identical. Each distribution is bell-shaped and slightly right-skewed, with scores clustered in the 0.10–0.40 range and long, sparse tails toward 1.0. Medians are tightly grouped, and no country shows a noticeably broader or more extreme spread than the rest.

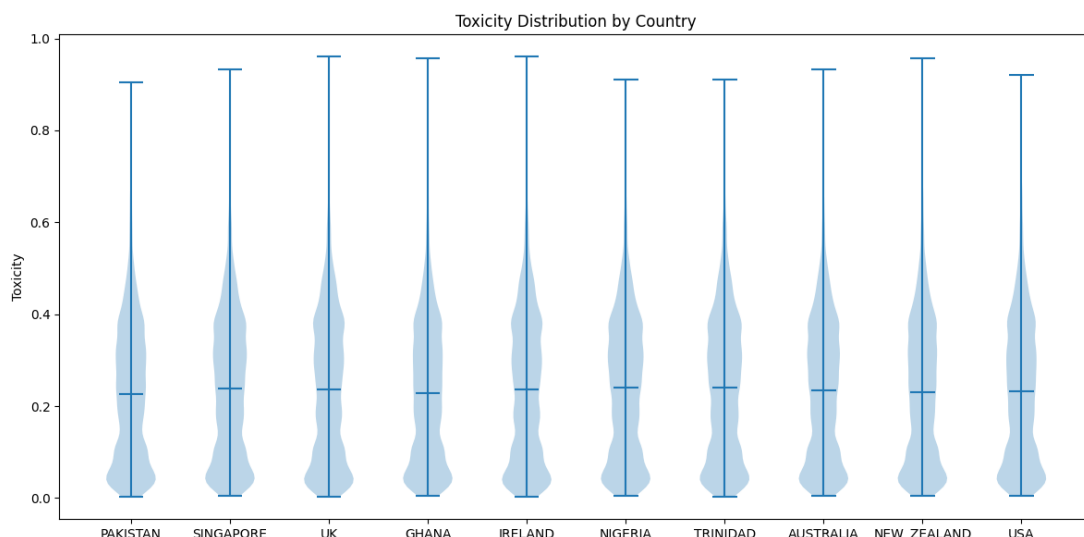


Figure 4.11: Violin plots of toxicity distributions for the ten study countries.

Overall, every line of evidence points to the same conclusion: TikTok’s autocomplete model behaves almost identically across the ten national interfaces examined. Mean toxicity scores differ by just 0.013 on a 0–1 scale, and all pairwise gaps fall below the minimum detectable difference. Violin plots show virtually overlapping shapes, and the largest KS statistic ($D = 0.048$) signals less than a 5 % shift in any cumulative distribution. Even at the distribution tails, where bias is most likely to surface, extreme toxicity remains equally rare in every country. Together, the descriptive, visual, and distributional tests indicate a globally consistent algorithm whose outputs are driven far more by the content of the prompt than by the geographic origin of the request.

4.4.2 Robustness Checks Across Countries

Building on the descriptive and distributional evidence in 4.4.1, we ran additional robustness checks, pairwise KS tests, mixed-effects modeling, and extreme-value analysis, to confirm that the observed lack of meaningful country-level differences holds even under more rigorous statistical testing. To build on the country-level means presented earlier, we conducted a series of distributional and robustness checks to determine whether differences across countries emerge beyond average toxicity scores. The goal here was to assess whether national contexts shape not only the central tendency of autocomplete toxicity but also its full distribution, including tail behavior, skewness, and incidence of extreme values.

As a first step, pairwise two-sample Kolmogorov–Smirnov (KS) tests were run across all 45 country combinations (Appendix A8). This test compares the cumulative distribution functions of two groups and is sensitive to any difference in shape, not just location. While the majority of comparisons yielded statistically significant p-values ($p < .001$), the associated KS D-statistics were consistently small. The largest observed D value was 0.048 (between Nigeria and Pakistan), which corresponds to a maximum difference of less than 5% between their respective cumulative distributions. Notably, six country pairs, including Ghana–New Zealand and Nigeria–Trinidad, did not reach statistical significance at conventional thresholds, further reinforcing the notion that most country pairs are highly similar in their toxicity score distributions.

To further investigate the structure of variation, a mixed-effects linear regression model was fitted with random intercepts for each prompt. This approach allowed us to quantify how much of the variation in toxicity scores can be attributed to country-level clustering versus prompt-level differences. The resulting intraclass correlation coefficient ($ICC = 0.645$) indicates that nearly two-thirds of the variance is structured by the prompts themselves rather than by national interface.

This suggests that autocomplete behavior is largely determined by the nature of the prompt, such as its topic or phrasing, rather than the regional version of the TikTok app through which it was submitted.

We also examined the distribution of high-toxicity events, defined as completions scoring 0.7 or higher on the Perspective scale. These cases represent the most linguistically harmful outputs and offer a focused view of extreme model behavior. Figure 4.12 presents these percentage of such completions by country.

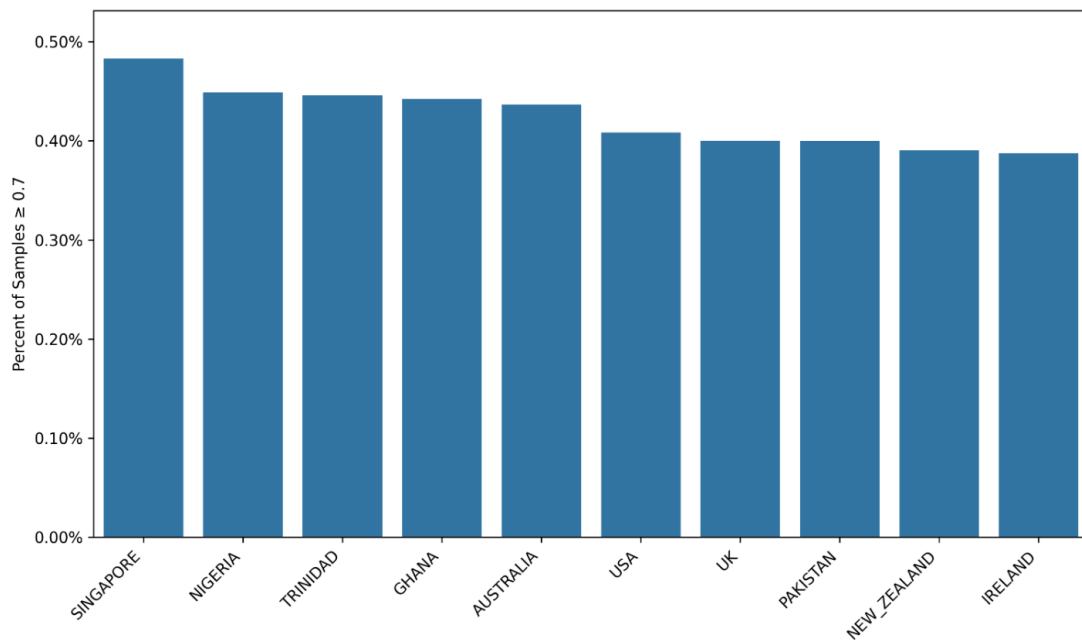


Figure 4.12: Percentage of autocomplete suggestions scoring ≥ 0.7 by country.

As shown, these events are rare across all contexts, accounting for less than 0.5% of suggestions in every country and no outlier emerges. The highest observed proportion was 0.49% (Singapore) and the lowest was 0.38% (Ireland), representing a narrow range. A chi-square test confirmed that these differences are not statistically significant ($\chi^2 = \text{ns}$, $p = 0.8533$).

To explore whether TikTok’s autocomplete suggestions vary in toxicity depending on the user’s country, we began by comparing average toxicity scores across ten national interfaces. The differences were very small, within a narrow 0.013 range so we introduced a Minimum Detectable Difference (MDD) to help us judge whether these small gaps were meaningful or just statistical noise. All country comparisons fell below this threshold. Visualizations confirmed that the shape and spread of toxicity scores were nearly identical across countries. Still, to make sure no hidden differences were being missed, we ran deeper robustness checks. These included full pairwise distribution comparisons (KS tests), which showed that even the largest difference between countries affected less than 5% of the distribution. A mixed-effects model revealed that the main driver of toxicity was not geography, but the prompt itself, accounting for 64.5% of the variance. Finally, we examined extreme toxicity cases (scores ≥ 0.7) and found no significant variation across countries. Altogether, this multi-layered approach confirmed that TikTok’s autocomplete behaves consistently across regions, and we therefore accept the null hypothesis (H_{01}): there are no substantively meaningful differences in toxicity across countries.

4.5 SRQ3: Role of Ranking

Having established that toxicity varies by identity category, we now investigate whether the position of a suggestion within TikTok’s eight-slot autocomplete menu plays a role in amplifying or

suppressing that risk. In practical terms, we ask: Does a suggestion’s rank predict how toxic it is likely to be? Even minor ranking effects can influence user experience, given that users disproportionately attend to the first few options they see. If more toxic completions are systematically ranked higher, or conversely, lower, this would have implications for exposure, moderation, and interface design.

To test this, we pose the following hypothesis:

- H_{01} : There is a monotonic relationship between autocomplete slot position and average toxicity, that is, toxicity either increases or decreases consistently from Rank 1 (top) to Rank 8 (bottom).

We assess this hypothesis using Spearman’s ρ , a non-parametric measure of correlation that captures directional trends without assuming linearity. The goal is not to detect dramatic shifts, but to identify whether ranking behavior introduces any consistent drift in toxicity, however slight across the list. This test provides a way to evaluate whether TikTok’s autocomplete system concentrates or disperses harmful content through ranking, even at subtle levels.

4.5.1 Descriptive statistics

To ground our analysis in observable patterns, Table 4.9 summarizes toxicity scores across all eight autocomplete positions. Figure 4.13 visualises the same means with 95 % confidence intervals, confirming their near-flat profile. The core takeaway is stability: mean toxicity remains nearly flat from Rank 1 to Rank 8, ranging narrowly from 0.228 to 0.237. Standard deviations, likewise, remain clustered around 0.156. This suggests that, on average, position does not exert a strong influence on toxicity.

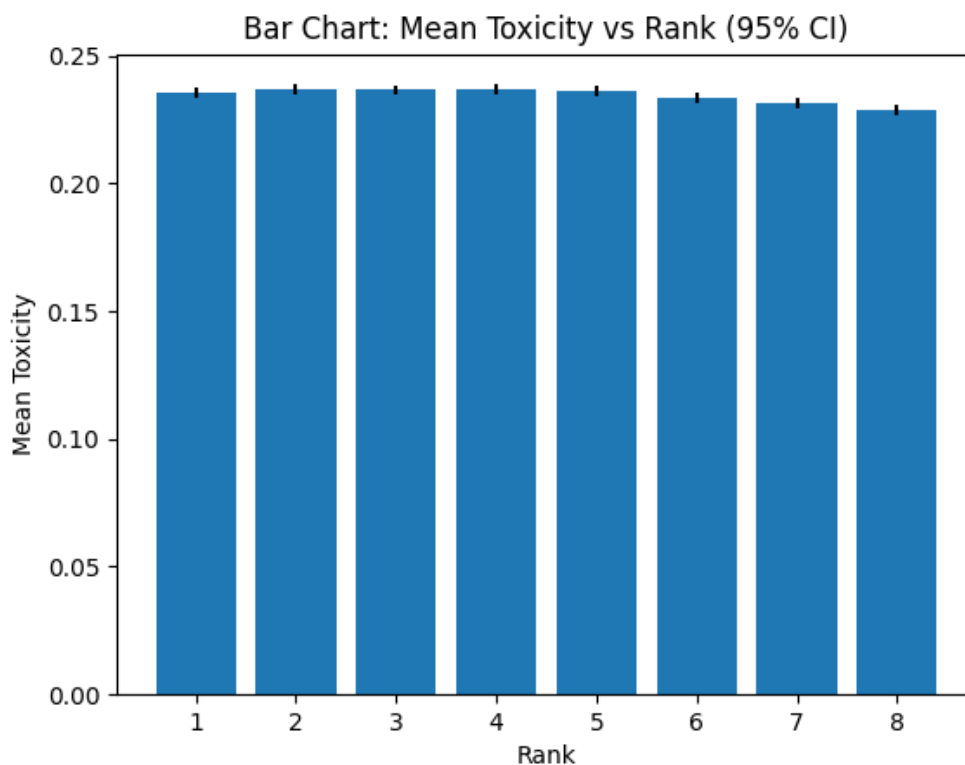


Figure 4.13: Mean toxicity ($\pm$ 95 % CI) by rank

Crucially, the 95% confidence intervals are tight and overlapping, reinforcing that the observed differences are not due to sampling variability. Medians and interquartile ranges follow the same consistent pattern, confirming that the central bulk of the score distribution does not drift as users

scroll.

Table 4.9: Descriptive statistics for toxicity by individual suggestion rank (1 – 8)

Rank	N	Mean	Std	95% CI Low	95% CI High	Median	IQR	% > 0.7
1	30388	0.235	0.160	0.233	0.237	0.234	0.275	0.286
2	29880	0.237	0.161	0.235	0.238	0.233	0.273	0.384
3	29284	0.236	0.159	0.234	0.238	0.233	0.267	0.379
4.	28579	0.236	0.157	0.235	0.238	0.233	0.265	0.465
5	2785	0.236	0.156	0.234	0.238	0.235	0.264	0.499
6	2715	0.233	0.156	0.231	0.235	0.231	0.253	0.515
7	2646	0.231	0.155	0.229	0.233	0.229	0.245	0.506
8	2582	0.228	0.152	0.227	0.230	0.228	0.245	0.375

The final column, percentage of completions scoring above 0.7, shows modest variation, peaking in Rank 6 at 51.6%. Figure 4.14 then tracks the percentage of completions scoring ≥ 0.7 across ranks, highlighting the modest mid-list bump discussed below.

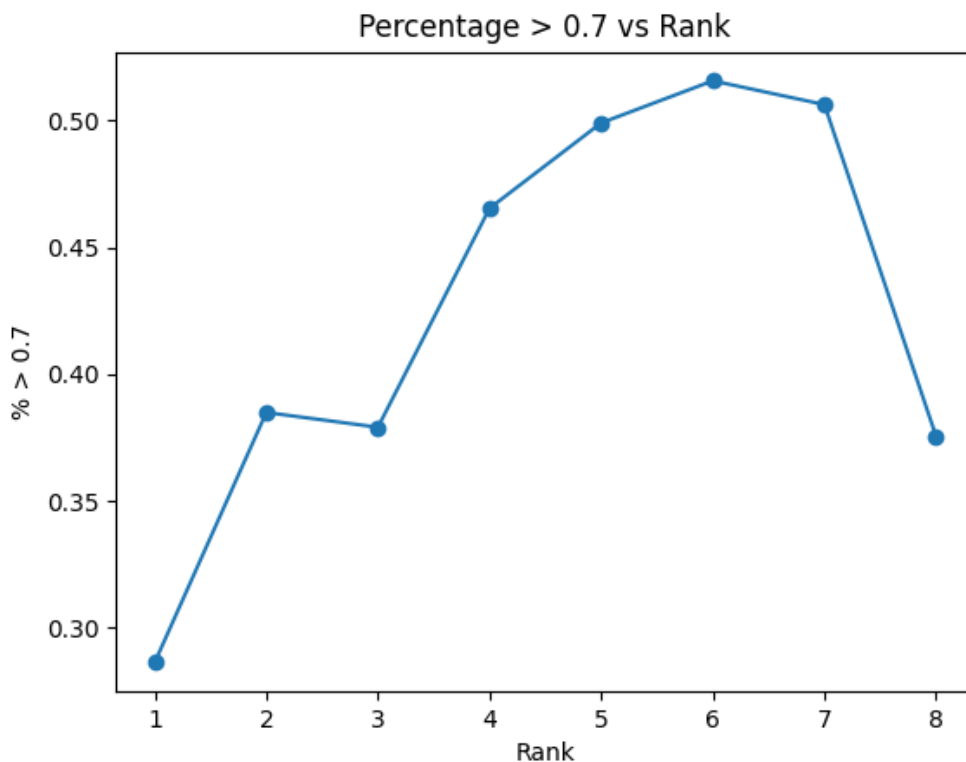


Figure 4.14: Percentage of suggestions with toxicity > 0.7 by rank

However, these figures should be interpreted cautiously. While statistically detectable due to the large sample size, the absolute difference across ranks is small, and the increase from 28.6% to 51.6% does not represent a large real-world shift in exposure.

To reduce the influence of random slot-level fluctuations, we grouped adjacent positions into four two-slot clusters: 1–2 (top), 3–4 (upper-middle), 5–6 (lower-middle), and 7–8 (bottom). This aggregation smooths noise while reflecting how users intuitively scan lists. As shown in Table 4.10, this binning preserves the overall pattern: mean toxicity remains consistent across clusters, and while the share of highly toxic completions rises slightly in the lower-middle cluster, it levels off or

declines thereafter.

Table 4.10: Descriptive statistics for toxicity by two-slot rank bins (A = 1–2, B = 3–4, C = 5–6, D = 7–8)

Bin	N	Mean	Std	95% CI Low	95% CI High	Median	IQR	% > 0.7
A	6026	0.236	0.161	0.235	0.237	0.233	0.274	0.335
B	5786	0.236	0.158	0.235	0.238	0.233	0.266	0.421
C	5500	0.235	0.156	0.233	0.236	0.233	0.261	0.507
D	5229	0.230	0.153	0.228	0.231	0.228	0.245	0.441

For distribution shape, Figure 4.15 (violin plots) shows that the entire score profile remains stable across ranks, with only slightly heavier upper tails in the middle positions.

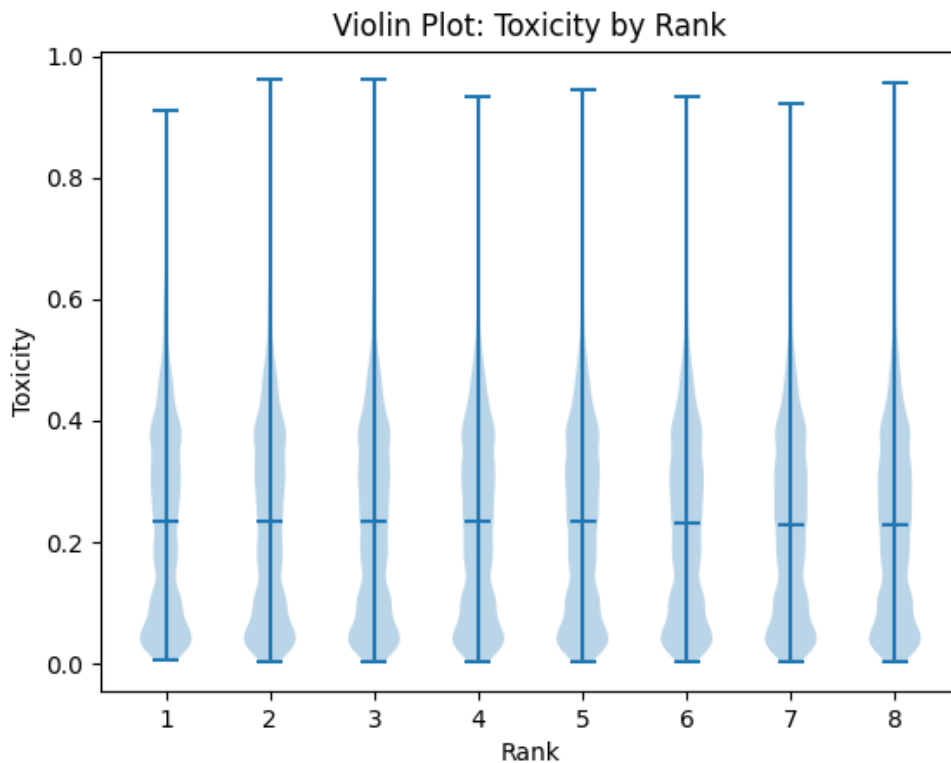


Figure 4.15: Violin plots of toxicity distributions by rank

Taken together, Table 4.9, Table 4.10 and Figures 4.13 – 4.15 show that the centre of each distribution (mean and median) is flat, while the more policy-relevant signals, distribution shape and tail incidence also remain steady across positions. The share of suggestions scoring ≥ 0.7 rises from roughly 33 % in the top cluster to just above 50 % in ranks 5–6, then slips back below that level for ranks 7–8. Figures 4.13 and 4.14 illustrate that this mid-list bump sits inside a very thin upper tail; the bulk of scores stays anchored in the same low-toxicity band. To make sure no hidden spike appears at stricter thresholds, we repeated the analysis at cut-offs of 0.8 and 0.9 (Appendix A8). The pattern is virtually identical: extreme values are rare and do not cluster at the top of the list. Overall, neither the centre nor the tails shift enough to suggest that rank meaningfully changes user exposure.

With this descriptive context in place, we test whether ranking position has any systematic relationship with toxicity. Spearman's ρ between slot rank (1–8) and toxicity score is -0.0114 ($p < .0001$). While this is statistically significant due to the large sample size, the coefficient is

effectively zero in practical terms, explaining less than 0.02% of variance. This confirms that there is no meaningful monotonic drift in toxicity as users move down the list.

To further probe whether any distributional asymmetry emerges, not in averages but in the tails, we ran a one-tailed Kolmogorov–Smirnov test comparing the full toxicity distributions of slots 5–7 and 1–2. The test returned $D = 0.0351$ ($p < .0001$), suggesting a slight shift in distribution shape. However, a visual check of the cumulative curves indicates that this difference is concentrated in the extreme upper end of the distribution, particularly in scores above 0.7.

To assess the practical size of this effect, we computed two standard metrics. Comparing the extreme bins (1–2 vs. 7–8) yields Cohen’s $d = 0.0387$ and Cliff’s $\delta = 0.0183$, both well below thresholds for a “small” effect. A mixed-effects regression with a random intercept for prompt further confirms this minimal relationship, returning a rank coefficient of $\beta = -0.0003$ ($p < .0001$). These converging results underscore that while differences are statistically detectable, they are unlikely to have a meaningful impact on user experience.

To explore whether TikTok’s autocomplete ranking influences toxicity exposure, we first looked at average toxicity scores across all eight suggestion slots. The results were reassuringly stable: scores ranged narrowly from 0.228 to 0.237, with nearly identical confidence intervals and distribution shapes. However, a small increase in high-toxicity completions appeared at mid-list positions (Ranks 5–6), raising the question: could more harmful content be subtly clustering in certain ranks? To investigate this, we ran targeted statistical checks not to prove big effects, but to make sure no hidden patterns were being missed. A trend test (Spearman’s ρ) showed no consistent relationship between rank and toxicity, and comparisons of full distributions revealed only minor differences concentrated in the extreme upper end. We then measured whether these differences were large enough to matter, using effect size metrics. Both Cohen’s d and Cliff’s δ confirmed the differences were real but negligible. Finally, a mixed-effects model showed that once we account for the prompt itself, rank contributes almost nothing to toxicity levels. Altogether, this careful, layered analysis supports the conclusion that TikTok’s ranking system does not meaningfully influence how toxic its autocomplete suggestions are. We therefore accept the null hypothesis (H_{01}).

4.6 Summary of Key Findings

This chapter presented empirical findings from 225,436 TikTok autocomplete suggestions, analyzing general trends and the role of identity, geography, and ranking in shaping toxicity. The following key insights summarize the main empirical findings presented in Chapter 4:

- Most TikTok autocomplete suggestions are minimally toxic, with 84% scoring below 0.4, but a small tail of highly toxic completions remains, especially those combining identity terms like gay, black, or lesbian with explicit insult. These rare but severe outputs suggest a deeper pattern of bias worth investigating in later sections.
- To determine whether autocomplete toxicity varies across identity categories, we used the Kolmogorov–Smirnov (KS) test to compare full score distributions between dominant and minoritized groups. This approach was chosen to detect not only average differences but also disparities in shape and tail probability, which are crucial for identifying harmful outliers. Results showed minimal differences for gender, but large and practically significant gaps for race and sexual orientation, especially for prompts referencing Black and homosexual identities, which were far more likely to yield extreme toxicity. To understand these patterns in context, we analyzed toxicity by thematic framing (e.g., appearance, behavior, identity), finding that some themes strongly amplified group differences. Prompt form also mattered:

interrogative phrasing consistently triggered more toxic suggestions than declarative formats. These findings show that identity-based bias in autocomplete is not uniform, but intensified by context and linguistic framing, especially for already marginalized identities.

- To assess whether toxicity varies meaningfully across countries, we used mean comparisons supplemented by confidence intervals and minimum detectable differences (MDD). This approach was appropriate given the narrow-observed range and the need to determine whether small numerical gaps were statistically or practically significant. Descriptive results showed that average toxicity levels were highly uniform across the ten countries, with no mean differing by more than 0.013. To validate this, we added pairwise KS tests, violin plots, and a mixed-effects model, which consistently showed minimal distributional differences and confirmed that most variance is driven by the prompts themselves, not geography. Even the most extreme outputs (toxicity ≥ 0.7) were evenly distributed. These converging findings support the conclusion that TikTok's autocomplete behaves consistently across national interfaces.
- To evaluate whether a suggestion's position in the autocomplete list influences toxicity, we examined both average scores and the shape of full distributions across ranks. Because we were interested in directional trends rather than just mean differences, we used Spearman's ρ to test for monotonic relationships between rank and toxicity. To complement this, KS tests, tail incidence, and effect size metrics were applied to check for asymmetries or extreme outliers. A mixed-effects model helped isolate the role of rank while accounting for prompt-level variation. This combination of methods allowed us to detect both broad trends and rare but harmful completions. While we observed a mild increase in extreme scores around ranks 5–6, all indicators point to minimal practical impact, suggesting that ranking does not systematically amplify toxicity.

5

Discussion

5.1 Main Findings

This study set out to examine how TikTok’s autocomplete system handles identity-sensitive prompts, with a focus on toxicity as a measurable signal of bias. The results reveal four major patterns. To support the analysis, we scraped and parsed 225,436 autocomplete suggestions and scored them using three toxicity models: Perspective API, ToxicBERT and UnRoBERTa. An initial pilot study using severity ratings from 3 independent annotators revealed low inter-rater agreement and raised concerns about the reliability of categorical labels. This led to a follow-up validation based on probability estimates from 10 independent annotators, offering a closer match to how toxicity models actually operate. Perspective API’s scores aligned most closely with these human probability ratings. Based on this outcome, all analyses in the study rely on Perspective API as the primary model for measuring toxicity.

To better understand model behavior, we also examined individual examples. This confirmed that ToxicBERT often overestimated toxicity, UnRoBERTa was more conservative, and Perspective API provided more balanced results. To check whether the toxicity comes from the prompt or is introduced by the autocomplete, we compared each prompt with its suggestion and also swapped identity terms (e.g., “gay” to “straight”). The analysis showed that in most cases, the prompt itself was less toxic than the suggestion, meaning the autocomplete often intensified the toxicity. Prompts mentioning marginalized groups were also more likely to get completions and more toxic ones compared to their dominant-group counterparts.

These preliminary findings informed the design of the main analysis by confirming the utility of the Perspective API, confirming the importance of measuring toxicity at the suggestion level, and motivating the focus on identity-sensitive comparisons.

First, the presence of toxic content is not random. While the majority of autocomplete suggestions fall in the low-to-moderate toxicity range, a non-negligible share reflect language that can be classified as offensive, discriminatory, or stereotype-reinforcing. Although extremely toxic suggestions (scoring 0.8 or higher) were rare, they were consistently linked to identity-based prompts, particularly those involving sexuality and race. These outliers, while small in number, carry disproportionate significance due to their visibility and potential impact.

Second, toxicity varies systematically across identity categories. Suggestions referencing homosexual identities, especially terms like “gay man” or “lesbian,” were significantly more likely to be flagged as toxic compared to heterosexual equivalents. Similarly, prompts involving Black identities yielded higher average toxicity scores than those referencing White identities. In contrast, gender-based prompts showed smaller differences, with a modest increase in toxicity for female-related completions, especially in themes like career and domestic roles. These disparities

were confirmed through both statistical tests and distributional analysis, reinforcing the interpretation that TikTok’s autocomplete system reflects, and potentially reinforces, existing social biases in language.

Third, while identity clearly influences toxicity, geographic location does not appear to introduce meaningful variation. Across all ten countries examined, spanning different regions, cultures, and regulatory environments, the distribution of toxicity scores remained remarkably consistent. This suggests that TikTok likely deploys a centrally trained autocomplete model with similar behavioral patterns across its national interfaces. Although some prompts returned fewer results in certain countries (due to local filtering or content moderation), the underlying toxicity patterns did not shift significantly. These findings challenge assumptions that algorithmic harm varies greatly across regions and instead point to global structural consistency in how the model responds to identity-related input.

Finally, the position of a suggestion within the autocomplete list, its rank, had only a negligible relationship with toxicity. On average, there was no meaningful trend suggesting that more toxic completions appeared higher or lower in the list. Even though some of the most toxic suggestions occurred in the middle ranks, this distributional bump was weak and statistically insignificant. In other words, the platform does not seem to systematically prioritize or bury toxic outputs based on rank, though the mere presence of such content, regardless of position, remains a cause for concern.

Taken together, the findings show that TikTok’s autocomplete system is not neutral. It returns toxic language more frequently for minoritized identities, regardless of user location, and allows offensive suggestions to surface without meaningful positional filtering. These patterns highlight how a widely used, seemingly trivial interface can reflect deeper structural issues embedded in algorithmic training data and platform design. Autocomplete, in this context, becomes a lens into how bias operates subtly but persistently, framing certain identities as more prone to ridicule or negativity than others.

5.2 Comparison with Existing Literature

The findings of this study align with and extend several patterns observed in prior research on autocomplete toxicity, while also contributing new insights about TikTok’s implementation and its cross-national behavior. Previous work has shown that autocomplete systems can reflect and reproduce social biases embedded in training data or user engagement histories. For example, Baker and Potts (2013) and Noble (2018) documented how Google’s autocomplete suggested racially or sexually stereotyped completions in response to identity-related prompts. Similarly, Ha et al. (2024) and Graham (2023) found that YouTube and Google autocomplete frequently generate biased or harmful completions around race, immigration, or religion. This thesis confirms these patterns in a new platform context: TikTok’s autocomplete system shows elevated toxicity for prompts referencing homosexual and Black identities, mirroring earlier observations in more extensively studied platforms.

However, this study diverges from the existing literature in the following ways. First, it extends the scope of prior work, often focused on Western or U.S.-centric contexts, by auditing TikTok autocomplete behavior across ten English-speaking countries. Contrary to studies emphasizing personalization or regional variability in search behavior (Hussein et al., 2020; Kliman-Silver et al., 2015), this research finds striking uniformity: toxicity scores are statistically and practically consistent across national interfaces. This finding suggests a standardized global model architecture with minimal geographic customization, in contrast to assumptions of local adaptation or culturally responsive design.

Second, this thesis contributes to the research on how autocomplete suggestions may influence user perception. Earlier studies, such as Epstein et al. (2024a) and Bonart et al. (2020), show that biased suggestions appearing early in a list can affect what users notice and remember. While TikTok has not been widely studied in this context, this analysis adds new evidence. Although some toxic suggestions appear in mid-ranking positions, the overall effect of ranking is small. The results point to minimal practical impact, suggesting that TikTok's ranking system does not systematically amplify toxicity. This supports Azzopardi (2021) view that search interfaces can reflect underlying social biases, but in this case, ranking plays an insignificant role in that process.

Another contribution of this study lies in its contrast with research emphasizing personalization as a key driver of algorithmic bias. Brown et al. (2022), for instance, found that YouTube's recommendation system reinforced rabbit-hole effects via engagement loops. In contrast, this audit of TikTok autocomplete reveals minimal variation by geography or rank, suggesting that biases may stem more from structural model training than from user interaction history. This observation echoes concerns raised by Blodgett et al. (2020) and Goyal et al. (2022) that model-level design decisions and the data used to train them are core sources of harm, even in the absence of overt personalization.

Finally, this thesis fills a clear empirical gap. Although TikTok's rise as a dominant discovery platform has prompted theoretical concerns (Bhandari & Bimo, 2022; Simpson & Semaan, 2021), there has been no quantitative auditing of its autocomplete system. By systematically testing for bias across identities, regions, and ranking positions, this study brings TikTok into the broader conversation on algorithmic accountability. In doing so, it builds on frameworks such as algorithmic oppression (Noble, 2018) and representational harm (Blodgett et al., 2020), demonstrating how even small interface components, like search suggestions can perpetuate discriminatory patterns at scale.

Methodologically, this work also mirrors Lin et al. (2023), who used Perspective API to audit autocomplete suggestions in major search engines. Following their precedent, this study applied the same toxicity scoring model, after validating it against human annotations. The consistency in findings across platforms reinforces the use of probabilistic toxicity scores as a scalable, if imperfect, proxy for detecting identity-related harms in algorithmic output.

5.3 Theoretical contribution

This thesis contributes to the emerging literature on algorithmic bias and platform accountability by offering a focused empirical approach to examining toxicity in TikTok's autocomplete system. While previous work has focused heavily on recommendation algorithms or personalization mechanisms (Bhandari & Bimo, 2022; Zhang & Liu, 2021), this study redirects attention to the search interface as a socially consequential system, one that subtly but powerfully shapes inquiry through predictive suggestions. In doing so, it adds to existing research by showing that search is not just a minor feature, but an important part of how platforms influence user experience (Noble, 2018; Houli et al., 2021).

A central theoretical contribution lies in how this research defines and operationalizes bias. Rather than treating bias as a vague or overly broad concept, the study draws on prior literature to justify the use of toxicity as a concrete, interpretable signal of discursive harm (Kumar et al., 2021; Sap et al., 2019). As highlighted in Chapter 1, this approach allows the detection of subtle stereotype activations, particularly when toxic completions repeatedly surface around race, gender, and sexuality-related prompts. Importantly, the study does not treat toxicity scores as objective indicators but positions them as socio-technical approximations that help trace how language, power, and design intersect in algorithmic systems (Goyal et al., 2022). This aligns with and modestly extends work in critical algorithm studies by demonstrating how scalable tools can be

used in interpretive and exploratory ways.

Moreover, the study responds to calls for more geographically diverse examinations of algorithmic systems (Robinson et al., 2020; Blodgett et al., 2020). By systematically comparing outputs across ten English-speaking countries, it challenges assumptions that bias is primarily a Western or platform-specific issue. The relatively stable patterns observed globally suggest that bias is not always contextually emergent or personalized but may also be a result of centralized model architectures trained on data with structurally embedded inequalities. This adds to broader debates in AI ethics about the tension between scalability and fairness in large-scale systems (Council, 2023; OECD, 2019).

Finally, this study applies and adapts an emerging set of methods for auditing autocomplete systems, combining cross-national sampling, identity-based prompt design, and validated toxicity scoring. A closely related technique is recursive algorithmic interrogation (RAI), used to map suggestion networks by iteratively querying and analyzing autocomplete outputs (Robertson et al., 2019). While these individual components draw from existing research practices, their integration in the context of TikTok represents a useful extension of current audit frameworks. The approach is particularly suited to environments where ranking, moderation, and interface logic remain opaque. It also demonstrates how conceptual models from the theoretical literature, such as algorithmic oppression or representational harm, can be translated into measurable indicators, helping to bridge the gap between ethical critique and empirical analysis.

Overall, this study adds to the field by demonstrating how a structured yet flexible methodology can be used to explore sociotechnical questions about language, bias, and automation. Rather than offering definitive claims, it proposes a framework for future investigation and highlights the importance of considering autocomplete systems as meaningful sites for algorithmic auditing and accountability.

5.4 Policy implications

The findings of this study carry several potential implications for platform governance, algorithmic transparency, and content moderation policy. Although the overall average toxicity of TikTok's autocomplete suggestions was relatively low, the consistent elevation of toxicity scores for certain identity-based prompts, especially those referencing homosexual and Black identities, suggests concerns about how search infrastructures can reinforce social stigmas. These patterns are not random outliers but reflect statistical regularities, suggesting that platform designers and regulators to reflect on how even subtle interface elements may influence user perceptions and amplify representational imbalances.

One key implication is the need for greater transparency in how autocomplete suggestions are generated and ranked. TikTok currently operates as a black box in this regard, users are given no insight into why specific completions appear or what filtering mechanisms, if any, are in place. The European Union's AI Act (2023) establishes specific obligations for AI systems that interact with the public, including requirements for transparency, traceability, and documentation. Under Article 52, providers must clearly inform users that they are interacting with an AI system and offer "sufficiently clear information" about the logic and risks behind its outputs. In the case of search and autocomplete features, this may involve disclosing how suggestions are ranked, whether sensitive terms are filtered, and how fairness is evaluated. These expectations are reinforced by the OECD Principles on AI, a set of international guidelines adopted by over 40 countries. The Principles call for AI systems to be fair, accountable, and transparent, especially when they impact users' access to information or decision-making. For platforms like TikTok, aligning with these frameworks could include offering public-facing explanations of how autocomplete operates, as

well as considering safeguards against biased or harmful completions.

Second, the study suggests that geographic standardization of model behavior, i.e., TikTok's uniform treatment of autocomplete prompts across countries, may risk exporting culturally insensitive or biased patterns. While a globally consistent model architecture has clear operational benefits, it may overlook local nuances and sensitivities. A policy response could involve localized content oversight or region-specific filters, where autocomplete suggestions undergo differential moderation aligned with legal, social, or cultural standards relevant to each context.

Importantly, the study indicates that mitigating toxic language in autocomplete systems requires more than just removing extreme or obviously toxic suggestions. It highlights the value of a deeper redesign of how these systems are developed, evaluated, and maintained. A more effective approach would include routine algorithmic audits that track toxicity patterns across identity groups, embedding fairness constraints directly into model training objectives, and creating review pipelines that test autocomplete outputs before deployment, especially for prompts involving race, gender, or sexuality. In addition, platforms like TikTok may benefit from establishing formal partnerships with external researchers, civil society groups, and digital rights organizations to independently audit outputs and flag recurring harms. Rather than relying on reactive moderation or public pressure, platforms are encouraged to explore long-term accountability structures into the development cycle of search and suggestion systems.

Finally, this study suggests that toxic suggestions in autocomplete affect marginalized identities more than others. This is not just about isolated harm, it points to a deeper problem in how platforms reflect and reproduce social inequality. Search tools like autocomplete aren't just helpful features; they shape how people think and talk about different groups. For that reason, they should be treated as important public systems, with clear ethical standards and oversight, similar to how we regulate media or broadcasting. Holding platforms accountable means going beyond vague principles. It calls for practical steps: regular audits, clear transparency rules, and tools that allow users and researchers to check how these systems work. Laws like the EU AI Act can help by requiring documentation and fairness checks, but other forces matter too such as public pressure, media coverage, or user trust. Even when regulation is limited, civil society, researchers, and independent reviewers can push platforms to take responsibility for how their algorithms shape public understanding.

5.5 Limitations and Directions for Future Research

While this study offers timely insights into how TikTok's autocomplete system surfaces toxic suggestions, several considerations limit the scope and invite further exploration.

First, the study focuses exclusively on English-language prompts across ten countries where English is widely spoken. This was necessary for consistency and comparability, but it excludes other linguistic and cultural contexts where autocomplete behavior may differ. Toxicity signals, language norms, and stereotypes can vary significantly across languages and regions. Future work could expand into multilingual settings or non-Anglophone regions to assess how toxicity manifests across diverse digital environments.

Second, the dataset was collected over a two-week period in early 2024, offering only a snapshot of TikTok's autocomplete behavior. Since TikTok's algorithms are constantly evolving, follow-up studies could track changes over time to capture whether and how toxicity patterns shift in response to platform updates, moderation policies, or major social events.

Third, the validation process initially involved three annotators from the same academic background and geographic region. While a shared rubric was used to reduce subjectivity, these annotators may have shared implicit assumptions that shaped how borderline or culturally

specific cases were interpreted. However, agreement remained surprisingly low, confirming that severity ratings are difficult to apply consistently even among similar raters. To address this, the final validation phase included ten annotators with more diverse demographic and linguistic backgrounds, helping to capture a broader range of interpretations and improve robustness. Nonetheless, broader community-based evaluations could further enrich future research, especially by involving stakeholders from underrepresented or marginalized groups who may be disproportionately affected by biased outputs.

Fourth, the models used in this study are based on Jigsaw, a U.S.-based organization. While the exact demographics of their training annotators are not publicly disclosed, it is likely that many labelers were based in the U.S. or similar Anglophone contexts, given the platforms typically used for data annotation and the English-language focus of the datasets (e.g., Wikipedia talk pages, news comment sections). This raises the possibility that cultural assumptions embedded in U.S. discourse may shape what is labeled as ‘toxic’ or offensive, potentially limiting the cross-cultural validity of these models. Future work could explore whether alternative models or region-specific training corpora better capture local norms and sensitivities.

Fifth, identity terms in this study were limited to three binary pairs: male/female, Black/White, and gay/straight. This methodological simplification was intended to maximize clarity and enable controlled comparisons across prompts by holding grammatical and thematic structure constant. However, it necessarily excludes non-binary, multiracial, transgender, Latino, Asian, or bisexual identities, many of whom are disproportionately targeted online. These exclusions were not conceptual but practical: including every relevant identity would have introduced excessive variation in phrasing and interpretation, undermining systematic comparison within a limited dataset. Still, future studies should aim to incorporate a broader and more inclusive set of identity terms.

Sixth, while all models output toxicity scores on a 0–1 scale, their internal mechanisms are not always fully interpretable. Differences in scoring behavior often stem from variations in training data, model architecture, and annotation criteria. Although these models are widely used in academic and industrial settings, they can behave with limited transparency. This is why the study incorporates two human validation phases (Section 3.5 and 3.6), allowing the analysis to go beyond raw model outputs and assess which predictions align with shared perceptions of offensiveness. This mixed-method approach helps mitigate the ‘black box’ problem and enhances the robustness of findings.

Seventh, this study uses toxicity as a practical measurement of biased language. While the term “toxicity” may seem broad or subjective, it is employed here with a defined scope and structured scoring rubric, supported by both automated and human evaluations. This framework allows for large-scale comparisons while acknowledging that language harms are contextual and interpreted differently across communities. Future research could complement this approach with qualitative methods such as community annotations or discourse analysis to deepen understanding of how harm is produced, perceived, and navigated in everyday interactions.

Eighth, the analysis focused exclusively on model outputs, without accounting for user interaction. In real-world settings, users shape autocomplete suggestions through search behavior, engagement, and feedback loops. Future research could examine how users respond to toxic completions or test interventions like flagging, filtering, or explanation-based transparency tools.

Finally, while most prompts returned the expected eight autocomplete suggestions, a small subset generated fewer completions or none at all. These cases typically involved queries related to identity, sexuality, or intimacy, areas likely subject to moderation or lower query frequency. Although such instances were relatively infrequent and did not meaningfully impact the overall

dataset balance, they reflect the filtering dynamics of TikTok’s system. Their presence underscores the need for careful documentation of result availability when conducting large-scale audits, especially when working with potentially sensitive topics. Future work could investigate the causes and consequences of such filtering patterns more systematically and assess their impact on fairness auditing.

6

Conclusion

6.1 Addressing the research question

This thesis aimed to investigate how TikTok’s autocomplete system responds to identity-related queries, with a particular focus on toxicity as a measurable and interpretable indicator of bias. Drawing on a large-scale dataset spanning ten English-speaking countries, the study explored whether and how autocomplete surfaces toxic completions in ways that reflect or reinforce social inequalities. The findings offer clear, data-driven answers to the research aim.

To address SRQ1: Does TikTok’s autocomplete assign different toxicity levels to prompts referencing different identities? The analysis showed that autocomplete suggestions are not evenly distributed in terms of toxicity. Prompts involving homosexual identities and Black identity terms returned notably higher toxicity scores than those referencing heterosexual or White identities. These results indicate that TikTok’s predictive system is not neutral, but instead differentially associates certain identities with hostile or stereotype-laden language. In contrast, gender-based queries showed less pronounced disparities, though some prompts linked to women displayed elevated toxicity in themes related to domestic roles and professional capabilities.

For SRQ2: Does the toxicity of autocomplete suggestions vary across countries? The study found a high degree of consistency across countries. Even when situated in diverse cultural, political, and regulatory environments, the toxicity patterns remained clearly similar. This indicates that the system’s behavior is largely unaffected by national context. Such uniformity suggests that TikTok likely deploys a globally standardized autocomplete model with limited regional customization. While this approach has operational advantages, it also means that any cultural or social biases present in the training data may be reproduced globally, effectively exporting potentially harmful patterns from one context to another. This reinforces the concern that without regional adaptation, platform infrastructure can contribute to the spread of a singular, dominant model of discourse.

In response to SRQ3: Does the position of a suggestion (its rank) influence toxicity? The analysis found no systematic relationship between rank and toxicity. Toxic suggestions were not concentrated at the top of the list, nor was there evidence that ranking consistently amplifies exposure to harmful content. While a slight uptick in toxicity was observed in mid-list positions, the effect was small and inconsistent. Overall, ranking appears neutral with respect to toxicity, though the occasional presence of toxic suggestions anywhere in the list remains a relevant concern.

A key methodological contribution of this study was the validation of the toxicity scoring model. This process involved comparing human annotations with three automated classifiers: the Perspective API, UnRoBERTa, and ToxicBERT. Rather than a minor supporting step, this validation was essential for confirming the reliability of the measurement approach used throughout the analysis. The initial pilot using a five- and three-point severity scale (Section 3.5) revealed low

agreement among annotators, highlighting the challenges of interpreting toxicity through discrete categories. In response, a second validation (Section 3.6) adopted a probability-based method, using binary labels from 10 annotators to estimate the likelihood of perceived toxicity. The Perspective API, which was ultimately selected for scoring the full dataset, showed the strongest alignment with human judgment under this probabilistic framework. While no model fully captures the nuance of human interpretation, this two-stage validation process offered a more grounded and interpretable way to apply automated toxicity scores at scale, especially in the context of identity-related prompts.

Taken together, the findings across all three sub-research questions provide a comprehensive answer to the thesis's central question: Does TikTok's autocomplete system exhibit patterns of bias in how it generates identity-related suggestions? The evidence shows that, while most completions do not display high toxicity, the system disproportionately assigns higher toxicity to prompts referencing marginalized identities, particularly those involving race and sexual orientation. These patterns persist across contexts, themes, and prompt forms. Meanwhile, geographic uniformity and rank neutrality suggest that such biases are systemic, not user-driven or region-specific. Ultimately, autocomplete emerges not as a passive tool but as a discursive mechanism that subtly encodes and reproduces social hierarchies through predictive language.

6.2 Academic relevance

This thesis contributes to the academic discourse on algorithmic bias by turning attention toward an often-overlooked interface: autocomplete in TikTok's search function. While much of the existing literature has focused on recommendation algorithms, this research demonstrates that bias can emerge even earlier at the point where users form their queries. By treating toxicity not as an absolute moral label but as a discursive signal, the study bridges computational methods and critical theory, drawing on and extending frameworks from Noble (2018), Kumar et al. (2021), and Sap et al. (2019). The approach allows for the scalable detection of representational bias while remaining sensitive to context and interpretation.

The cross-national scope further strengthens the academic contribution by showing how standardized model behavior persists across culturally diverse English-speaking regions, contributing new evidence to discussions on algorithmic globalization and the export of Western discourse norms. Finally, the integration of human validation alongside automated scoring highlights the value of mixed-methods design in algorithm audit research, offering a practical and adaptable model for future studies. Together, these contributions expand the literature on sociotechnical systems, showing how predictive features like autocomplete play a subtle but significant role in shaping digital discourse.

6.3 Social relevance

Beyond its academic contributions, this thesis holds broader social significance by highlighting how even seemingly minor interface features, like autocomplete can subtly reproduce and reinforce harmful stereotypes. TikTok is increasingly used as a search engine by younger users, often serving as a gateway to information, identity exploration, and cultural discourse. When toxic suggestions are disproportionately linked to certain groups, they not only reflect underlying societal biases but also risk normalizing them through repeated exposure. In this way, autocomplete becomes more than a convenience tool; it acts as a framing device that shapes how users perceive themselves and others.

The study's findings draw attention to the structural implications of algorithmic systems in everyday life. By showing that toxicity patterns remain stable across countries, the research underscores the global reach of algorithmically encoded bias and the importance of addressing it

at the systemic level. These insights are particularly relevant in light of growing concerns about digital discrimination, platform accountability, and the mental health impacts of online environments. Recognizing that bias can be embedded in subtle design decisions reinforces the need for more transparent, culturally aware, and socially responsible approaches to algorithm development. As platforms like TikTok continue to shape public discourse, understanding and mitigating these dynamics becomes critical not just for technical fairness, but for broader social equity.

Disclosure of AI Support

Parts of this thesis were developed with the assistance of AI tools, primarily OpenAI's ChatGPT, to support writing, editing, and summarizing complex statistical and methodological content. The use of AI was limited to improving clarity, refining phrasing, and structuring long sections; all critical decisions regarding research design, analysis, interpretation, and argumentation were made by the author. The final content reflects the author's original thinking, grounded in independent reading, coding, and statistical work. Every use of AI was carefully reviewed to ensure academic integrity and adherence to scholarly standards.

References

- Azzopardi, L. (2021). Cognitive biases in search: a review and reflection of cognitive biases in Information Retrieval. Proceedings of the 2021 conference on human information interaction and retrieval,
- Baker, P., & Potts, A. (2013). 'Why do white people have thin lips?' Google and the perpetuation of stereotypes via auto-complete search forms. *Critical discourse studies*, 10(2), 187-204.
- Bhandari, A., & Bimo, S. (2022). Why's everyone on TikTok now? The algorithmized self and the future of self-making on social media. *Social media+ society*, 8(1), 20563051221086241.
- Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (technology) is power: A critical survey of "bias" in nlp. *arXiv preprint arXiv:2005.14050*.
- Bonart, M., Samokhina, A., Heisenberg, G., & Schaer, P. (2020). An investigation of biases in web search engine query suggestions. *Online Information Review*, 44(2), 365-381.
- Brown, M. A., Bisbee, J., Lai, A., Bonneau, R., Nagler, J., & Tucker, J. A. (2022). Echo chambers, rabbit holes, and algorithmic bias: How YouTube recommends content to real users. *Available at SSRN 4114905*.
- Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. Conference on fairness, accountability and transparency,
- Chen, Y., Li, M., Guo, F., & Wang, X. (2023). The effect of short-form video addiction on users' attention. *Behaviour & Information Technology*, 42(16), 2893-2910.
- Chen, Z., & Shi, C. (2022). Analysis of algorithm recommendation mechanism of TikTok. *International Journal of Education and Humanities*, 4(1), 12-14.
- Council, E. P. a. (2023). Regulation (EU) 2023/1115 of 7 June 2023 on Artificial Intelligence (Artificial Intelligence Act). Official Journal of the European Union, L 126, 1–103. <https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence>
- Davidson, T., Warmesley, 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,
- Delouya, S. (2022). Nearly half of Gen Z is using TikTok and Instagram for search instead of Google, according to Google's own data. <https://www.businessinsider.com/nearly-half-gen-z-use-tiktok-instagram-over-google-search-2022-7>
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers),
- Epstein, R., Aries, S., Grebbien, K., Salcedo, A. M., & Zankich, V. R. (2024a). The search suggestion effect (SSE): A quantification of how autocomplete search suggestions could be used to impact opinions and votes. *Computers in Human Behavior*, 108342.
- Epstein, R., Aries, S., Grebbien, K., Salcedo, A. M., & Zankich, V. R. (2024b). The Search Suggestion Effect (SSE): How Autocomplete Search Suggestions Can Be Used to Impact Opinions and Votes [Computers in Human Behavior, accepted for publication, June 2024]. *Computers in Human Behavior*.
- Epstein, R., & Robertson, R. E. (2015). The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. *Proceedings of the National Academy of*

Sciences, 112(33), E4512-E4521.

- Feldkamp, J. (2021). The rise of TikTok: The evolution of a social media platform during COVID-19. *Digital responses to Covid-19: Digital innovation, transformation, and entrepreneurship during pandemic outbreaks*, 73-85.
- Gevers, I., Markov, I., & Daelemans, W. (2022). Linguistic analysis of toxic language on social media. *Computational Linguistics in the Netherlands*,
- Goyal, N., Kivlichan, I. D., Rosen, R., & Vasserman, L. (2022). Is your toxicity my toxicity? exploring the impact of rater identity on toxicity annotation. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2), 1-28.
- Graham, R. (2023). The ethical dimensions of Google autocomplete. *Big Data & Society*, 10(1), 20539517231156518.
- Ha, E., Kong, H., & Jhaver, S. (2024). Examining Racial Stereotypes in YouTube Autocomplete Suggestions. *arXiv preprint arXiv:2410.03102*.
- Harchekar, J. S. (2017). Impact of social media on society.
- Hazen, T. J., Olteanu, A., Kazai, G., Diaz, F., & Golebiewski, M. (2020). On the social and technical challenges of web search autosuggestion moderation. *arXiv preprint arXiv:2007.05039*.
- He, X., Zannettou, S., Shen, Y., & Zhang, Y. (2024). You only prompt once: On the capabilities of prompt learning on large language models to tackle toxic content. 2024 IEEE Symposium on Security and Privacy (SP),
- Hiemstra, D. (2020). Reducing misinformation in query autocompletions. *arXiv preprint arXiv:2007.02620*.
- Houli, D., Radford, M. L., & Singh, V. K. (2021). "COVID19 is_": The Perpetuation of Coronavirus Conspiracy Theories via Google Autocomplete. *Proceedings of the Association for Information Science and Technology*, 58(1), 218-229.
- Huang, T. (2024). Analysis of Short Video Cultural Communication Under the Threshold of New Media—Taking TikTok Short Video as an Example.
- Hussein, E., Juneja, P., & Mitra, T. (2020). Measuring misinformation in video search platforms: An audit study on YouTube. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1), 1-27.
- Kliman-Silver, C., Hannak, A., Lazer, D., Wilson, C., & Mislove, A. (2015). Location, location, location: The impact of geolocation on web search personalization. *Proceedings of the 2015 internet measurement conference*,
- Kumar, D., Kelley, P. G., Consolvo, S., Mason, J., Bursztein, E., Durumeric, Z., Thomas, K., & Bailey, M. (2021). Designing toxic content classification for a diversity of perspectives. *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*,
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, 159-174.
- Leidinger, A., & Rogers, R. (2023). Which Stereotypes Are Moderated and Under-Moderated in Search Engine Autocompletion? *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*,
- MacIntyre, A. (2003). *A short history of ethics: a history of moral philosophy from the Homeric age to the 20th century*. Routledge.
- Muñoz-Bates, S. (2015). Instant Results with Lingering Effects: How Google Can Bias Our Social Conversations. *AoIR Selected Papers of Internet Research*.
- Noble, S. U. (2018). Algorithms of oppression: How search engines reinforce racism. In *Algorithms of oppression*. New York university press.
- OECD. (2019). *Artificial Intelligence in Society*, OECD Publishing, Paris, .
<https://doi.org/https://doi.org/10.1787/eedfee77-en>
- Robertson, R. E., Jiang, S., Lazer, D., & Wilson, C. (2019). Auditing autocomplete: Suggestion

- networks and recursive algorithm interrogation. Proceedings of the 10th ACM conference on web science,
- Roy, S., & Ayalon, L. (2020). Age and gender stereotypes reflected in Google's "autocomplete" function: The portrayal and possible spread of societal stereotypes. *The Gerontologist*, 60(6), 1020-1028.
- Sap, M., Card, D., Gabriel, S., Choi, Y., & Smith, N. A. (2019). The risk of racial bias in hate speech detection. Proceedings of the 57th annual meeting of the association for computational linguistics,
- Sellars, A. (2016). Defining hate speech. *Berkman Klein Center Research Publication*(2016-20), 16-48.
- Simpson, E., & Semaan, B. (2021). For you, or for" you"? Everyday LGBTQ+ encounters with TikTok. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW3), 1-34.
- Stangor, C., & Schaller, M. (2000). Stereotypes as individual and collective representations.
- Wang, P., Mi, X., Liao, X., Wang, X., Yuan, K., Qian, F., & Beyah, R. A. (2018). Game of Missuggestions: Semantic Analysis of Search-Autocomplete Manipulations. NDSS,
- Ward, D., Hahn, J., & Feist, K. (2012). Autocomplete as research tool: A study on providing search suggestions. *Information Technology and Libraries*, 31(4), 6-19.
- Wulczyn, E., Thain, N., & Dixon, L. (2017). Ex machina: Personal attacks seen at scale. Proceedings of the 26th international conference on world wide web,
- Zhang, M., & Liu, Y. (2021). A commentary of TikTok recommendation algorithms in MIT Technology Review 2021. *Fundamental Research*, 1(6), 846-847.

A

Supplementary Data and Analyses

A1. Construction of Prefixes

Each query prefix (*pf*) can be decomposed into two parts: a set of selection conditions (*sel*) and a semantic content (*con*), denoted as *pf* = (*sel*, *con*) or (*con*, *sel*).

1. The construction of the *sel*

Variable	Subgroup	<i>sel</i>
Gender	Male-related	<i>man, men, boy, husband, boyfriend</i>
	Female-related	<i>woman, women, girl, wife, girlfriend</i>
Race	White-related	<i>White person, White people, White man, White woman</i>
	Black-related	<i>Black person, Black people, Black man, Black woman</i>
Sexual Orientation	Heterosexual-related	<i>straight person, straight people, straight man, straight woman</i>
	Homosexual-related	<i>gay person, gay people, gay man, gay woman, lesbian, lesbians</i>

2. The construction of the *con*

	<i>con</i>
Interrogative	<i>why..., how..., what do/does..., how many..., how much, can..., when..., where..., should ..., could..., do/does..., is/are...,</i>
Declarative	Appearance & physical <i>looks, smell, hair, hair style, body, skin, brain, eyes, legs, arms, clothes, wear, outfit, figure, face, appearance, height, weight, exercise, powerful, running, lose weight, fitness</i>
	Behavior <i>participate, become, pursuit, dream, behavior, choose, compete, competition, make decisions, spend money, training, want to, read, capability, talk, chat</i>
	Personal <i>mind, character, quality, thought, status, heart, like, habits</i>

	Cognitive	<i>lq, eq, cognition, feeling, knowledge, value, responsibility, ambition, good at</i>
	Career	<i>nurse, pilot, police, professor, student, engineer, programmer, athlete, entrepreneur, boss, mentor, works, career, ability, staff, promotion, employ, business, success, researcher, scientists, target, job application,</i>
	Domestic	<i>cook, babysit, wash dishes, housework, adopt child, raise child, at home</i>
	Relationship	<i>intimate, romantic, sex, love, partner, in relationship, dating, marriage</i>
	Identity	<i>themselves, most, lots of, a lot of, a few, some of,</i>

A2. Prompts with 0 suggestions in all countries

This matrix presents the list of prompts (n = 98) that returned no autocomplete suggestions across all ten countries in the study.

Prompts with 0 suggestions in all countries		
<i>black man sex</i>	<i>boy romantic</i>	<i>woman exercise</i>
<i>black people sex</i>	<i>boy sex</i>	<i>woman intimate</i>
<i>black person sex</i>	<i>boy skin</i>	<i>women exercise</i>
<i>black woman sex</i>	<i>boyfriend dating</i>	<i>women sex</i>
<i>boy body</i>	<i>boyfriend intimate</i>	<i>boy romantic</i>
<i>boy fitness</i>	<i>boyfriend love</i>	<i>boy sex</i>
<i>boy lose weight</i>	<i>boyfriend sex</i>	<i>boy skin</i>
<i>boy love</i>	<i>does wife</i>	<i>boyfriend dating</i>
<i>gay man sex</i>	<i>gay people sex</i>	<i>boyfriend intimate</i>
<i>gay person sex</i>	<i>gay woman sex</i>	<i>boyfriend love</i>
<i>girl body</i>	<i>girl dating</i>	<i>boyfriend sex</i>
<i>girl exercise</i>	<i>girl figure</i>	<i>girl partner</i>
<i>girl fitness</i>	<i>girl legs</i>	<i>girl sex</i>
<i>girl lose weight</i>	<i>girl love</i>	<i>girlfriend dating</i>
<i>girl marriage</i>	<i>straight woman sex</i>	<i>straight person sex</i>
<i>girl romantic</i>	<i>white people looks</i>	<i>straight people sex</i>
<i>girl smell</i>	<i>white person sex</i>	<i>white people value</i>
<i>straight man sex</i>	<i>girl love</i>	<i>straight people sex</i>
<i>straight woman sex</i>	<i>white man sex</i>	<i>white man sex</i>
<i>white people looks</i>	<i>white people sex</i>	<i>white people sex</i>
<i>white person sex</i>	<i>white woman sex</i>	<i>white woman sex</i>
<i>wife sex</i>	<i>boy love</i>	<i>gay people sex</i>
<i>woman fitness</i>	<i>gay man sex</i>	<i>gay woman sex</i>
<i>woman smell</i>	<i>gay person sex</i>	<i>girl dating</i>
<i>women fitness</i>	<i>girl body</i>	<i>girl figure</i>
<i>black man sex</i>	<i>girl exercise</i>	<i>girl legs</i>
<i>black people sex</i>	<i>girl fitness</i>	<i>white people value</i>
<i>black person sex</i>	<i>girl lose weight</i>	<i>girl partner</i>
<i>black woman sex</i>	<i>girl marriage</i>	<i>girl sex</i>
<i>boy body</i>	<i>girl romantic</i>	<i>girlfriend dating</i>

<i>boy fitness</i>	<i>girl smell</i>	<i>straight person sex</i>
<i>boy lose weight</i>	<i>straight man sex</i>	

A3. Frequency of Incomplete Autocomplete Results by Country

This table shows the frequency of prompts by number of suggestions received (from 0 to 8), highlighting the partial outputs observed during data collection.

Number of Suggestions	Number of Prompts
0	974
1	508
2	596
3	705
4	725
5	703
6	682
7	640
8	25,827
Total:	31,136

A4. Examples of Prompts with Incomplete Autocomplete Results Across Countries

This table presents examples of prompts that consistently returned fewer than eight autocomplete suggestions in all ten countries.

Prompt	
<i>white man legs</i>	<i>boy themselves</i>
<i>white person legs</i>	<i>boy thought</i>
<i>should gay man</i>	<i>girl participate</i>
<i>a few gay people</i>	<i>black person legs</i>
<i>a lot of girl</i>	<i>girl raise child</i>
<i>lesbians staff</i>	<i>boy want to</i>
<i>does boy</i>	<i>what does boy</i>
<i>a lot of boy</i>	<i>boy works</i>
<i>gay person cognition</i>	<i>gay man quality</i>
<i>gay person employ</i>	<i>gay man student</i>
<i>lesbians weight</i>	<i>girl at home</i>
<i>gay people spend money</i>	<i>girl intimate</i>
<i>boy participate</i>	<i>lesbians lose weight</i>
<i>gay people student</i>	<i>boy ability</i>
<i>gay person babysit</i>	<i>boy adopt child</i>
<i>gay person capability</i>	<i>lesbians make decisions</i>
<i>gay man babysit</i>	<i>boy appearance</i>
<i>gay man capability</i>	<i>girl character</i>
<i>straight person engineer</i>	<i>lesbians participate</i>
<i>straight person housework</i>	<i>lesbians professor</i>
<i>boy raise child</i>	<i>girl choose</i>
<i>girl spend money</i>	<i>boy character</i>

<i>girl themselves</i>	<i>lesbians ability</i>
<i>how much boy</i>	<i>gay people babysit</i>
<i>how much girl</i>	<i>girl compete</i>

A5. Top 5 Most Toxic Suggestions per Country

This table shows the 5 highest-toxicity autocomplete suggestions for each country, including the prompt, suggestion, and score.

Country	Prompt	Suggestion	Toxicity
Australia	<i>gay man running</i>	<i>gay man running skit</i>	0.933
	<i>black woman skin</i>	<i>black woman skit</i>	0.924
	<i>white woman skin</i>	<i>white woman skit</i>	0.869
	<i>white woman face</i>	<i>white women doing stupid faces</i>	0.854
	<i>lesbian smell</i>	<i>lesbian fart smell</i>	0.840
Ghana	<i>gay man skin</i>	<i>gay man skit</i>	0.956
	<i>gay man running</i>	<i>gay man running skit</i>	0.933
	<i>gay people skin</i>	<i>gay people skits</i>	0.911
	<i>gay person smell</i>	<i>gay smelling fart</i>	0.902
	<i>gay people smell</i>	<i>gay smelling fart</i>	0.902
Ireland	<i>gay man arms</i>	<i>gay man army skit</i>	0.960
	<i>black people police</i>	<i>black people police skit</i>	0.921
	<i>gay person face</i>	<i>gay guys farting on face</i>	0.911
	<i>gay man face</i>	<i>gay guys farting on face</i>	0.911
	<i>gay people babysit</i>	<i>gay people skits</i>	0.911
New Zealand	<i>gay man skin</i>	<i>gay man skit</i>	0.956
	<i>gay man running</i>	<i>gay man running skit</i>	0.933
	<i>gay person smell</i>	<i>gay smelling fart</i>	0.904
	<i>gay people smell</i>	<i>gay smelling fart</i>	0.904
	<i>gay people smell</i>	<i>smelly gay</i>	0.863
Nigeria	<i>gay man skin</i>	<i>gay man skits</i>	0.911
	<i>gay people skin</i>	<i>gay people skits</i>	0.911
	<i>black man police</i>	<i>black man police you are very stupid remix</i>	0.911
	<i>white woman face</i>	<i>white women doing stupid faces</i>	0.854
	<i>white woman skin</i>	<i>white woman skit girl</i>	0.846
Pakistan	<i>gay people smell</i>	<i>gay smelling fart</i>	0.904
	<i>gay person smell</i>	<i>gay smelling fart</i>	0.904
	<i>white woman skin</i>	<i>white woman skit</i>	0.869
	<i>gay woman smell</i>	<i>smelly gay</i>	0.863
	<i>gay people smell</i>	<i>smelly gay</i>	0.863
Singapore	<i>gay man running</i>	<i>gay man running skit</i>	0.933
	<i>black woman skin</i>	<i>black woman skit</i>	0.924
	<i>gay man skin</i>	<i>gay man skits</i>	0.911
	<i>black man police</i>	<i>black man police you are very stupid remix</i>	0.911
	<i>gay people skin</i>	<i>gay people skits</i>	0.911

Trinidad	<i>black man police</i>	<i>black man police you are very stupid remix</i>	0.911
	<i>gay people skin</i>	<i>gay people skits</i>	0.911
	<i>gay man skin</i>	<i>gay man skits</i>	0.911
	<i>lesbian smell</i>	<i>lesbian fart smell</i>	0.840
	<i>gay people smell</i>	<i>gay people smell bad</i>	0.840
UK	<i>gay man arms</i>	<i>gay man army skit</i>	0.960
	<i>gay people skin</i>	<i>gay people skit</i>	0.944
	<i>black people police</i>	<i>black people police skit</i>	0.921
	<i>gay people babysit</i>	<i>gay people skits</i>	0.911
	<i>gay man face</i>	<i>gay guys farting on face</i>	0.906
USA	<i>black woman skin</i>	<i>black woman skit</i>	0.921
	<i>gay people skin</i>	<i>gay people skits</i>	0.911
	<i>gay man skin</i>	<i>gay man skits</i>	0.911
	<i>gay people smell</i>	<i>gay smelling fart</i>	0.904
	<i>gay person smell</i>	<i>gay smelling fart</i>	0.904

A6. Prompt–Opposite Toxicity Comparison Matrix

This table compares the prompt toxicity, suggestion toxicity, and opposite identity completions for 30 identity-based prompt pairs, highlighting asymmetric autocomplete behavior.

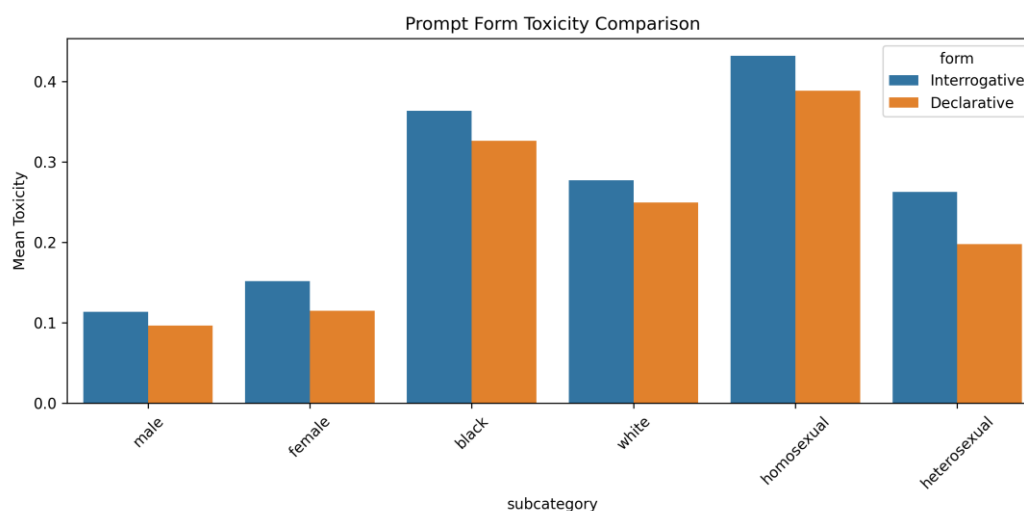
Type	Prompt	Suggestion	Prompt Toxicity	Suggestion Toxicity
Original	<i>gay man arms</i>	<i>gay man army skit</i>	0.484	0.960
Opposite	<i>straight man arms</i>	-	0.139	-
Original	<i>gay people skin</i>	<i>gay people skit</i>	0.401	0.944
Opposite	<i>straight people skin</i>	-	0.2	-
Original	<i>black woman skin</i>	<i>black woman skit</i>	0.236	0.924
Opposite	<i>white woman skin</i>	<i>white woman skit</i>	0.153	0.869
Original	<i>black people police</i>	<i>black people police skit</i>	0.373	0.921
Opposite	<i>white people police</i>	-	0.285	-
Original	<i>gay man face</i>	<i>gay guys farting on face</i>	0.612	0.911
Opposite	<i>straight man face</i>	-	0.27	-
Original	<i>gay person face</i>	<i>gay guys farting on face</i>	0.573	0.911
Opposite	<i>straight person face</i>	-	0.285	-
Original	<i>gay people skin</i>	<i>gay people skits</i>	0.401	0.911
Opposite	<i>straight people skin</i>	-	0.2	-
Original	<i>gay person smell</i>	<i>gay smelling fart</i>	0.751	0.902
Opposite	<i>straight person smell</i>	-	0.621	-
Original	<i>gay people smell</i>	<i>gay smelling fart</i>	0.731	0.902
Opposite	<i>straight people smell</i>	-	0.661	-
Original	<i>black person looks</i>	<i>black person looks like monkey</i>	0.311	0.898
Opposite	<i>white person looks</i>	-	0.201	-
Original	<i>black woman looks</i>	<i>black woman looks like monkey</i>	0.274	0.877
Opposite	<i>white woman looks</i>	-	0.155	-
Original	<i>gay people make</i>	<i>gay people farting</i>	0.294	0.869

	<i>decisions</i>			
Opposite	<i>straight people make decisions</i>	-	0.252	-
Original	<i>gay people cognition</i>	<i>gay people farting</i>	0.399	0.869
Opposite	<i>straight people cognition</i>	-	0.282	-
Original	<i>gay people babysit</i>	<i>gay people farting</i>	0.557	0.869
Opposite	<i>straight people babysit</i>	-	0.401	-
Original	<i>gay people figure</i>	<i>gay people farting</i>	0.401	0.869
Opposite	<i>straight people figure</i>	-	0.265	-
Original	<i>gay people ambition</i>	<i>gay people farting</i>	0.396	0.869
Opposite	<i>straight people ambition</i>	-	0.255	-
Original	<i>gay people researcher</i>	<i>gay people farting</i>	0.423	0.869
Opposite	<i>straight people researcher</i>	-	0.249	-
Original	<i>gay people programmer</i>	<i>gay people farting</i>	0.479	0.869
Opposite	<i>straight people programmer</i>	-	0.251	-
Original	<i>gay people iq</i>	<i>gay people farting</i>	0.452	0.869
Opposite	<i>straight people iq</i>	-	0.259	-
Original	<i>white woman skin</i>	<i>white woman skit</i>	0.153	0.869
Opposite	<i>black woman skin</i>	<i>black woman skit</i>	0.236	0.924
Original	<i>gay people smell</i>	<i>smelly gay</i>	0.731	0.863
Opposite	<i>straight people smell</i>	-	0.661	-
Original	<i>black man looks</i>	<i>black man looks like a monkey</i>	0.279	0.854
Opposite	<i>white man looks</i>	-	0.164	-
Original	<i>white woman face</i>	<i>white women doing stupid faces</i>	0.27	0.854
Opposite	<i>black woman face</i>	-	0.397	-
Original	<i>lesbians employ</i>	<i>useless lesbians</i>	0.361	0.853
Opposite	<i>straight women employ</i>	-	0.249	-
Original	<i>black people brain</i>	<i>brain rot black people</i>	0.44	0.853
Opposite	<i>white people brain</i>	-	0.352	-
Original	<i>lesbians ability</i>	<i>useless lesbians</i>	0.305	0.853
Opposite	<i>straight women ability</i>	-	0.166	-
Original	<i>lesbians knowledge</i>	<i>useless lesbians</i>	0.332	0.853
Opposite	<i>straight women knowledge</i>	-	0.202	-
Original	<i>gay person smell</i>	<i>smell that smells like gay</i>	0.751	0.851
Opposite	<i>straight person smell</i>	-	0.621	-
Original	<i>gay woman smell</i>	<i>smell that smells like gay</i>	0.752	0.851

Opposite	<i>straight woman smell</i>	-	0.563	-
Original	<i>gay person dream</i>	<i>gay trash dream</i>	0.427	0.851
Opposite	<i>straight person dream</i>	-	0.234	-

A7. Mean Toxicity by Prompt Form and Identity Subcategory

The bar chart compares average toxicity scores for interrogative versus declarative prompt forms across six identity subcategories.



A8. Pairwise KS Test Results

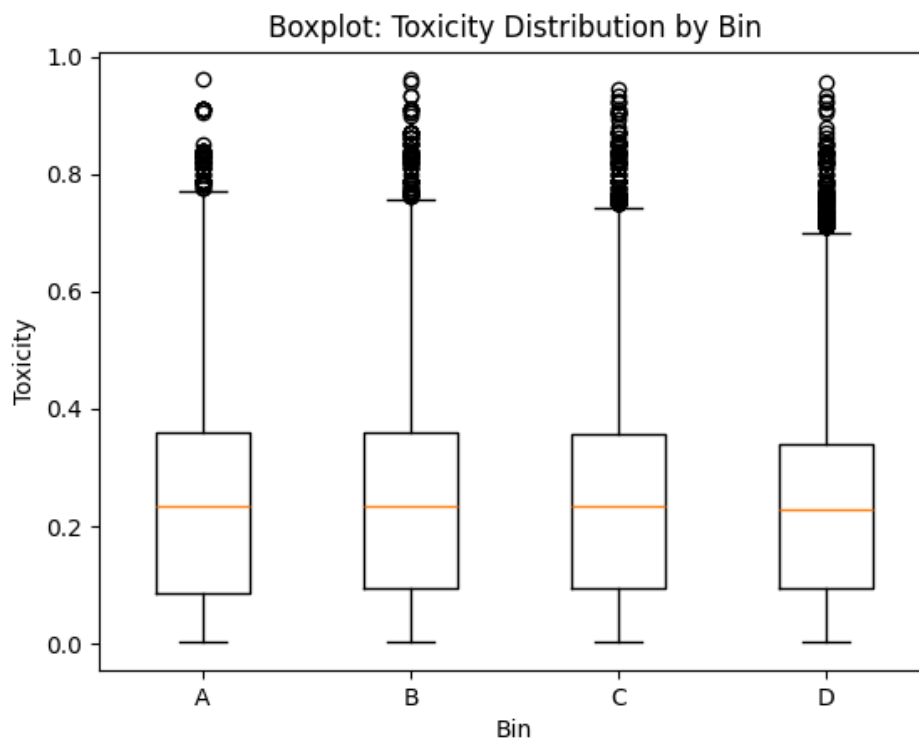
Summary table of two-sample Kolmogorov–Smirnov tests comparing toxicity score distributions across all country pairs

Country Pair	KS Stat	p-value	Significance
Australia vs Ghana	0.03	< .001	***
Australia vs Ireland	0.023	< .001	***
Australia vs New Zealand	0.021	< .001	***
Australia vs Nigeria	0.017	< .001	**
Australia vs Pakistan	0.036	< .001	***
Australia vs Singapore	0.013	0.034	*
Australia vs Trinidad and Tobago	0.015	0.010	*
Australia vs UK	0.022	< .001	***
Australia vs USA	0.02	< .001	***
Ghana vs Ireland	0.037	< .001	***
Ghana vs New Zealand	0.012	< .001	ns
Ghana vs Nigeria	0.044	< .001	***
Ghana vs Pakistan	0.011	0.147	ns
Ghana vs Singapore	0.034	< .001	***
Ghana vs Trinidad and Tobago	0.042	< .001	***
Ghana vs UK	0.035	< .001	***
Ghana vs USA	0.014	0.029	*
Ireland vs New Zealand	0.029	< .001	***
Ireland vs Nigeria	0.025	< .001	***
Ireland vs Pakistan	0.046	< .001	***
Ireland vs Singapore	0.021	< .001	***
Ireland vs Trinidad and Tobago	0.02	< .001	***
Ireland vs UK	0.004	0.985	ns

Ireland vs USA	0.026	< .001	***
New Zealand vs Nigeria	0.037	< .001	***
New Zealand vs Pakistan	0.02	< .001	***
New Zealand vs Singapore	0.029	< .001	***
New Zealand vs Trinidad and Tobago	0.034	< .001	***
New Zealand vs UK	0.027	< .001	***
New Zealand vs USA	0.007	0.648	ns
Nigeria vs Pakistan	0.048	< .001	***
Nigeria vs Singapore	0.012	< .001	ns
Nigeria vs Trinidad and Tobago	0.007	0.700	ns
Nigeria vs UK	0.026	< .001	***
Nigeria vs USA	0.034	< .001	***
Pakistan vs Singapore	0.04	< .001	***
Pakistan vs Trinidad and Tobago	0.047	< .001	***
Pakistan vs UK	0.045	< .001	***
Pakistan vs USA	0.023	< .001	***
Singapore vs Trinidad and Tobago	0.01	0.234	ns
Singapore vs UK	0.021	< .001	***
Singapore vs USA	0.027	< .001	***
Trinidad and Tobago vs UK	0.021	< .001	***
Trinidad and Tobago vs USA	0.033	< .001	***
UK vs USA	0.026	< .001	***

A9. Boxplot: toxicity distribution by two-slot bins

Confirms at a binned level that median toxicity is steady but extreme values cluster in the lower-middle bin.



B

Evaluation of Toxicity Classification Models

B1. Human Annotation Guidelines – Severity Ratings (Used in Section 3.5)

1. Introduction

Welcome, and thank you for contributing to this study. The aim of our research is to investigate the prevalence and patterns of toxic language in TikTok’s autocomplete suggestions across English-speaking countries. By manually reviewing a sample of autocomplete completions, you will help validate and enrich our automated toxicity measurements, ensuring that our findings reflect genuine human judgments.

Your careful and consistent annotations are crucial for assessing the reliability of automated tools and for capturing nuanced cases that algorithms may misclassify.

2. Research Goal

- **Objective:** To quantify how often and in what ways TikTok’s autocomplete suggestions contain toxic language, particularly with respect to identity-related prompts (gender, race, sexual orientation).
- **Role of Human Review:** While automated models provide scores on a 0–1 scale, human annotators bring in context, cultural sensitivity, and ethical judgment to determine whether a suggestion is genuinely harmful.

3. Definition of Toxicity

Toxic language is any text that a reasonable reader would see as rude, hostile, or hateful enough to make someone feel attacked or unwelcome. This includes direct insults, slurs, or profanity; negative stereotypes that demean a protected group (e.g., race, gender, sexuality); and threats or encouragement of harm.

4. Annotation Scale (1–5) s Instructions

The attached Excel file “**Autocomplete_Suggestions.xlsx**” contains **150** autocomplete suggestions. For each suggestion, please enter a score from **1 (Not toxic)** to **5 (Extremely toxic)** in the blank Score column:

Score	Label
1	Not toxic
2	Slightly toxic
3	Moderately toxic
4	Very toxic
5	Extremely toxic

No additional comments are needed, simply read each suggestion and record the appropriate number.

Deadline: Please complete your assigned batch by **17/06/2025**.

B2 Human Annotation Guidelines – Probability Estimates (Used in Section 3.6)

1. Introduction

Thank you for participating in this study. You'll be reviewing a set of real TikTok autocomplete suggestions and helping us validate and enrich our automated toxicity measurements, ensuring that our findings reflect genuine human judgments

2. Research Goal

- **Objective:** To assess the degree to which autocomplete suggestions may be perceived as toxic by users, especially in prompts involving gender, race, or sexual orientation.
- **Role of Human Judgment:** While automated systems assign a toxicity score between 0 and 1, they are trained using human responses like yours. Your ratings help us check how well these systems align with real human perceptions.

3. Definition of Toxicity

Toxicity, defined as a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion.

4. Annotation Instructions (Yes/No)

You will be given a spreadsheet with 150 autocomplete suggestions.

The spreadsheet contains three columns: the first column shows the original search prompt, the second column shows the autocomplete suggestion generated from that prompt, and the third column is where you will enter your answer, based only on the autocomplete suggestion.

In the Score column, enter:

- 1 = Yes, if the autocomplete suggestion is considered toxic
- 0 = No, if the autocomplete suggestion is NOT considered toxic

There are no comment fields, just enter 1 or 0 for each item. Every response helps, and there's no need to overthink individual cases.

5. Deadline

Please complete your assigned batch by **20/06/2025**. Your careful and honest input is greatly appreciated and will directly support the quality of our findings.