UNIVERSITY OF TARTU
Faculty of Science and Technology
Institute of Computer Science
Data Science Curriculum

Triin Pohla

Toxicity in Google Play Store Reviews: What, Where and Why?

Master's Thesis (15 ECTS)

Supervisors: Vigneshwaran Shankaran, MSc
Rajesh Sharma, PhD

Tartu 2023

Toxicity in Google Play Store: What, Where and Why?

Abstract:

With an ever-growing user base, more and more people are using mobile applications and actively providing feedback on application stores, influencing app quality and user experiences. Despite ongoing efforts in moderating online content, offensive language in online comments is a common phenomenon. This thesis presents a large-scale study that explores the prevalence of toxicity in Google Play Store using nearly 60M application reviews from over 5800 applications over the span of nine years from January 2014 to January 2023. We finetune a RoBERTa-based multi-class toxic comment classifier that distinguishes between four types of reviews, including toxic and toxic-critical comments, with an accuracy of (88%). We find that, on average, 3.5% of all reviews contain toxic content and while the share of outright toxic comments has remained around 1% over the span of nearly a decade, the share of toxic-critical reviews shows subtle increase from below 2% of all reviews in 2014 to over 3% in January 2023. Major changes to the UX/UI or policies can increase the share of toxic-critical comments while the effect of external events, such as COVID-19 pandemic and Russian invasion of Ukraine, appear to have a limited contribution to toxic content in application reviews. This study contributes to the broader understanding of digital communication and user behavior, different facets of toxic content, and the implications for enhancing online platforms' content moderation strategies and user engagement policies.

Disclaimer: This thesis contains examples of harmful language used for illustration purposes only. These examples do not reflect the opinions or beliefs of the author, and they are not intended to promote or endorse any form of harmful language or behavior.

Keywords:
Toxicity, Google Play store, application reviews, RoBERTa

CERCS: P170 Computer science, numerical analysis, systems, control

Toksilisus Google Play poes: mis, kus ja miks?

Lühikokkuvõte:

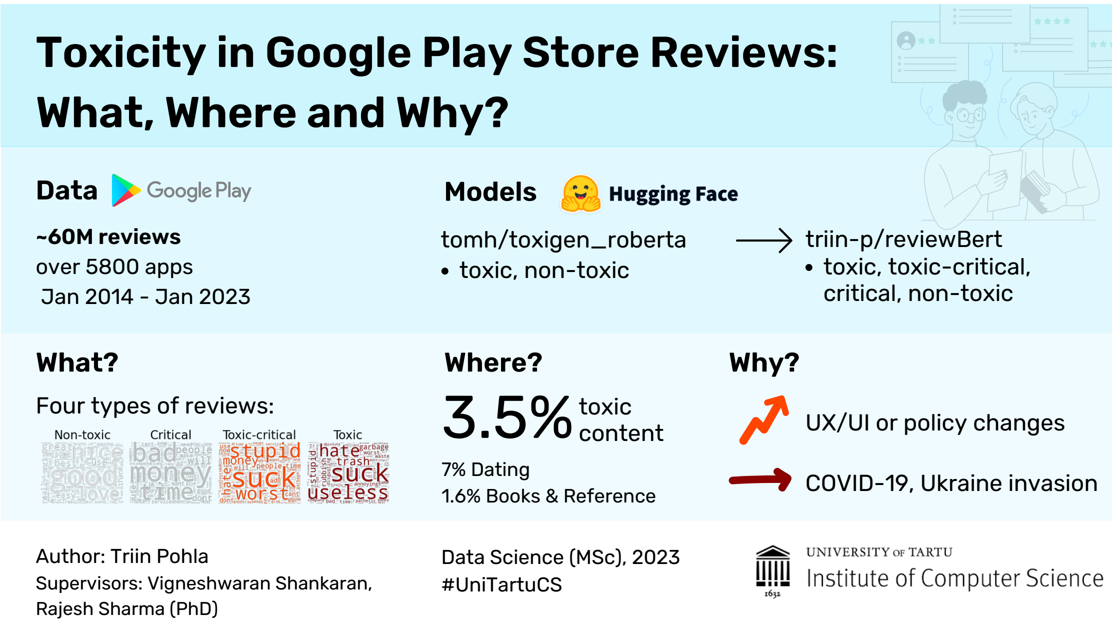
Üha kasvava kasutajaskonnaga mobiilirakendusi kasutab igapäevalt üha enam inimesi, kes annavad aktiivselt tagasisidet rakenduste kohta, mõjutades nii nende kvaliteeti kui teiste kasutajate kogemusi. Vaatamata jätkuvatele jõupingutustele veebisisu modereerimisel, on solvav sõnakasutus kommentaarides tavaline nähtus. Käesolev magistritöö tutvustab laiapõhjalist uuringut toksilise sisu levimusest Google Play poes, kasutades üheksa-aastase perioodi jooksul (jaanuar 2014 kuni jaanuar 2023) enam kui 5800 rakendusele kirjutatud peaaegu 60 miljonit arvustust. Selleks treenime RoBERTal põhineva mitmeklassilise toksilisuse klassifikaatori, mis eristab 88%-täpsusega nelja tüüpi arvustusi, sealhulgas toksilisi ja toksilis-kriitilisi kommentaare. Analüüsi tulemustest selgub, et keskmiselt sisaldab 3,5% kõikidest arvustustest toksilist sisu ning kuigi otseselt halvustavate kommentaaride osakaal on peaaegu dekaadi jooksul püsinud 1% lähedal kõikidest arvustustest, siis toksilis-kriitiliste arvustuste osakaal on vaikselt kasvanud 2014. aasta jaanuaris vähem kui kahelt protsendilt 2023. aasta jaanuariks veidi enam kui kolmeni. Suured muudatused kasutajakogemuses või eeskirjades võivad suurendada toksilis-kriitiliste kommentaaride osakaalu. Välised sündmused, nagu COVID-19 ja Venemaa sissetung Ukrainasse, ei näi avaldavat rakenduste kommentaariumile märgatavat negatiivset mõju. Käesolev töö aitab laiemalt mõista digitaalset suhtlust ja kasutajakäitumist, toksilise sisu erinevaid tahke ning mõju veebiplatvormide sisu modereerimise strateegiate ja kasutajate kaasamispoliitika tõhustamisele.

Hoiatus: See lõputöö sisaldab näiteid toksilistest kommentaaridest, mis on esitatud vaid illustreerival eesmärgil. Need näited ei kajasta autori arvamusi ega tõekspidamisi ning nende eesmärk ei ole propageerida ega toetada mis tahes vormis kahjulikku keelekasutust või käitumist.

Võtmesõnad:

Toksilisus, Google Play pood, rakenduste arvustused, RoBERTa

CERCS: P170 - arvutiteadus, arvuline analüüs, süsteemid, kontroll



Toxicity in Google Play Store Reviews: What, Where and Why?

Data Google Play

~60M reviews
over 5800 apps
Jan 2014 - Jan 2023

Models Hugging Face

tomh/toxigen_roberta
- toxic, non-toxic

→ triin-p/reviewBert
- toxic, toxic-critical, critical, non-toxic

What?

Four types of reviews:

Non-toxic Critical Toxic-critical Toxic

nice
good
love
bad
people
money
time
stupid
money
suck
worst
hate
trash
suck
useless
garbage

Where?

3.5% toxic content

7% Dating
1.6% Books & Reference

Why?

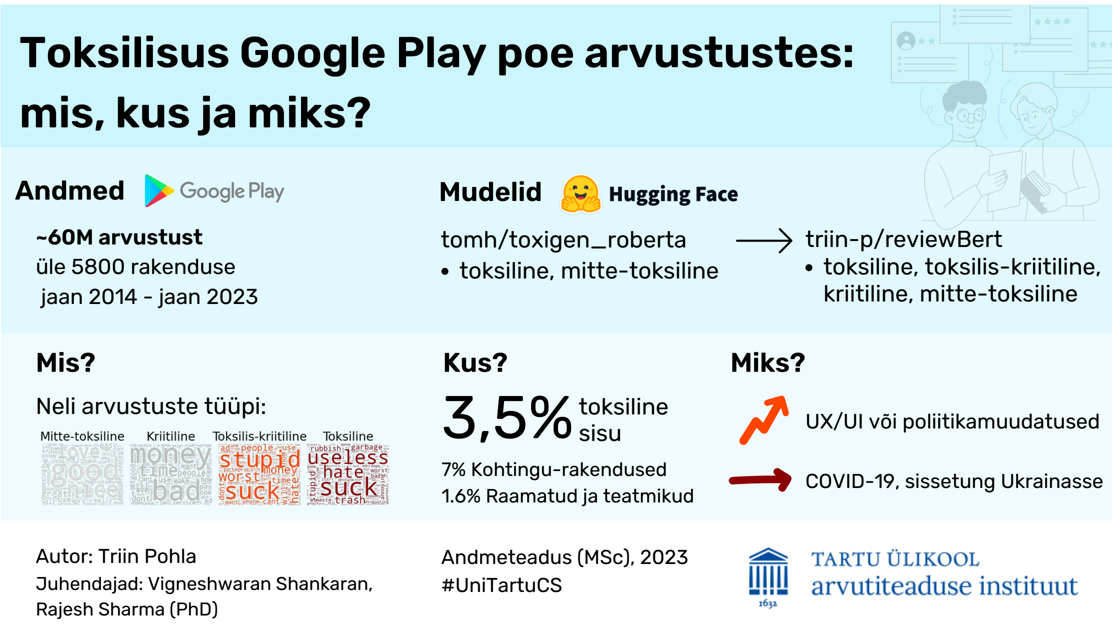
UX/UI or policy changes

COVID-19, Ukraine invasion

Author: Triin Pohla
Supervisors: Vigneshwaran Shankaran, Rajesh Sharma (PhD)

Data Science (MSc), 2023
#UniTartuCS

UNIVERSITY OF TARTU
Institute of Computer Science



Toksilisus Google Play poe arvustustes: mis, kus ja miks?

Andmed Google Play

~60M arvustust
üle 5800 rakenduse
jaan 2014 - jaan 2023

Mudelid Hugging Face

tomh/toxigen_roberta
- toksiline, mitte-toksiline

→ triin-p/reviewBert
- toksiline, toksilis-kriitiline, kriitiline, mitte-toksiline

Mis?

Neli arvustuste tüüpi:

Mitte-toksiline Kriitiline Toksilis-kriitiline Toksiline

Kus?

3,5% toksiline sisu

7% Kohtingu-rakendused
1.6% Raamatud ja teatmikud

Miks?

UX/UI või poliitikamuudatused

COVID-19, sissetung Ukrainasse

Autor: Triin Pohla
Juhendajad: Vigneshwaran Shankaran, Rajesh Sharma (PhD)

Andmeteadus (MSc), 2023
#UniTartuCS

TARTU ÜLIKOOL
arvutiteaduse instituut

Contents

1 Introduction 6

2 Background 8
2.1 Toxicity in online media 8
2.2 Models and tools for toxicity detection 10
2.3 App store reviews for data analysis 12

3 Data and Methodology 13
3.1 Scraping app reviews and metadata 13
3.2 English language and toxicity detection 15
3.2.1 English language detection 15
3.2.2 Toxicity detection with ToxiGen RoBERTa 16
3.3 Finetuning a multi-class toxicity detection model 17
3.3.1 Labeling training data 17
3.3.2 Toxic content classification models 18
3.3.3 reviewBERT 21
3.4 Scaling up the dataset 23
3.5 Analysing the data and presenting the results 25

4 Results and Discussion 29
4.1 Toxicity and app genres 29
4.2 Toxicity and other variables 31
4.3 Temporal trends 33
4.4 External events 36
4.4.1 Chinese apps and COVID-19 pandemic 37
4.4.2 Russian apps and invasion of Ukraine 38

5 Conclusion 42

References 50

Appendix 51
I. GitLab repository 51
II. Licence 52

1 Introduction

In today's interconnected world, where mobile apps have seamlessly integrated into our daily routines, user feedback in the form of app reviews holds significant influence. However, this digital landscape is not without its challenges. One such challenge is the presence of toxic content within these reviews, encompassing a range of negative behaviors such as hate speech, harassment, and offensive language. This study takes a glance at the Google Play Store reviews to gain insight into the "what, where, and why" of toxicity, skimming the surface of its intricacies and implications.

Toxicity, within the context of this study, encompasses a range of negative and harmful behaviors expressed through user reviews. We find that distinguishing between toxic and toxic-critical reviews is crucial. While toxic reviews solely contain offensive language or hurtful comments, toxic-critical reviews also offer pointed critique on an app's features, ease of use, or other significant factors. The removal of explicitly toxic comments can enhance the quality of app store review section without sacrificing meaningful input. Conversely, erasing toxic-critical reviews could deprive developers of valuable user feedback.

In this thesis, we seek to answer the following research questions:

> ***RQ1*:** How prevalent is toxicity in Google Play Store app reviews, and does it exhibit similarities or differences across various app genres?
>
> ***RQ2*:** Are specific variables in app metadata more strongly correlated with the presence of toxicity?
>
> ***RQ3*:** Has the prevalence of toxicity in app reviews increased over the past decade?
>
> ***RQ4*:** Do external events contribute to the presence of toxicity in app reviews? More specifically, did the COVID-19 pandemic and Russian invasion of Ukraine lead to an upsurge of negative comments in the review section of Chinese and Russian apps?

In this thesis, our focus is on understanding how levels of toxicity vary across diverse app types. We explore the connections between four distinct toxicity categories and app metadata, including its genre, installation count, release year, payment model, presence of advertisements, and more. Additionally, we analyse temporal trends, revealing that the proportion of toxic comments has remained consistently around 1% of all reviews for almost a decade (from January 2014 to January 2023). However, the prevalence of toxic-critical reviews has exhibited a slight increase, rising from just below 2% of all reviews in 2014 to over 3% by January 2023.

While examining the impact of toxicity, we explore some of the causal factors that may contribute to its prevalence. The study brings two examples of the role of app-related dynamics, such as major changes in user interface and experience (UI/UX), as well as shifts in user policies. Additionally, we assess the limited impact of external events, like the COVID-19 pandemic and the Russian invasion of Ukraine, on the occurrence of toxic content within reviews.

Understanding the significance of this study is crucial for several reasons. In an era where mobile apps have become an integral part of daily life, the quality of user experiences holds immense importance. Toxic content within Google Play Store reviews can severely impact these experiences, dissuading potential users and tarnishing an app's reputation. Moreover, developers rely on user feedback to improve their apps and cater to user needs. The presence of toxic comments hinders this process, potentially leading to the loss of valuable insights. By delving into the world of app reviews and toxicity, we uncover insights that not only shed light on user behavior but also offer actionable information to enhance app quality and user engagement. This study's findings have the potential to shape content moderation strategies, facilitate informed decision-making, and ultimately contribute to creating a healthier online environment for both developers and users.

This thesis makes a multifaceted contribution to the field. Initially, we embarked on the training of a multiclass review classification model called reviewBERT, proficient in identifying four distinct types of reviews with an accuracy rate of 88%. Subsequently, we curated an extensive dataset encompassing close to 60 million Google Play Store reviews, originating from a diverse range of over 5800 apps. This dataset serves as a comprehensive resource, enabling an in-depth analysis of user feedback and the identification of toxic content within the app ecosystem. Lastly, our investigation into the data provides empirical insights into the prevalence of toxicity in app reviews, temporal trends and factors influencing the occurrence of toxic content.

The subsequent chapters are structured as follows: Chapter 2 provides background information on toxicity in online media, models and tools for toxicity detection, and how app reviews have been used in research. Chapter 3 focuses on the data and methodology, providing an overview of all the steps taken from scraping the data, developing a custom toxicity detection model, preparing and analysing the data, and presenting the results. In Chapter 4, we present and discuss the main findings. The thesis concludes with Chapter 5 that summarises the key insights from this study.

2 Background

This section introduces key aspects related to toxicity in online media, tools for detecting toxicity, and the importance of app store reviews for data analysis. We begin by defining toxicity and its range of negative behaviors in user-generated content. We then examine toxicity's variations across different online platforms. Subsequently, we discuss models and tools used to detect toxicity, from rule-based systems to advanced machine learning models like BERT. Lastly, we underline the significance of app store reviews as a valuable source for understanding user sentiments and experiences. These discussions provide a foundation for the analysis of toxicity in Google Play Store reviews.

Definition of toxicity Toxicity is often considered an umbrella term for various antisocial behaviour, including hate speech, harassment and cyberbullying [MCK$^+$22], or related phenomena, including offensive, abusive, hateful or other types of unsafe content [PSD$^+$20] that can offend or harm its recipients [KBA19] or make someone leave a discussion [Goo17]. In different contexts and use cases, definitions of toxicity may vary, reflecting the diverse nature of digital spaces and their unique user dynamics. Within the focus of this thesis, Google Play Store comment posting policy[1] discourages the promotion of violence and inciting hatred, and the posting of obscene, profane or offensive language and reviews that harass, bully or attack others.

2.1 Toxicity in online media

In the last few years, toxicity has received a lot of attention in the context of online platforms, especially social media [WK15, SGK$^+$23]. Toxicity, however, manifests itself differently on various platforms and online communities [MCK$^+$22]. Studies have shown that most people do not respond to explicit aggression [ZBQ14], and toxic behavior, such as harassment, can decrease user participation and retention [Wik15]. On the other hand, when analysing Reddit data, [XZL$^+$20] found that toxicity in a comment significantly increased the number of its replies in 4/5 subreddits. Other differences between platforms also emerge, including in the prevalence of toxicity: while it is considered endemic in Twitter [GSH16], in four prominent SE platforms – GitHub, Gitter, Slack and Stack Overflow, it ranges from 0.07% to 0.43% of the posts [CSC21b].

In their comparative analysis across five online platforms (Reddit, Wikipedia, Twitter, Stack Overflow, and GitHub), [MCK$^+$22] examine various forms of toxicity, interventions, and their effectiveness, and emphasise the necessity of tailoring interventions to suit the unique characteristics of each platform. To the best of my knowledge, large-scale studies of toxic content have not yet encompassed reviews from app stores like Google Play. In a specific study focusing on Google Play Store, [MS22] analyse 1200 reviews

[1] `https://play.google.com/about/comment-posting-policy/` (last visited 05.07.2023)

from 18 apps across three categories (Service, Pharma, and Travel), totaling 21,600 reviews. Their findings suggest a correlation between higher reviewer scores and positive content sentiment, along with reduced comment toxicity. However, their study did not explore the prevalence, trends, or the interplay between external factors and toxic reviews which forms the central focus of this thesis.

User comments play a central role in social media, online discussion forums and platforms connecting service and product providers with their users. Reviews for mobile apps, hosted by Google Play and Apple App Store, are crowdsourcing knowledge of user experience with the apps, providing valuable information for app release planning, such as major bugs to fix and important features to add [GLQ$^+$22]. [VHMN12] found that users tend to leave short, yet informative reviews, but tend to leave longer messages when they rate an app poorly. Most of the feedback is provided shortly after new releases and the quality and constructiveness varies widely, from helpful advice and innovative ideas to insulting offenses [PM13].

User content, however, can also be abusive [CDNML15]. This puts the hosting platforms under pressure to combat this type of content, which can damage their reputation and make them liable to fines, e.g., when hosting comments encouraging illegal actions [PMA17]. Content moderation, including curbing the spread of toxic and hateful comments, requires a systemic balancing of individual speech rights against other societal interests and values [Dou21]. To this end, various platforms use different strategies from automated detection to moderation [JRSM20, CSC21a], deletion, suspension and warning [MJB$^+$15] or combinations of these[MCK$^+$22] while others (like Gab.com) present themselves as champions of freedom of speech with a 'lax' moderation policy for harmful content [SGK$^+$23].

Most app markets, including Google Play and Apple App Store, have issued strict guidelines or policies for user review submission[2] and the option for all visitors to flag individual reviews as inappropriate. However, it is quite challenging to achieve compliance as large numbers of undesired reviews pop up in the app markets from time to time. Google claimed to have removed millions of fake reviews and thousands of bad apps in 2018 [Fei18], but some developers felt that legitimate positive reviews were caught in the crossfire [Wil18]. Analysis of removed reviews in Apple App Store suggests that the app store is flooding with spam reviews: [WWL$^+$22] studied 33k popular apps over the course of a year and found that more than 30 million reviews were removed from the App Store, accounting for 77.4% of all the reviews the apps received. It can be expected that Google Play Store is under similar pressure, as reports of large quantities of reviews being taken down every few months still continued in 2022 [Ash22].

Practical reasons for removing app reviews include fake, off-topic and offensive reviews, advertising, conflict of interest or other reasons, such as sexually explicit content,

[2] Google Play: https://play.google.com/about/comment-posting-policy/, Apple App Store: https://developer.apple.com/app-store/review/guidelines/ (last visited 05.07.2023)

or personal and confidential information) [WWL$^+$22]. This means that identifying toxic reviews, albeit an important one, is just one task for maintaining user trust in the ratings and reviews of app in the app store [Fei18].

2.2 Models and tools for toxicity detection

Toxicity detection has been an important research topic in natural language processing (NLP), aiming to identify abusive language such as calumniation (e.g., false accusations), discrimination, disrespect, hooliganism, insult, irony, swearing, or threat [PMA17]. Over the years, various machine learning (ML) classifiers, linguistic features, and datasets have been explored for this task. In its early stages, toxicity detection primarily relied on rule-based systems, which used predefined linguistic patterns and keywords to identify potentially harmful content. These systems, while effective to a certain extent, struggled with nuanced language and context. [Noe18] compared 62 classifiers representing 19 major algorithmic families for identifying toxic comments based on statistically significant differences in accuracy and relative execution time. They found that tree-based algorithms provide the most transparently explainable rules and a simple bad word list proves most predictive of offensive commentary. However, for all classifiers in their study, the overall accuracy remained below 40%.

Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) improved the accuracy of toxicity detection by allowing models to capture sequential patterns and contextual information. LSTMs or Long Short-Term Memory networks are a type of RNN which are capable of learning long-term dependencies and CNN are neural networks that can compress large amounts of information into smaller pools [PPS21]. The development of transformers, particularly exemplified by models like BERT (Bidirectional Encoder Representations from Transformers), took toxicity detection to even greater heights. BERT and its variants can effectively grasp the context of words by considering their surrounding text, leading to remarkable improvements in accuracy. Comparative studies revealed that both BERT-based models and Perspective API outperformed traditional approaches on F1 score and accuracy metrics [KSA22, BMST$^+$23].

One of the strong baseline systems for toxicity detection is Perspective API, developed and operated by Google's Jigsaw. This black-box tool uses machine learning models to score a phrase based on its perceived impact in a conversation [Goo17]. It provides scores for attributes such as Toxicity, Severe Toxicity, Insult, Profanity, Identity attack, Threat, and Sexually explicit. Perspective API is freely available in 18 languages and helps websites moderate their forums and comment sections [Goo17]. However, the inner workings of Perspective API, including its architecture and training datasets, are not publicly disclosed, making it challenging to understand its decision-making process.

In the past, [HKZP17] showed that Perspective API misclassifies toxic comments with simple text obfuscation (e.g., inserting punctuation or spaces, doubling letters). However, they also acknowledged that it could easily be defended via text preprocessing, which [RG18] found only increases processing time by a factor of two. Thanks to extensive experiments on covert toxicity, emoji-based hate, human-readable obfuscation and bias evaluation settings, Perspective API is not as easily fooled by such attacks anymore [JBC+18, LTT+22].

Toxicity detection gained significant momentum in 2018 when Kaggle hosted a competition on toxicity detection[3] where over 5000 individuals participated in 4539 teams and submitted 92,230 entries. This challenge aimed to create a multi-headed model capable of identifying various forms of toxicity in text, including threats, obscenity, insults, and identity-based hate, with greater precision than Perspective's existing models. Participants were tasked with leveraging a dataset of comments from Wikipedia's talk page edits to develop models that could contribute to more constructive and respectful online discussions. Notably, many top-ranking teams in the competition employed innovative blends of different approaches, including diverse pre-trained embeddings, translations as train/test-time augmentation (TTA), and robust cross-validation with stacking frameworks, etc.

The introduction of BERT marked another breakthrough in NLP and toxicity detection by pretraining deep bidirectional representations from unlabeled text [DCLT18]. Finetuning BERT models for various tasks, including toxicity detection, became popular from 2019, after the publication of the article introducing it in October 2018 [DCLT18]. Finetuned BERT models for toxicity detection have been made available on platforms like HuggingFace, with hundreds of models covering different base models (e.g., BERT base uncased, DistilBERT, RoBERTa) and supporting multiple languages [4]. However, BERT-based models have limitations, being less universally applicable across different languages and tasks, and are considered more or less static once trained [LTT+22]. This makes it difficult to apply a single model across a diverse range of languages, domains, and tasks.

As of July 2023, there are over 200[5] BERT-based toxicity detection models on HuggingFace[6] - a platform that provides state-of-the-art models, datasets, and tools for natural language processing (NLP). These are finetuned on different base models (e.g., BERT base uncased [DCLT18], DistilBERT [SDCW19], RoBERTa [LOG+19]) and for different languages (e.g., English, Russian, German, French, Arabic).

[3] More information at: https://www.kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge/overview/description (last visited 07.08.2023)

[4] See more here: https://huggingface.co/models (last visited 07.08.2023

[5] A keyword search of the term 'toxi' on the model page https://huggingface.co/models results in 267 models (07.07.2023).

[6] https://huggingface.co/

2.3 App store reviews for data analysis

Following the widespread adoption of smartphones and mobile apps that led to the establishment of app stores, including Google Play Store, researchers began to recognise the potential of app store reviews as a valuable source of user-generated content for analysis. In the late 2000s and early 2010s, researchers started to focus more on analysing app store reviews to gain insights into user feedback, sentiment, and app quality. Some of the earliest papers using app review data explored the relationship between review length and rating [VHMN12] and investigated how and when users provide feedback, inspected the feedback content, and analysed its impact on the user community [PM13].

As NLP techniques improved and computing resources became more accessible, the analysis of app store reviews gained more attention. Researchers began developing sentiment analysis [Isl14, SFE+19, VKV20, PRW+20, MS22] and topic modeling methods [FLL+13, NL22, Çal23] specifically tailored to app store reviews and conducting case studies for different languages [SÖO+20, CHB21].

The use of Google Play Store data for research purposes, including sentiment analysis, toxicity detection, and feedback analysis, continues to be an ongoing trend. Researchers and industry professionals regularly leverage app store reviews to understand user satisfaction, identify pain points, and improve app experiences. This thesis takes advantage of application reviews data to analyse the prevalence of toxicity in Google Play Store app reviews, temporal trends, and the possible role of external events in shaping the observed patterns.

3 Data and Methodology

This section describes the tools and methods used for obtaining, preprocessing and analysing the data, and presenting the results. I used Google Colab (Pro+ from 24.02.2023), Google Drive for file storage and commonly used Python libraries[7] (e.g., pandas, numpy, matplotlib) throughout the process. In accordance with the Master's thesis guidelines, I acknowledge utilising ChatGPT [Ope22] for various tasks, such as providing conceptual and technical explanations, code snippets and suggestions for formatting in LaTeX.

The rest of this chapter introduces the practical steps taken to gather, preprocess and analyse data and present the results. We start by explaining how we collected app reviews and their details from online sources. Then we describe the existing models for English language and toxicity detection, and present the details of training a new multi-class toxic content classification model. We continue with describing the final datasets and statistical tests used for exploring associations between categorical variables.

3.1 Scraping app reviews and metadata

This section describes the process of gathering app reviews and metadata from the Google Play Store. Google Play Store offers a wide range of apps from different genres (categories), however, it can be challenging to identify the category of recommended apps during navigation. All apps belong to one of 32 genres, such as Art & Design, Comics, Education, Personalization[8], Productivity, Weather, etc. However, when navigating the app store, apps are sometimes sorted and presented differently, showing suggestions from groups such as Recommended for You, Popular Apps, Premium Apps, Podcasts & Radio, etc. Therefore, we used all-time download statistics (as of 14th November 2022) from AppBrain.com for a more structured approach for collecting app reviews from the app store.

AppBrain.com is an online platform that enables discovering, exploring, and managing Android apps. It provides daily updated information on Android apps and offers helpful filters, including sorting by category and number of downloads. We wrote a custom Python scraper using BeautifulSoup library that enabled to gather the ID, name and category for 200 most downloaded apps from each of the 32 categories.

The App Detail (app), App Reviews (reviews) and App All Reviews (reviews_all) functions from the Google Play Scraper[9] enabled us to gather further details about the apps and reviews for the given app ID. For the preliminary analysis, on the 22nd of December 2022, we scraped 1000 most recent reviews for the 10 most downloaded apps from each category and ended up with a dataset of 289,709 reviews from 304 apps

[7] The latest version available in the Python Package Index (PyPI)

[8] While the work follows British English, genre names (e.g., News& Magazines and Personalization) are retained in their original US English form with a 'z'.

[9] Available at: `https://pypi.org/project/google-play-scraper/` (last visited 27.06.2023)

Genre	DS10 scraped (N)	DS10 apps (N)	DS200 scraped (N)	DS200 apps (N)
Art & Design	10,000	10	2,114,855	200
Auto & Vehicles	7181	9	1,383,249	198
Beauty	9034	10	474,011	200
Books & Reference	10,000	9	6,664,734	199
Business	9000	10	8,743,594	200
Comics	9321	10	668,493	196
Communication	10,000	10	24,544,898	199
Dating	10,000	10	1,859,481	200
Education	10,000	10	9,284,233	199
Entertainment	8347	9	25,806,730	200
Events	8130	10	262,432	195
Finance	9000	9	19,038,881	199
Food & Drink	10,000	10	7,286,748	192
Health & Fitness	9000	9	7,864,289	199
House & Home	8696	10	1,308,151	197
Libraries & Demo	2282	7	314,455	196
Lifestyle	9569	10	7,854,676	198
Maps & Navigation	10,000	10	5,698,625	200
Medical	9126	10	2,122,283	197
Music & Audio	10,000	10	16,839,957	199
News & Magazines	9000	9	4,522,255	200
Parenting	10,000	10	639,021	200
Personalization	8000	8	6,229,074	199
Photography	8000	8	13,161,148	199
Productivity	8023	9	15,092,780	200
Shopping	10,000	10	21,437,701	199
Social	10,000	10	17,092,316	200
Sports	10,000	10	3,657,312	200
Tools	10,000	10	21,330,475	200
Travel & Local	10,000	10	9,063,087	200
Video Players & Editors	9000	9	17,229,966	199
Weather	9000	9	3,870,213	199
Total	289,709	304	283,460,123	6358

Table 1. Dataset statistics for datasets DS10 and DS200: information about the number of scraped reviews from distinct apps

(referred to as DS10 from here on). Scraping for the final dataset (referred to as DS200 from here on, containing all available reviews for the 200 most downloaded apps for each of the 32 categories) was completed between the 21st and 27th of February 2023. It took over 300 working hours with multiple scraping sessions running simultaneously, and the scraped dataset contained over 283 million reviews for 6358 apps. The distribution of reviews and apps across categories is presented in Table 1.

3.2 English language and toxicity detection

When analysing toxicity in app reviews, understanding the language and detecting toxic content are pivotal tasks. This subsection delves into the processes of language identification and toxicity detection, highlighting their significance in shaping the ensuing analysis.

3.2.1 English language detection

In this thesis, only reviews in the English language are analysed. Multilingual toxicity classification has gained more popularity in recent years [SHX21, CjSR+22, Ous21, KDB+23]. However, the three main challenges associated with toxicity detection in a multilingual setting - multilingual characteristics, lack of annotated data and imbalanced sample distribution [SHX21] - would have posed a significant obstacle to achieving the research objectives of this thesis. The pragmatic decision to remove non-English reviews reduced the dataset size by approximately 10% but also enabled a more streamlined and manageable analysis.

For this task, several Python libraries for language detection models were tested, including langdetect[10], PYCLD2[11] and Lingua[12]. The developers of Lingua also compared the language detection results of Lingua to other models, including fastText, langdetect, langid, CLD2 and CLD3, and reached a similar conclusion: Lingua outperforms the other models, especially in identifying the language for the shortest texts.

As explained by the developers, Lingua draws on both rule-based and statistical methods to first determine the alphabet of the input text and only then, in a second step, the probabilistic n-gram model is taken into consideration. Additionally, most libraries only use n-grams of size 3 (trigrams) which is satisfactory for detecting the language of longer text fragments consisting of multiple sentences. "For short phrases or single words, however, trigrams are not enough. The shorter the input text is, the less n-grams are available. The probabilities estimated from such few n-grams are not reliable. This is why Lingua makes use of n-grams of sizes 1 up to 5 which results in much more accurate prediction of the correct language." [Pet22] This is important also

[10] `https://pypi.org/project/langdetect/` (last visited 28.06.2023)

[11] `https://pypi.org/project/pycld2/` (last visited 28.06.2023)

[12] `https://pypi.org/project/lingua-language-detector/` (last visited 28.06.2023)

when analysing reviews in app stores because users tend to leave mostly short comments [VHMN12, PM13].

3.2.2 Toxicity detection with ToxiGen RoBERTa

Toxigen is a large-scale machine-generated dataset that contains 274k toxic and benign statements about 13 minority groups. It uses adversarial generation to reduce the bias from spurious correlations of minority group mentions: since those groups are often targets of online hate, toxic language detection systems often falsely flag text that contains minority group mentions as toxic [HGP$^+$22]. This dataset was used to pretrain ToxiGen RoBERTa[13] - a RoBERTa-based model specifically fine-tuned for toxicity detection, making it suitable for identifying toxic content, hate speech, offensive language, and other forms of harmful communication. So, this model was also used for toxicity detection on DS10 data.

ToxiGen RoBERTa classifies 5823 reviews (2%) of DS10 (N=289,709) as toxic and nearly 284k as non-toxic. However, when looking through the classification results, two things can be noted:

1. Two different types of toxic reviews emerge: **toxic** reviews that only contain insults, vulgarity, cursing and/or hate speech; and **toxic-critical** reviews that contain toxic content but also criticism about a certain feature, aspect or development of the app.
2. There is a substantial amount of reviews that do not come across as toxic, though classified as such by the model. From the 5823 reviews classified as toxic by ToxiGen RoBERTa, 1855 (over 30%) were later classified as **critical** and **non-toxic** (more information in the next subsection). This finding underscores the significance of domain-specific nuances in toxicity detection [KSA22], highlighting the need for diverse and contextually relevant data sources to ensure enhanced model performance.

Different harmful language detection models focus on various aspects of toxicity. A two-fold typology, which has also found practical use [PTDA19], assesses whether (i) the abuse targets a specific entity and (ii) the degree of its explicitness or implicitness [WDWW17]. In certain contexts, it may be important to discern whether a comment singles out a particular target or holds a more general nature, as well as whether the toxicity is conveyed directly or inferred from the text. For instance, Google's Perspective API offers scores for Severe Toxicity, Insult, Profanity, Identity attack, Threat, and Sexually explicit content[14]. However, in the context of moderating app reviews, we

[13] The data and information on training the model are available in a GitHub repositoryhttps://github.com/microsoft/TOXIGEN (last visited 06.08.2023)

[14] For more information, visit the developer's site: https://perspectiveapi.com/how-it-works/. The original (capitalised) names of the attributes are retained.

contend that the *target*, *severity*, and *type* of toxicity are less significant than the rest of the content itself (if present). While removing outright **toxic** comments preserves the integrity of the review section and eliminates content policy violations, **toxic-critical** reviews, which contain valuable user-reported issues, should be treated differently. Mechanisms that allow developers to access comments violating content policies while keeping them shielded from public view could be considered.

3.3 Finetuning a multi-class toxicity detection model

3.3.1 Labeling training data

The quality and constructiveness of app reviews varies widely [PM13], making detecting toxicity and criticism in reviews a challenging task. So, in this section, we examine the notions of toxicity and criticality, the annotation process, disagreements, and the steps taken to label the training data for finetuning a multi-class toxicity detection model.

Toxicity can manifest it differently as exemplified here:

- **Profanity**: "This game sucks"
- **Hate speech**: "I hate this app"
- **Insult**: "Absolute garbage"
- **Cursing**: "[...] I hope all of the executives choke on their own feces. [...]"
- **General abuse**: "Fu#@ing terrible"
- **Targeted abuse**: "The NFL is a Monster, they need to have their Anti-Trust removed and reviewed every year by Fans, Investors, Owners, then Congress!!"
- **Explicit remarks**: "Sucks dick."
- **Implicit remarks**: "Was a great app now pile of male bovine"

Similarly, **criticism** can explicitly mention or imply an aspect about the app or its developers that they found upsetting, not meeting their expectations or that they feel should be improved on. This could be anything from but not limited to the following examples:

- **Bugs**: "Keeps crashing. Says it has a bug. This is very bad as I need in my Gmail many times a day for work."
- **Compatibility issues**: "I can't use this app on my phone, which is why."
- **Privacy or security concerns**: "There is no privacy for the child."

- **Changes in UI or UX**: “Now is very bad app, not open, just loading.”
- **Changes in features or monetisation strategy**: “So disappointed that you are now making the scan barcode function a premium paid function. Ridiculous!””
- **Updates**: “It seems that is a very nice app, sadly for Colombia, it needs to be updated a lot.”
- **Experiences with customer service**: “I wish the customer service agents weren’t outsourced! They don’t speak English well and just don’t get it. A few glitches but overall I’m thankful for the app. Please get better agents!”

These examples would later serve as guidelines in the annotation process where 5823 toxic and a random subset of 1677 non-toxic reviews - as classified by ToxiGen RoBERTa - reviews from dataset DS10 were labeled and then used for finetuning a new multiclass model. The data was classified into four categories: **toxic** (T), **toxic-critical** (TC), **critical** (C), and **non-toxic** (NT).

The initial round of annotations served as the foundational mapping of diverse example types and toxicity categories. Additionally, a subset of 1000 reviews was annotated by one of the thesis supervisors with ensuing discussions to develop more nuanced annotation guidelines. The inter-annotator agreement score, measured by Cohen’s kappa coefficient (κ) [Coh60], exhibited a moderate level of agreement at 0.54. Among the instances of disagreement (N=340), there were comments featuring expressions of hate (e.g., *“I hate it”*), insults (e.g., *“This is app is the worse than every other app”*, *“U guys nuts”* or *“Good app but has a lot of stupid peope”*) and various single-word comments where comprehending the intended meaning was challenging (e.g., *“Diabolical”*, *“dope”* or *“Silly”*).

These disagreements were also labeled by an expert annotator. Subsequently, 20 reviews that all three annotators classified differently were discussed in a dedicated meeting. After a thorough review of the labeling disparities and group deliberations, all 7500 reviews were subjected to a final round of annotation. In this stage, 902 reviews, primarily those containing or not containing hate speech and insults, were reevaluated and assigned revised labels. The final distribution and illustrative examples of reviews for the newly defined categories are presented in Table 2.

3.3.2 Toxic content classification models

There are several approaches that can be used to train a multi-class text classifier. This section presents an overview of training and comparing six different models that can be categorised into three main groups based on their underlying methodologies:

1. Traditional Machine Learning (ML) Models:

Label	Description	Examples	N	Share
Toxic	Reviews that only contain insults / vulgarity / cursing / hate speech	“Sucks” “Screw Google” “Garbage.”	1286	17%
Toxic-critical	Reviews that contain both toxic content and criticism about a certain feature/aspect/development of the app	“New update is terrible. Hate it. The app is so dumbed down it’s no longer usable.” “Too many f****** advertisements” “crashes. sucks.”	2756	37%
Critical	Reviews that only contain criticism about a certain feature / aspect / development of the app	“Stop ad please” “Full of bots. Not worth your time” “It is Full of bug”	2092	28%
Non-toxic	Reviews that do not contain criticism nor toxicity, i.e., are neutral, positive (including praise) or illegible (do not make any sense).	“Ok” “on time service and great professionalism by agents working under different fields” “brilliant aap”	1366	18%

Table 2. Classification and examples of reviews

- Support Vector Machine (SVM)
- Logistic Regression (LR)
- Multinomial Naive Bayes

2. Recurrent Neural Network (RNN) Models:

   - Long Short-Term Memory Networks (LSTMs)

3. Transformer-Based Language Models:

   - DistilBERT
   - ToxiGen RoBERTa

Support Vector Machine (SVM), Logistic Regression (LR), and Multinomial Naive Bayes are traditional machine learning approaches used for text classification. These models rely on feature extraction from text data and are trained to differentiate between different classes based on these features. In these methods, the text is typically transformed into a numerical representation using techniques such as TF-IDF (Term Frequency-Inverse Document Frequency) or bag-of-words. These representations are then used as input features for the models. Before training, these approaches require several preprocessing steps that help improve the performance of the model. Often this involves cleaning and tokenising the text, removing punctuation, and converting words to lowercase. Stopwords may be removed, and stemming or lemmatisation can be applied to reduce words to their root forms.

Long Short-Term Memory Networks (LSTMs) are a type of Recurrent Neural Network (RNN) that can capture sequential dependencies in text data. LSTMs are well-suited for tasks where the order of words matters. They process text data in a step-by-step manner, considering the current word along with the information from previous words. Similarly, they require text preprocessing steps to be taken beforehand, involving tokenisation, converting words to numerical embeddings, and padding sequences to a consistent length.

Transformer-Based Language Models (also referred to as attention-based models), such as DistilBERT and RoBERTa, are advanced neural networks that excel at understanding context and relationships within text. They utilise the attention mechanism, allowing them to prioritise essential words, even considering their positions within the sequence. These models often include their own text preprocessing steps and tokenisation procedures. They are pre-trained on massive text corpora and have their own mechanisms for handling casing, stopwords, and special characters. Removing stopwords and lemmatisation are often not necessary due to the contextualised embeddings they generate.

Training the models To ensure robust results, the dataset was split into distinct subsets for training, validation, and testing. The data was randomly shuffled, and a train/test split was performed, allocating 75% of the reviews (N=5625) to the training set, 13% (N=975) to the validation set, and 12% (N=900) to the test set. The division preserved class distribution across the subsets, enabling the models to generalise effectively.

The model training process involved a systematic approach to optimising various hyperparameters to achieve the best performance for each classification model. For example, in the case of Support Vector Machine (SVM), Logistic Regression (LR), and Naive Bayes (NB), a grid search was conducted over a range of hyperparameter values, including regularisation strength and kernel choice for SVM, and regularisation strength and solver for LR. The best ML models had the following hyperparameters: SVM: $c = 1$, kernel=rbf; LR: $c = 1$, solver=lbfgs; and NB: $\alpha = 0.5$, fit_prior=False. Similarly, the Long Short-Term Memory Network (LSTM) underwent grid search, optimising model architecture by exploring various LSTM units and dropout rates, where 50 LSTM units and 0.2 achieved the best performance. The best transformer-based models (DistilBERT and RoBERTa) were trained with learning rates of 1×10^{-7} and 2×10^{-7}, and weight decay of 0.8 and 0.9, respectively. Furthermore, necessary preprocessing steps, including lowercasing, tokenisation, and padding, were applied to models that required it to ensure a fair comparison.

Base model	**Accuracy**	**F1**	**Precision**	**Recall**
SVM	0.77	0.78	0.79	0.77
Logistic Regression	0.73	0.73	0.73	0.73
Naive Bayes	0.72	0.72	0.73	0.72
LSTM	0.72	0.72	0.73	0.72
DistilBERT	0.87	0.87	0.87	0.87
ToxiGen RoBERTa	**0.88**	**0.88**	**0.88**	**0.88**

Table 3. Classification metrics for different models

Evaluating the models From Table 3 we can see that the selected traditional machine learning models (SVM, Logistic Regression, Naive Bayes) and neural network models (LSTM) performed similarly on the held-out test set of 900 reviews: all reached accuracy and F1 scores between 0.72 and 0.77. The best finetuned BERT-based models (DistilBERT and ToxiGen RoBERTa) got 0.87 and 0.88 weighted accuracy and F1 scores.

While the performance metrics for both DistilBERT and ToxiGen RoBERTa-based models are very similar with only 0.01 difference on the test set, the finetuned model based on ToxiGen RoBERTa was chosen (referred to as ReviewBERT from now on) over DistilBERT's due to task relevance and domain-specific pretraining. ToxiGen RoBERTa was designed for a toxicity detection task and has already learned relevant features and representations that are useful for identifying toxic content, including mitigating the bias for minority group mentions.

3.3.3 reviewBERT

The selected model - reviewBERT - achieved 0.88 weighted accuracy and F1 score on the test set. Among the 900 reviews reserved for final testing, 768 reviews had matching predicted and actual labels. Figure 1 provides valuable insights into the classification performance of the model. Each cell in the heatmap displays the percentage of instances that belong to the true label and were predicted as the corresponding predicted label. Notably, the model faced challenges in distinguishing between critical and toxic-critical reviews, as 15.7 and 16.7% of critical and toxic reviews respectively were classified as toxic-critical.

A detailed analysis of the mismatches (N=132) between predicted and actual labels reveals some weaknesses of reviewBERT:

1. **Spelling mistakes**: the model occasionally struggled to comprehend comments with spelling mistakes, classifying terms like "Fabolous" or "Very Exllant" as toxic or "stints" as non-toxic.

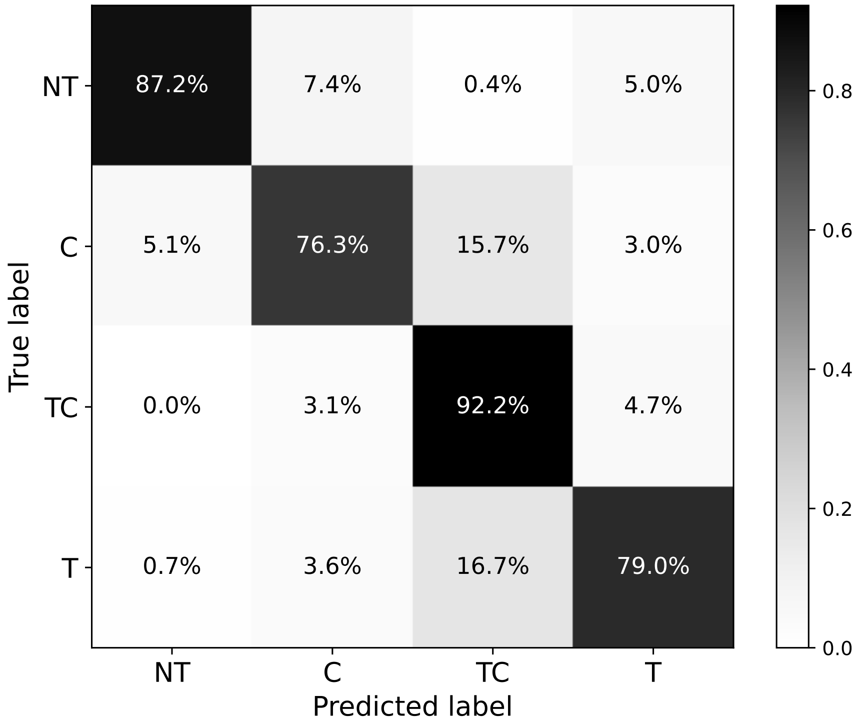


Figure 1. Predicted and true labels for test data (N=900) for reviewBERT

2. **Not detecting criticism**: reviewBERT can sometimes fail to detect criticism in a review, misclassifying feedback like "It crashes frequently" or "Uses to much memory" as non-toxic.

3. **Misclassification of targeted insults**: the model can categorise targeted insults as toxic-critical feedback, for instance, in cases such as "Disney and espn are trash", "U suck Ticketmaster ! ! !" and "Nosey friggin Google aps".

In conclusion, the analysis presented in this section can underscore the performance of reviewBERT in the context of toxicity classification for app reviews. With a weighted accuracy and F1 score of 0.88 on the test set, reviewBERT demonstrates a high level of proficiency in identifying toxic content, critical feedback, and other sentiment nuances. While a few limitations were observed, such as occasional struggles with spelling mistakes and targeted insults, these minor shortcomings do not diminish the overall effectiveness and utility of the model.

3.4 Scaling up the dataset

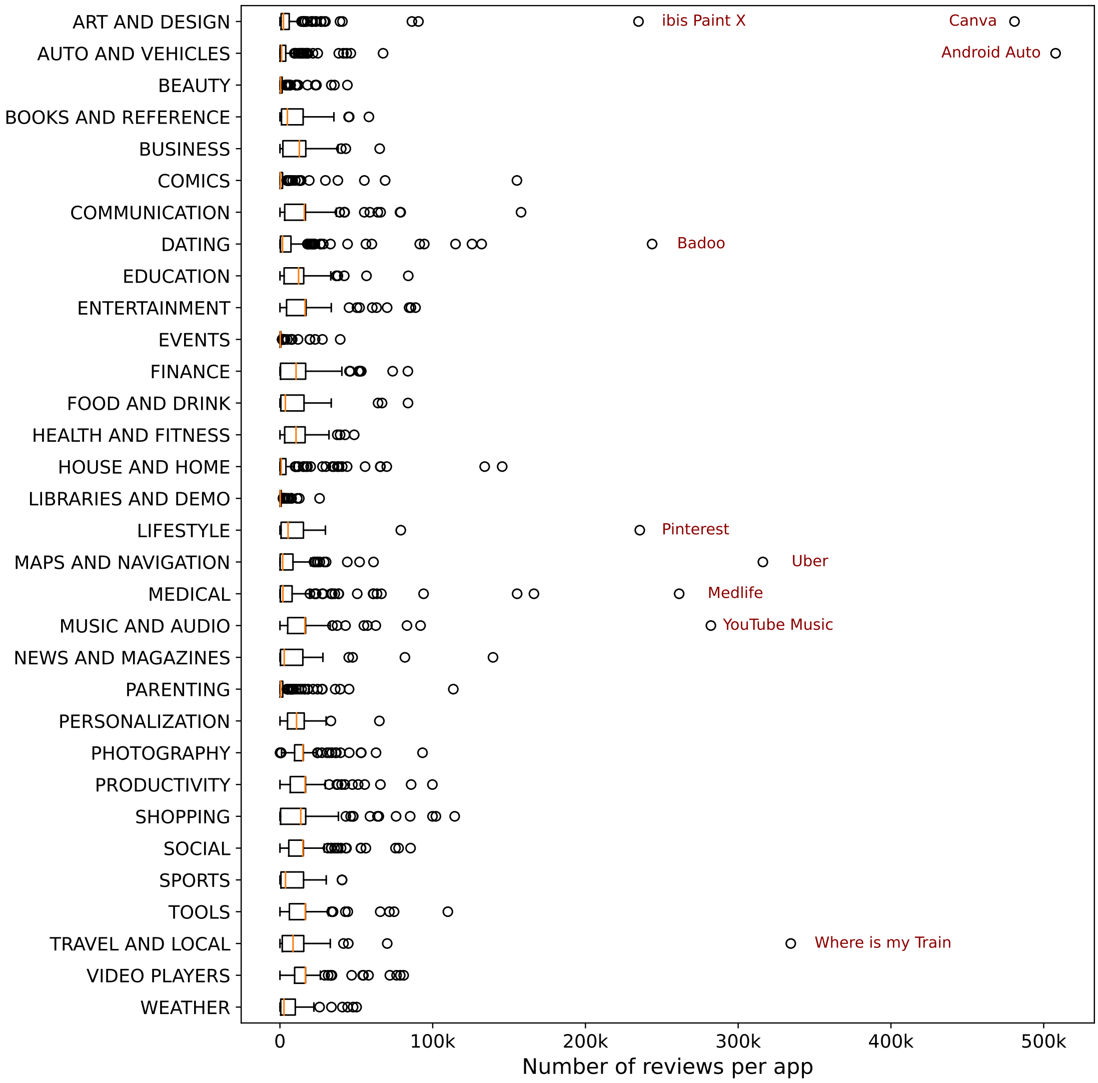


Figure 2. Number of reviews per app in different genres

To conduct an extensive analysis of toxicity in Google Play Store reviews, data from dataset DS200, comprising of 283 million scraped reviews across 6358 apps, was used. To streamline the dataset size prior to deploying the language detection (Lingua) and toxicity classification (reviewBERT) models, a method of stratified quota sampling was employed. In categories where the total review count remained below 2.5 million (N=10[15]), the entire set of reviews was retained in the final dataset. For categories with

[15] This includes the following genres: Art & Design, Auto & Vehicles, Beauty, Comics, Dating, Events,

Genre	N reviews	N apps
Art & Design	1,828,166	193
Auto & Vehicles	1,227,726	164
Beauty	382,282	178
Books & Reference	1,685,216	191
Business	2,221,021	193
Comics	581,366	145
Communication	2,750,440	187
Dating	1,797,327	193
Education	2,203,187	193
Entertainment	2,810,576	195
Events	231,789	145
Finance	2,377,083	198
Food & Drink	1,626,793	184
Health & Fitness	2,155,138	192
House & Home	1,171,097	163
Libraries & Demo	166,125	157
Lifestyle	1,717,937	172
Maps & Navigation	1,363,450	176
Medical	1,853,634	171
Music & Audio	2,918,864	191
News & Magazines	1,395,167	173
Parenting	641,105	184
Personalization	2,093,828	190
Photography	2,759,110	189
Productivity	2,704,583	186
Shopping	2,679,786	194
Social	2,727,508	191
Sports	1,446,537	192
Tools	2,803,049	191
Travel & Local	2,145,714	182
Video Players & Editors	3,018,115	191
Weather	1,183,435	179
Total	58,667,154	5823

Table 4. Dataset statistics for the final dataset used for data analysis: information about the number of reviews and distinct apps

larger review volumes (N=22[16]), a randomised subset of 15k reviews was chosen for apps hosting up to 1 million reviews. In cases where apps exceeded the 1 million review mark, a random subsample equivalent to 10% of the total reviews was selected. This approach ensured that while apps featuring up to 15k reviews retained all their entries, an increment in review count saw a proportional reduction in the proportion of reviews included in the analysis, ensuring that the share never dipped below 10% of the complete review corpus.

During the preprocessing of data, apps with fewer than 10 reviews were excluded to improve dataset reliability. To enhance data quality and reduce redundancy, duplicate reviews (based on review ID) were removed. Furthermore, a time-based filter was applied to ensure dataset relevance. Reviews prior to 2014, up to the end of 2013, were omitted, while reviews starting from January 2014 onwards were included in the final dataset. Table 4 provides an overview of the total number of reviews and apps included in the final dataset, while Figure 2 shows the distribution in the number of reviews for different genres. The names of apps with more than 200k reviews in the final dataset are marked on the plot.

Finally, another subset of all reviews from four popular apps by Chinese and four by Russian developers was selected for exploring the relationship between app reviews and external events. More details about the selected apps are provided in Table 5.

3.5 Analysing the data and presenting the results

Data analysis includes both descriptive statistics, trends and statistical tests. All plots were made using the matplotlib library[17]. From statistical tools, the Chi-squared (χ^2) test and Cramér's V were used to analyse the association between the app's metadata and the share of reviews in different toxicity categories and Bonferroni correction to counteract the multiple comparisons problem. χ^2 test and Cramér's V are statistical tools that are commonly used in the field of statistics to analyse the association between categorical variables while Bonferroni correction helps in family-wise error rate (FWER) control.

The χ^2 test is a fundamental statistical technique used to determine if there is a significant association between two categorical variables, proposed by Karl Pearson in 1900 [Pea00]. It is particularly valuable when working with nominal or ordinal data, such as survey responses, where variables have distinct categories but lack a natural numerical order. The test assesses whether the observed frequencies of categorical data

House & Home, Libraries & Demo, Medical, Parenting

[16] This includes the following genres: Books & Reference, Business, Communication, Education, Entertainment, Finance, Food & Drink, Health & Fitness, Lifestyle, Maps & Navigation, Music & Audio, News & Magazines, Personalization, Photography, Productivity, Shopping, Social, Sports, Tools, Travel, Video Players & Editors, Weather

[17] More information available here: `https://pypi.org/project/matplotlib/` (last visited (07.08.2023)

Table 5. Selected Chinese and Russian apps

Chinese Apps					
App name	Developer	App score	Downloads	N reviews	Genre
UC Browser-Safe, Fast, Private	UCWeb Singapore Pte. Ltd.	4.3	1B+	2,078,386	Communication
SHAREit: Transfer, Share Files	Smart Media4U Technology Pte.Ltd.	4.2	1B+	1,548,960	Tools
Likee - Community of Interests	Likeme Pte. Ltd.	4.4	500M+	1,041,415	Social
SHEIN-Fashion Shopping Online	Roadget Business PTE. LTD.	4.7	100M+	353,409	Shopping
Russian Apps					
App name	Developer	App score	Downloads	N reviews	Genre
Kaspersky Antivirus & VPN	Kaspersky ME	4.7	100M+	255,607	Tools
VK: music, video, messenger	VK.com	3.7	100M+	22,337	Social
Yandex Browser	Intertech Services AG	4.4	100M+	14,500	Personalization
OK: Social Network	Odnoklassniki Ltd	4.2	100M+	3899	Social

significantly deviate from the expected frequencies that would occur under a specified hypothesis of independence between the variables. χ^2 is defined by:

$$\chi^2 = \sum \left(\frac{O_i - E_i}{E_i} \right)^2$$

where O_i is the observed value (actual value) and E_i is the expected value. The resulting statistic is then compared to a critical value from the χ^2 distribution to determine whether the association is statistically significant. If the calculated statistic exceeds the critical value, it suggests that the variables are dependent, indicating a potential relationship between them.

Cramér's V, on the other hand, is a measure of association that complements the χ^2 test by quantifying the strength of the relationship between categorical variables that was introduced by Harald Cramér in 1946. While the χ^2 test provides information about the presence or absence of an association, Cramér's V goes a step further by indicating the magnitude of the association. Cramér's V is computed by taking the square root of the chi-squared statistic divided by the sample size and the minimum of the two dimensions minus 1:

$$V = \sqrt{\frac{\varphi^2}{\min(k-1, r-1)}} = \sqrt{\frac{\chi^2/n}{\min(k-1, r-1)}}$$

where:

- φ is the phi coefficient.
- χ^2 is derived from Pearson's chi-squared test.
- n is the total number of observations.
- k represents the number of columns.
- r represents the number of rows.

The resulting value ranges from 0 to 1, with higher values indicating a stronger association between the variables. This makes Cramér's V a valuable tool for comparing the strength of associations across different datasets or scenarios.

We also apply Bonferroni correction to counteract the multiple comparisons problem when analysing the relationship between app genres and other metadata, and toxicity categories. Statistical hypothesis testing is based on rejecting the null hypothesis if the likelihood of the observed data under the null hypotheses is low. If multiple hypotheses are tested, the probability of observing a rare event increases, and therefore, the likelihood of incorrectly rejecting a null hypothesis (i.e., making a Type I error) increases [MJM00].

The Bonferroni correction compensates for that increase by testing each individual hypothesis at a significance level of $\frac{\alpha}{m}$, where α is the desired overall alpha level and m is the number of hypotheses [RJ+12].

4 Results and Discussion

In this chapter, we turn our attention to the research questions that guided our study. We begin by examining the extent to which toxicity appears in Google Play Store app reviews, considering whether it appears consistently across different types of apps. We then explore whether certain details in the app information have a stronger connection to the presence of toxic content. Additionally, we investigate whether the prevalence of toxicity has changed over the last decade, shedding light on any potential trends. Finally, we look into the impact of external events on toxic content in app reviews. Specifically, we explore whether significant events, such as the COVID-19 pandemic and the Russian invasion of Ukraine, influenced a rise in negative comments for Chinese and Russian apps. Throughout, we establish links to relevant existing research in the field.

4.1 Toxicity and app genres

In this section, we revisit our first research question:

RQ1: *How prevalent is toxicity in Google Play Store app reviews, and does it exhibit similarities or differences across various app genres?*

We explore the relationship between toxicity categories (toxic, toxic-critical, critical, and non-toxic) and app genres (e.g., Art & Design, Dating, Lifestyle, Weather, etc.): how the prevalence of toxicity varies across different genres and discuss potential factors influencing these patterns. Figure 3 below shows that there are only small differences in the shares of toxic and toxic-critical reviews between app genres. On average, about 1% of all reviews are toxic, 2.5% toxic-critical, 16.5% critical and 80% non-toxic (neutral or positive comments). We can see that apps in the Dating and Food & Drink categories have slightly larger shares of toxic-critical (5-6%) and critical (28-29%) reviews. These differences, however, appear marginal as an overwhelming majority (63-85%) of the reviews in all genres are non-toxic.

Statistical tests, such as Pearson's χ^2-test[18], can be used to determine whether there is a statistically significant difference between the expected frequencies and the observed frequencies, while measures of association like Cramér's V[19] can be used to quantify the strength of association between two variables. The results of the χ^2-test (1077856.71, $p =$ *<0.001*) show that the resulting distribution of reviews between different app genre and toxicity categories is highly unlikely to have occurred by chance alone and there is a strong statistical evidence of an association between the variables. However, the Cramér's V value of 0.08 falls into the **negligible category** and suggests a discrepancy between the statistical significance of the association and the practical significance of

[18] More info at: `https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.chisquare.html` (last visited 24.07.2023)

[19] More info at: `https://www.statology.org/cramers-v-in-python/`(last visited 24.07.2023)

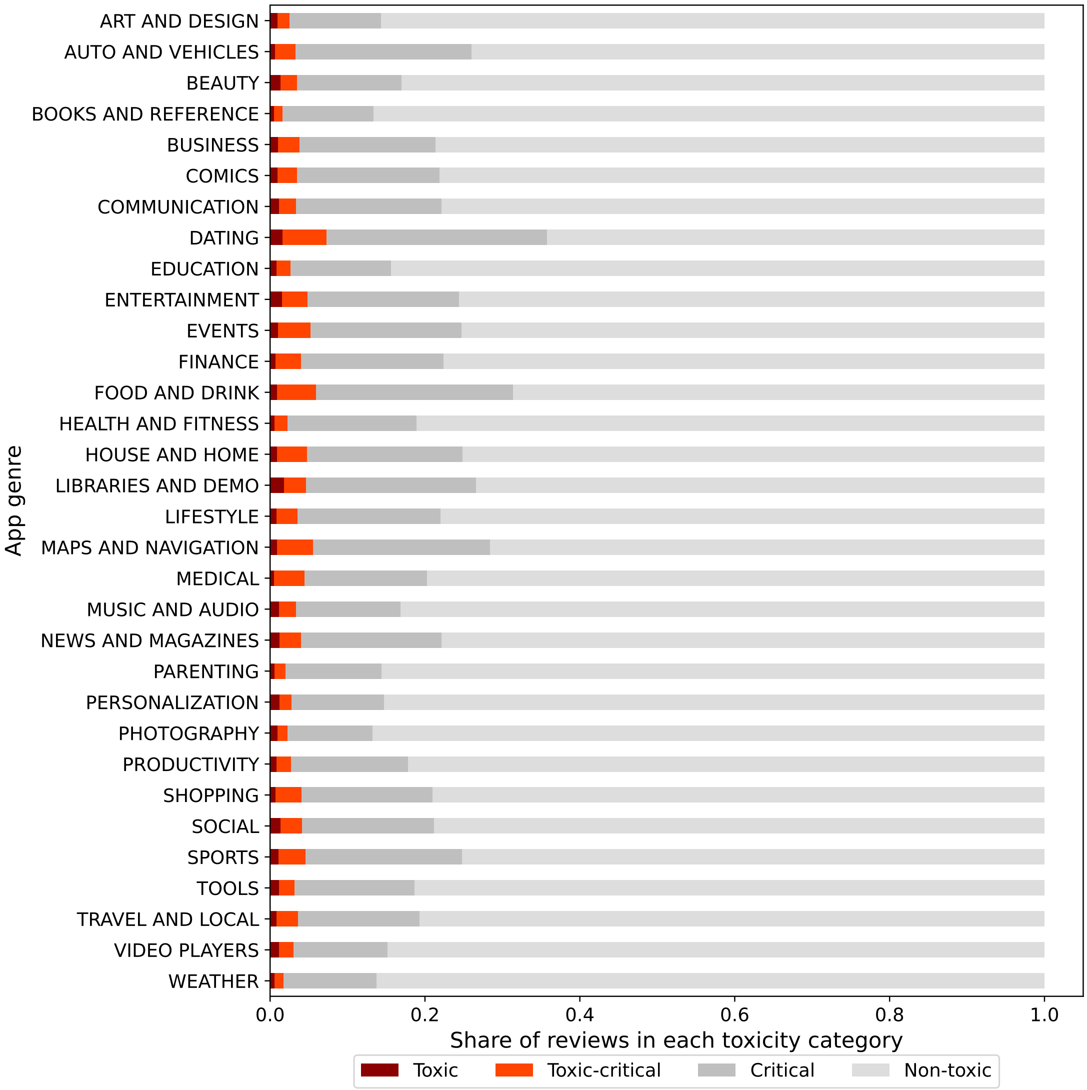


Figure 3. Share of reviews in each toxicity category for 32 app genres

the relationship. In other words, there may exist a relationship between app genres and toxicity categories, but the practical or substantive significance of that association is limited.

[VHMN12] faced a somewhat comparable puzzle when they compared app reviews from 22 categories from Apple App Store. Given their dataset size (8.7M reviews from 17,330 apps), they anticipated that the law of large numbers would to come into play and hence expected review lengths to be similar across categories. Interestingly, though, the ANOVA test clearly showed that review lengths differed significantly across the 22 categories (p < *0.01*). Although they could identify potential causes for short reviews and why poor ratings tend to elicit longer reviews, they were not able to present a strong hypothesis that would explain why the category of an app has an influence on the review length.

The variation in the levels of toxic or toxic-critical reviews across different genres of apps could be influenced by a range of factors or combinations thereof, including:

1. **Controversial content**: Categories and apps that deal with sensitive or controversial subjects, such as politics, religion, or social issues, could attract more polarised opinions and potentially toxic discussions.

2. **High Expectations**: Categories involving education, productivity, or utility apps (tools) might have higher expectations from users. If these apps don't meet expectations, users find them difficult to use or if they experience technical issues, users might express frustration more readily, leading to an increase in toxic or toxic-critical comments.

3. **Competitiveness**: Categories with intense competition could also result in users being more critical in their reviews, especially if they are comparing analogous apps and finding flaws.

Examining these relationships, however, would require careful research design and additional analysis.

4.2 Toxicity and other variables

This section focuses on the second research question:

RQ2*: Are specific variables in app metadata more strongly correlated with the presence of toxicity?*

We explore associations between toxicity categories and other variables corresponding to the available app metadata: app content rating (*e.g., for adults, everyone, teens*), number of installs (*downloads*), app score (*i.e., how many stars the app has*), cost of the app (*free or paid*), ads (*whether it contains ads or not*), or release year when the app was first launched. As in the previous section, we find that there is evidence of association for all variables, even when applying Bonferroni correction to minimise Type I errors in a multiple comparison setting, but the practical significance of the association is limited (Table 6).

Even though apps with lower scores (i.e., fewer stars) appear to have more toxic, toxic-critical and critical reviews (Figure 4), the effect size is still small (Table 6). This is interesting as one might expect that an app that has gathered a substantial amount of negative ratings would also have amassed a fair share of toxic comments. This is even more puzzling as there is a significant association between the review score (i.e.

	χ^2	Cramér's V	Effect size*
Categories	1077856.72	0.078	small
Content rating	190220.36	0.033	negligible
Installs	128614.86	0.027	negligible
App score	1421733.03	0.090	small
Free or paid	239.05	0.002	negligible
Contains ads	380591.48	0.081	negligible
Release year	227249.64	0.036	negligible
Review score	25988293.07	0.384	large

* The effect size is determined based on the degrees of freedom, which is the smallest of #rows-1 or #columns-1 of the underlying contingency table.

Table 6. Statistical test results and effect sizes. All test results are statistically significant at *p<0.001*.

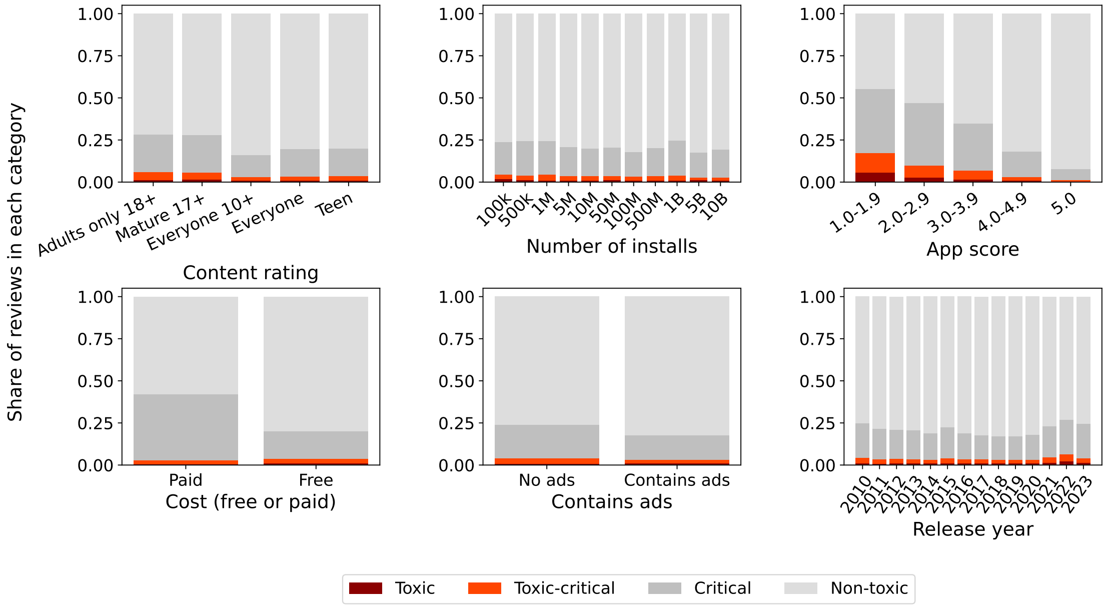


Figure 4. Share of reviews in each toxicity category for different variables

the number of stars accompanying the specific review) and toxicity groups: users who rate the app poorly also tend to leave more toxic and/or critical reviews (Table 6). This finding corresponds to that of [MS22], who compared 1200 reviews from 18 apps in Google Play Store (a total of 21,600 reviews) from three selected categories (Service, Pharma and Travel), and found that higher review scores correlate with lower toxicity levels. Similarly, based on manual content analysis of 528 reviews[20] in the Apple App Store, they concluded that users tend to become insulting quickly, especially when they have spent money. Therefore, we could also expect paid apps to contain more toxic content than free apps. Also, as overall attitudes of app users are negative towards advertisements [Ayd16, GNH17], we could expect that apps that contain adds also have more toxic content due to that. Based on the results of this study, however, no meaningful relationships between toxicity and app metadata can be confirmed.

4.3 Temporal trends

In this section, we analyse the trends related to comment dynamics in different categories of apps across nearly a decade, covering reviews from January 2014 to January 2023, and aim to answer our third research question:

RQ3: *Has the prevalence of toxicity increased over the past decade?*

Figure 5 illustrates that the share of **toxic** comments among all reviews has remained relatively constant for nearly a decade. The share of **toxic-critical** comments appears to be increasing gradually, showing a growth of over one percentage point, from below 2% to slightly over 3%, over a span of nine years. This could be due to increased experience and variety of apps as users become more demanding or have higher expectations for quality, performance, or features.

Additionally, anonymity and social dynamics or changes in user demographics can influence the types of comments being expressed. For instance, if the user base becomes more diverse in terms of age, background, or interests, this could contribute to a wider range of critical viewpoints being expressed. On the other hand, online anonymity can embolden individuals to express their opinions more strongly, including negative ones.

Finally, changes in platform policies, content moderation, or guidelines could impact the types of comments that are prevalent. Users may adapt their communication style in response to these policy changes and deploy newer and more subtle techniques [SGK^{+}23].

These trends, however, are not universal across different app genres. In some categories, e.g., Books & Reference, Personalization, and Photography, the share of both toxic and toxic-critical reviews has remained constantly low throughout the years (Figure 6). In others, like Dating, Finance, House & Home, Maps & Navigation, Medical, and

[20] This included 12 reviews from each of the 22 app categories for both free and paid apps

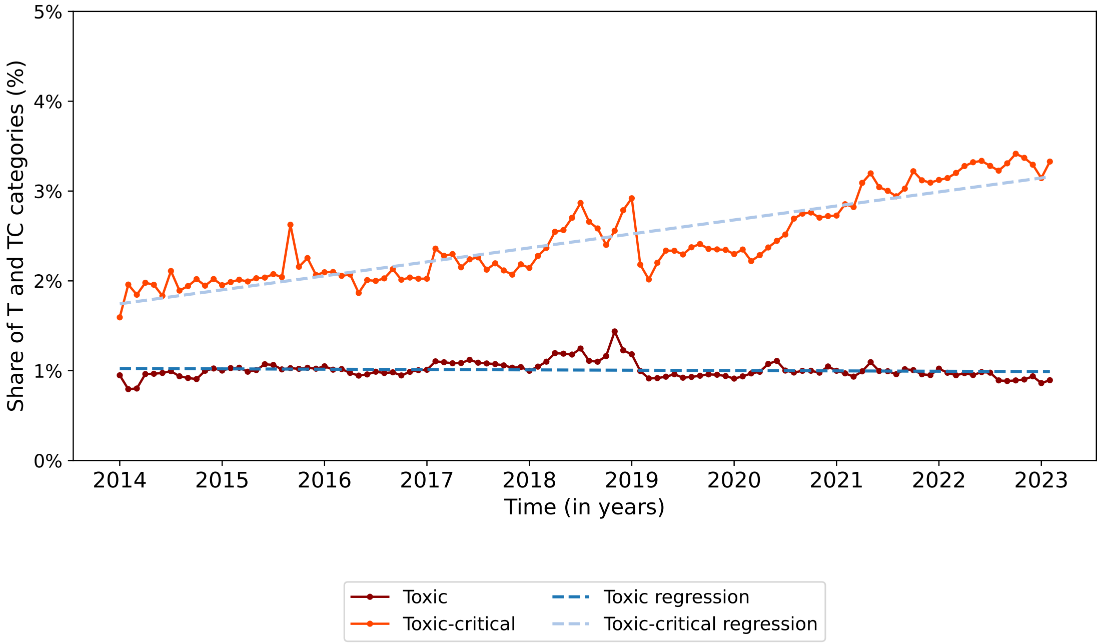


Figure 5. Share of toxic and toxic-critical reviews across time from 2014 to 2023

Shopping, a similar trend to the overall pattern appears: the share of toxic comments remains constant at around 1% of all reviews while the share of toxic-critical reviews increases over time from around 2% to 5-7%.

Additionally, we can observe noticeable fluctuations in the Events and Libraries and Demo categories, which are also the smallest with only 166,125 and 231,789 reviews. We can also see some distinct peaks and valleys in other categories, such as in Dating in 2017, Finance and Medical in 2021 and Maps and Navigation in 2022.

Looking at the sharpest peaks and valleys in the toxic-critical reviews, some of them can be more easily explained than others. Fluctuations at the beginning of the observable period (2014-2015) and in certain categories (e.g., Events and Libraries&Demo) are mostly due to sparse data: a few toxic-critical comments more or less can already have a significant effect on the total trend.

It is also worthwhile to remember that when we are looking at the share of toxic and toxic-critical reviews and how it has changed over time, then the trend lines (marked with orange and dark red in Figure 6) for each genre represent the average monthly shares of individual apps that may have completely different and distinct patterns. From looking into trend lines for individual apps, we can see that a single app with a large number of reviews can have an outstanding effect on the displayed results for the whole category. For example, both *Badoo: Dating. Chat. Meet.*[21] in the Dating category and

[21] More info from: `https://play.google.com/store/apps/details?id=com.badoo.mobile` (last

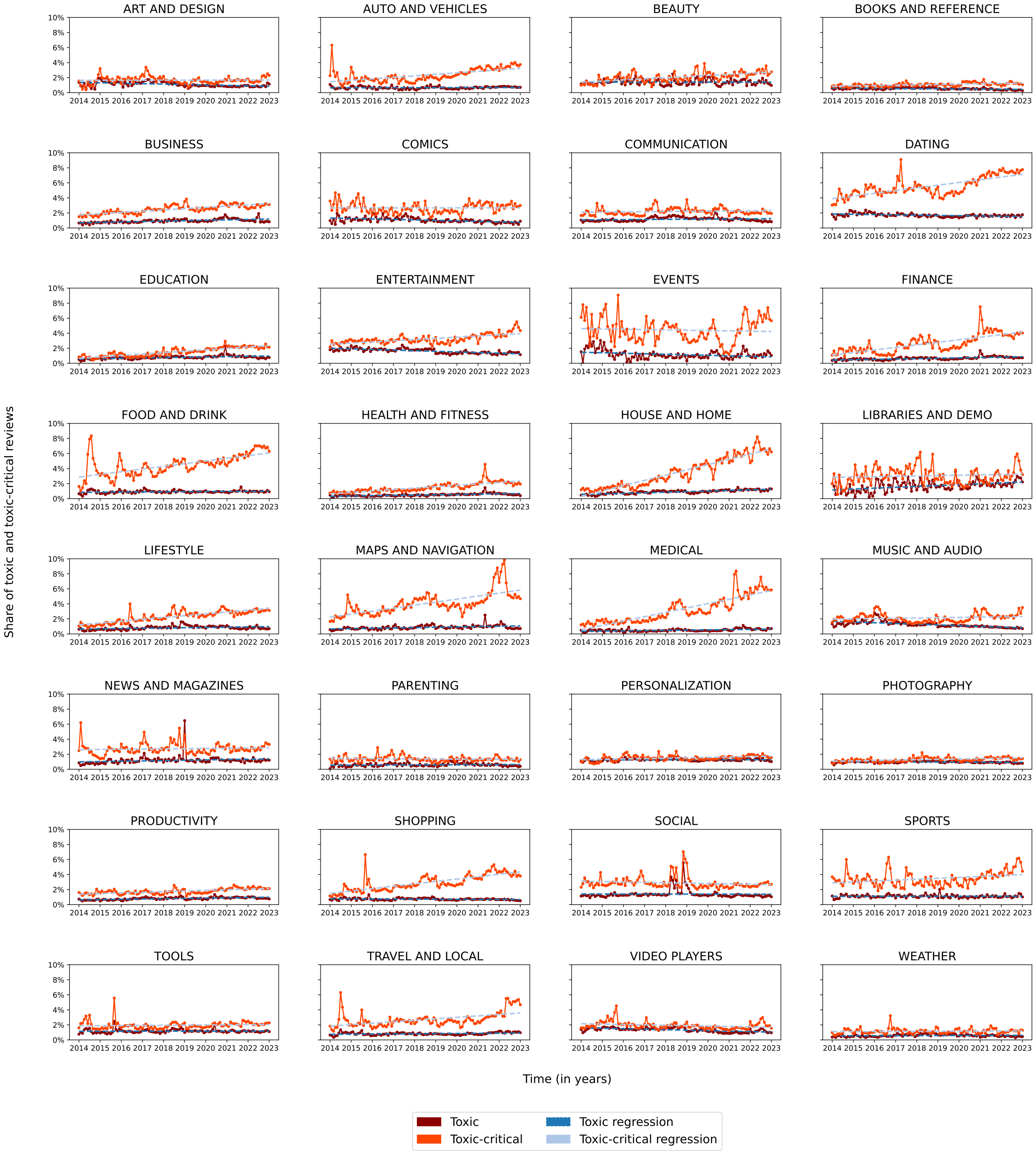


Figure 6. Share of toxic and toxic-critical reviews across time from 2014 to 2023 in different app genres

Tumblr - a microblogging and social networking website from Social category - saw an unprecedented surge in both the number and share of toxic-critical reviews reaching nearly 20% of all reviews in April 2017 and December 2018 respectively, resulting in noticeable peaks for the whole category (Figure 6). Both of these apps undertook major changes at the time: *Badoo* launched a newly redesigned app and brand in April 2017 [Coo17] while *Tumblr* introduced a stricter content policy with heavier restrictions on adult content [Lia19], both of which resulted in the users' display of disapproval of the recent changes.

[HBH18] focus their study on the top 250 bad updates[22] from 26,726 updates of 2526 top free-to-download apps in the Google Play Store, and argue that app-level analysis misses the point that users post reviews to provide their feedback on a certain update. They find that feature removal and UI issues have the highest increase in the percentage of negative reviews. It can be expected that numerous toxic comments are also born out of frustration with unexpected and unwanted changes, particularly shortly after new releases when most of the feedback is provided [PM13]. We are also likely to miss many of them by only looking at trend lines on app or app category level.

4.4 External events

In this section we explore our fourth research question:

RQ4: *Do external events contribute to the presence of toxicity in app reviews? More specifically, did the COVID-19 pandemic and Russian invasion of Ukraine lead to an upsurge of negative comments in the review section of Chinese and Russian apps?*

While it could be expected that user feedback is influenced by changes in the UX/UI, bugs and incompatibility issues or changes in the overall monetisation strategy, there can be external factors not related to the app or its functioning, influencing the review sections of apps. Therefore, in this subsection we explore the relationship between app reviews and two major events in recent history: first, the beginning of the COVID-19 pandemic[23] and global lockdowns (March-April 2020)[24]; and second, the Russian invasion of Ukraine that began on the 24th of February 2022[25].

visited 30.07.2023

[22] They defined them as: updates with the highest increase in the percentage of negative reviews relative to the prior updates of the app

[23] More information available from World Health Organisation (WHO): https://www.who.int/emergencies/diseases/novel-coronavirus-2019 (last visited 02.08.2023)

[24] Information on the the exact dates of lockdown by country with references to the source information has been gathered for a Kaggle competition and can be accessed here: https://www.kaggle.com/datasets/jcyzag/covid19-lockdown-dates-by-country (last visited 02.08.2020)

[25] More information available from the United Nations (UN): https://press.un.org/en/2022/sc14803.doc.htm (last visited 02.08.2023)

More specifically, we analyse the reviews from four popular apps by Chinese (*UC Browser*, *SHAREit*, *Likee* and *SHEIN*) and four by Russian developers (*Kaspersky Antivirus & VPN*, *VK (VKontakte)*, *Yandex Browser* and *OK (Odnoklassniki)*).

From Figure 7 we can notice a distinct peak in the second quarter of 2020 for toxic, toxic-critical and critical reviews for Chinese apps, which falls right to the time of first complete and partial lockdowns in several countries across the world, and the then US president Trump using the term **"Chinese virus"** that had several experts fear for xenophobia [RJS20]. There is no such distinct peak in toxic and toxic-critical reviews for Russian apps in the second quarter of 2022, the first months from the beginning of the invasion of Ukraine.

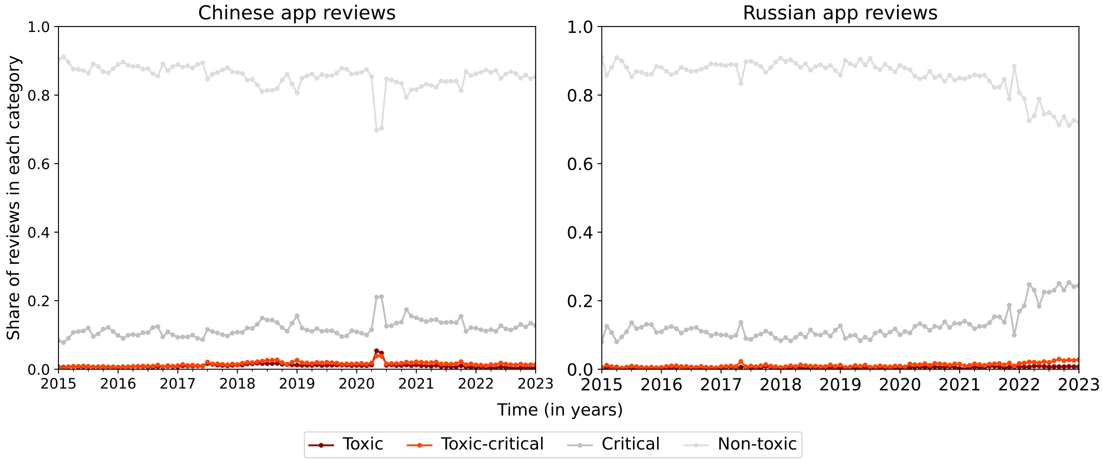


Figure 7. Share of toxicity categories from 2015 to 2023 for Chinese and Russian apps

4.4.1 Chinese apps and COVID-19 pandemic

Based on Figure 7 we might want to reason that Chinese apps received more toxic and/or critical reviews during and because of the pandemic. However, another perspective emerges from looking at the total number of reviews (Figure 8) instead of the share in monthly reviews. We can see that all four apps experienced a significant drop in the number of reviews, presumably at least partially because all four apps were banned in India, one of the largest markets for Chinese apps [Aha20], in June 2020 by the Indian government for national security concerns [Hem22].

In March and April 2020, there were a total of 207,711 reviews for the four Chinese apps. Out of these, 151 reviews (0.07%) contained at least one of COVID-19-related search terms[26] (Table 7). Out of those only 31 were toxic-critical (and one toxic). There

[26] The search terms included: corona, covid and chinese virus. All search terms were transformed to regular expression patterns for case-insensitive search

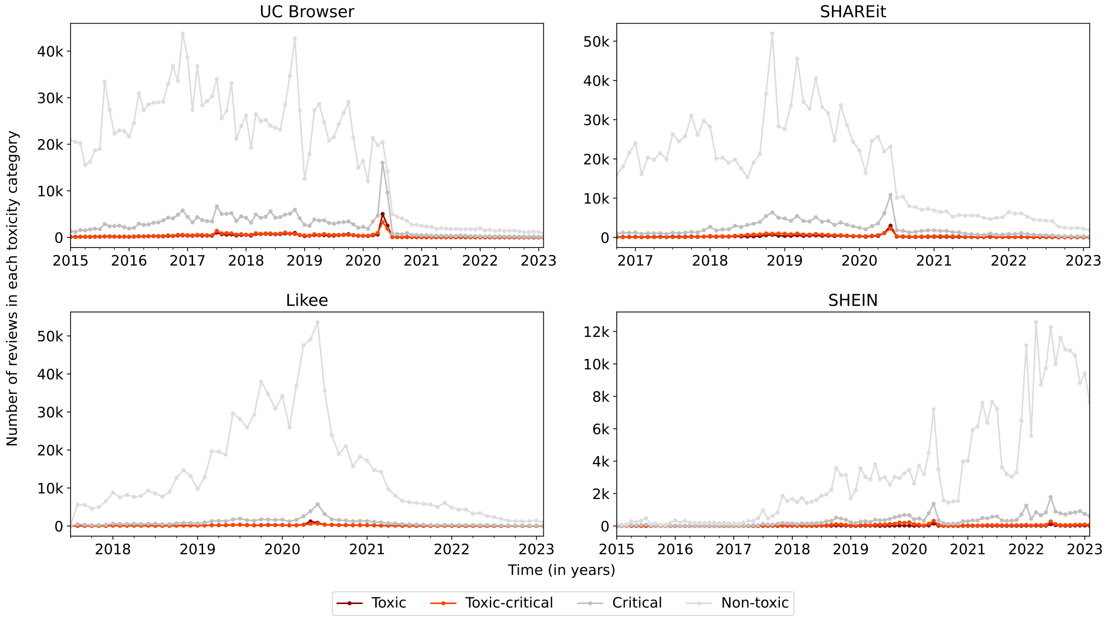


Figure 8. Number of reviews for each toxicity category from 2015 to 2023 for four Chinese apps

were also nearly the same amount of reviews (N=25) that rose concerns about data safety and privacy[27] for the same two-month period.

It has been established that anti-Asian attitudes were activated in the early stages of the pandemic [GHL20] and Asian Americans reported a surge in racially motivated hate crimes involving physical violence and harassment in the Spring months of 2020 [RB22]. However, we found that only approximately 30 reviews (0.015% of more than 208k reviews for all four apps in March and April 2020) blamed the Chinese for the pandemic and called to uninstall and boycott their apps. We can therefore deduce that the overall negative effect of the pandemic on the review space for Chinese apps is rather limited. This could be either due to Google's moderation efforts or users choosing alternative platforms, such as Twitter, to voice their sentiments [HZS$^+$21].

4.4.2 Russian apps and invasion of Ukraine

While we could notice a peak in the share and number of toxic and toxic-critical reviews for Chinese apps in the second quarter of 2020, there is no such distinct peak for Russian apps in the spring of 2022, the first months following the beginning of the invasion. The four Russian apps under review, had a total of 3103 reviews in March and April 2022, 50

[27] The search terms included: safe, security and privacy. All search terms were transformed to regular expression patterns for case-insensitive search.

Toxicity category	N reviews	Example(s)
Non-toxic	50	"Super bro. U can see videos in the app bro. Be safe from corona" "5* for spreading awareness of Corona. Edit replying: Yes, I do like the feature, it is very useful. Much appreciated! :) Also I like the app in general, it sends files rapidly."
Critical	69	"Very nice app but 1 problem is that if you want to shoot English song it doesn't have any of English song ...and guys be safe #corona virus" "Bad not happy no communication regarding my order paid on the 26th February till now my clothes and boots not yet delivered. In regardless of the Corona virus I still need to know how far is my delivery is it safe. Very disappointed."
Toxic-critical	31	"The quality of stuff I ordered was pathetic. Also, They evade taxes of the country they send stuff too. Thanks but no thanks, the Chinese Corona virus is enough. Dont need more trash from them." "I am going to uninstall "shareit" because it was made by China... It is my first step to ban Chinese products... They are responsible for the epidemic COVID-19 Me and all the people of this world just hate China. We are going to ban you M...F..."
Toxic	1	"I hate China and Chinese #CORONA"

Table 7. Example reviews from four toxicity categories for four Chinese apps between March and April 2020

of which (1.6%) contained search terms related to the invasion[28] (Table 8). Out of the 50 reviews three were toxic-critical and one was toxic. Majority (N=44) expressed that they would be uninstalling or boycotting the app because of the developer's ties to the leaders of Russia and invasion of Ukraine, but also did so without explicit hate speech, vulgarity, cursing or calumniation.

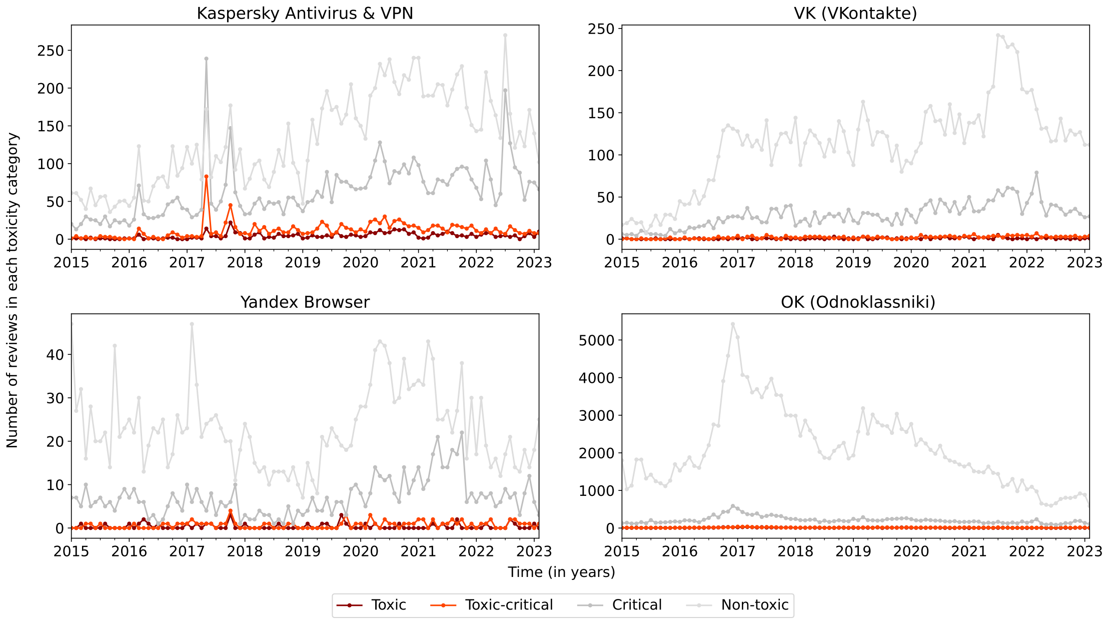


Figure 9. Share of toxicity categories from 2015 to 2023 for four Russian apps

These findings indicate that the reviews sections of Russian apps were affected by the invasion of Ukraine, but not necessarily in terms of toxic content. However, it is worthwhile to note that only reviews in the English language were analysed in this thesis, as a different kind of perspective could emerge from analysing reviews in Slavic languages.

[28] The search terms included: ukraine, war, invasion and slava. All search terms were transformed to regular expression patterns for case-insensitive search.

Toxicity category	N reviews	Example(s)
Non-toxic	4	"Peace for Ukraine" "can see Ukraine Russia war in different view. nil cencorship"
Critical	42	"Eugene Kaspersky's statement about the invasion of Ukraine is totally unacceptable. I no longer want to use any of your products anymore." "Uninstalling the app due to the Russian invasion of the Ukrain"
Toxic-critical	3	"Excellent antivirus.! (with the Ukraine war I'm rethinking my future with Kaspersky). At this moment I keep it on my mobile and iMac but that may change. Retards." "Kaspersky would be good if the owner of the company would not be the best friend of head of Russian Intelligence service. But in war times, they can make you a traitor by stealing info from your phone or even using it to attack network from the inside. Please delete it before squad team would storm your door just because last night your phone attacked a nearby nuclear plant ..."
Toxic	1	"Bad app. Putin War Criminal"

Table 8. Example reviews from four toxicity categories for four Russian apps between March and April 2022

5 Conclusion

This thesis delves into understanding the nature of toxicity within app store review sections, examining "what, where, and why" toxic content arises. By analysing nearly 60 million reviews spanning almost a decade (from January 2014 to January 2023) on the Google Play Store from over 5800 apps, we shed light on key aspects of toxicity in this digital landscape.

What?

Toxicity, encompassing hate speech, harassment, and cyberbullying, manifests differently across various online platforms [MCK$^+$22]. For app reviews, we identify two types of toxic comments: explicit **toxic** reviews with insults, vulgarity, or hate speech, and **toxic-critical** reviews blending toxicity with criticism about app features. We emphasise that, beyond *target*, *severity*, and *type* of toxicity, the content itself matters. We advocate swift removal of **toxic** reviews violating content policies, but **toxic-critical** reviews, containing valuable user-reported concerns, may require a distinct approach. Implementing mechanisms allowing developers to access policy-violating comments privately could be considered.

Where?

Toxic content is present in about 3.5% of reviews on average, surpassing rates on prominent platforms like GitHub and Stack Overflow [CSC21b]. However, toxicity's prevalence varies across app genres and time periods. While over 7% of reviews in the Dating category are toxic, fewer than 2% in Books & Reference and Weather are. Over nearly a decade, the share of **toxic** reviews remains around 1%, while the share of **toxic-critical** reviews gradually grows, increasing from below 2% to slightly over 3% in nine years.

Why?

This study only scratches the surface of causality in Google Play Store toxicity. Two examples illustrate that changes in UX/UI or policies can lead to significant spikes in the proportion of toxic-critical reviews, increasing to almost 20% of all reviews, which is nearly a tenfold rise compared to the overall average of 2.5%. Conversely, external events appear to have a minimal impact on the prevalence of toxicity in Google Play Store reviews. Investigating the potential effects of the COVID-19 pandemic and Russian invasion of Ukraine on toxicity levels for Chinese and Russian applications, we observe that only a small fraction of reviews during the initial global lockdowns (March and April 2020) were pandemic-related, and even fewer contained toxic content (0.07% and 0.015% respectively). Similarly, for the invasion, approximately 1.6% of reviews in March and April 2022 were relevant, with only 0.12% containing toxic content.

This thesis yields a multifold contribution: training the reviewBERT model, compiling a large-scale dataset for extensive research, and offering insights into toxicity's prevalence, trends, and influencing factors. Further research could delve into causal dynamics and quantification. For instance, probing why certain apps or genres foster toxicity can uncover reasons like controversial content, high expectations, or negative user experiences. Additional case studies may unveil nuances in user feedback dynamics, particularly around new releases and quantify frustration-triggered toxicity.

References

[Aha20] Abhijit Ahaskar. India a major market for chinese apps on banned list. *Livemint*, July 2020.

[Ash22] Ashley. Why Did My Google Reviews Get Removed?, October 24 2022.

[Ayd16] Gökhan Aydın. Attitudes towards digital advertisements: Testing differences between social media ads and mobile ads. *International Journal of Research*, 1, 2016.

[BMST$^+$23] Andrea Bonetti, Marcelino Martínez-Sober, Julio C Torres, Jose M Vega, Sebastien Pellerin, and Joan Vila-Francés. Comparison between machine learning and deep learning approaches for the detection of toxic comments on social networks. *Applied Sciences*, 13(10):6038, 2023.

[Çal23] Levent Çallı. Exploring mobile banking adoption and service quality features through user-generated content: The application of a topic modeling approach to google play store reviews. *International Journal of Bank Marketing*, 41(2):428–454, 2023.

[CDNML15] Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. Antisocial behavior in online discussion communities. *Proceedings of the International AAAI Conference on Web and Social Media*, 9, 04 2015.

[CHB21] Asma Chader, Leila Hamdad, and Abdesselam Belkhiri. Sentiment analysis in google play store: Algerian reviews case. In *Modelling and Implementation of Complex Systems: Proceedings of the 6th International Symposium, MISC 2020, Batna, Algeria, October 24-26, 2020 6*, pages 107–121. Springer, 2021.

[CjSR$^+$22] Marta R Costa-jussà, Eric Smith, Christophe Ropers, Daniel Licht, Javier Ferrando, and Carlos Escolano. Toxicity in multilingual machine translation at scale. *arXiv preprint arXiv:2210.03070*, 2022.

[Coh60] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[Coo17] James Cook. The reclusive ceo of dating app badoo on his app's redesign, bumble, and why he won't ipo any time soon. *Business Insider*, April 2017.

[CSC21a] Jithin Cheriyan, Bastin Tony Roy Savarimuthu, and Stephen Cranefield. Norm violation in online communities–a study of stack overflow comments. In *Coordination, Organizations, Institutions, Norms, and Ethics for*

Governance of Multi-Agent Systems XIII: International Workshops COIN 2017 and COINE 2020, Sao Paulo, Brazil, May 8-9, 2017 and Virtual Event, May 9, 2020, Revised Selected Papers, pages 20–34. Springer, 2021.

[CSC21b] Jithin Cheriyan, Bastin Tony Roy Savarimuthu, and Stephen Cranefield. Towards offensive language detection and reduction in four software engineering communities. In *Proceedings of the 25th International Conference on Evaluation and Assessment in Software Engineering*, EASE '21, page 254–259, New York, NY, USA, 2021. Association for Computing Machinery.

[DCLT18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[Dou21] Evelyn Douek. Governing online speech: From "posts-as-trumps" to proportionality and probability. *Columbia Law Review*, 121(3):759–834, 04 2021.

[Fei18] Fei Ye and Kazushi Nagayama. In reviews we trust — Making Google Play ratings and reviews more trustworthy, December 17 2018.

[FLL$^+$13] Bin Fu, Jialiu Lin, Lei Li, Christos Faloutsos, Jason Hong, and Norman Sadeh. Why people hate your app: Making sense of user feedback in a mobile app store. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1276–1284, 2013.

[GHL20] Angela R Gover, Shannon B Harper, and Lynn Langton. Anti-asian hate crime during the covid-19 pandemic: Exploring the reproduction of inequality. *American journal of criminal justice*, 45:647–667, 2020.

[GLQ$^+$22] Cuiyun Gao, Yaoxian Li, Shuhan Qi, Yang Liu, Xuan Wang, Zibin Zheng, and Qing Liao. Listening to users' voice: Automatic summarization of helpful app reviews. *IEEE Transactions on Reliability*, pages 1–13, 2022.

[GNH17] Jiaping Gui, Meiyappan Nagappan, and William GJ Halfond. What aspects of mobile ads do users care about? an empirical study of mobile in-app ad reviews. *arXiv preprint arXiv:1702.07681*, 2017.

[Goo17] Google. Google Perspective API. Website, 2017.

[GSH16] Joshua Guberman, Carol Schmitz, and Libby Hemphill. Quantifying toxicity and verbal violence on twitter. In *Proceedings of the 19th ACM*

Conference on Computer Supported Cooperative Work and Social Computing Companion, pages 277–280, 2016.

[HBH18] Safwat Hassan, Cor-Paul Bezemer, and Ahmed E Hassan. Studying bad updates of top free-to-download apps in the google play store. *IEEE Transactions on Software Engineering*, 46(7):773–793, 2018.

[Hem22] Asha Hemrajani. CO22102 | The Indian Government Ban on Chinese Apps and the Singapore Connection. *RSIS Commentary*, October 2022.

[HGP$^+$22] Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association of Computational Linguistics*, 2022.

[HKZP17] Hossein Hosseini, Sreeram Kannan, Baosen Zhang, and Radha Poovendran. Deceiving google's perspective api built for detecting toxic comments. 02 2017.

[HZS$^+$21] Bing He, Caleb Ziems, Sandeep Soni, Naren Ramakrishnan, Diyi Yang, and Srijan Kumar. Racism is a virus: Anti-asian hate and counterspeech in social media during the covid-19 crisis. In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 90–94, 2021.

[Isl14] Mir Riyanul Islam. Numeric rating of apps on google play store by sentiment analysis on user reviews. In *2014 International Conference on Electrical Engineering and Information & Communication Technology*, pages 1–4. IEEE, 2014.

[JBC$^+$18] Edwin Jain, Stephan Brown, Jeffery Chen, Erin Neaton, Mohammad Baidas, Ziqian Dong, Huanying Gu, and Nabi Sertac Artan. Adversarial text generation for google's perspective api. In *2018 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 1136–1141, 2018.

[JRSM20] Prerna Juneja, Deepika Rama Subramanian, and Tanushree Mitra. Through the looking glass: Study of transparency in reddit's moderation practices. 4(GROUP), jan 2020.

[KBA19] Keita Kurita, Anna Belova, and Antonios Anastasopoulos. Towards robust toxic content classification. *arXiv preprint arXiv:1912.06872*, 2019.

[KDB+23] Arvind S Kapse, Anamay Dubey, Harshvardhan Bisen, Kapil Kumar, and Md Tamheed. Multilingual toxic comment classifier. In *2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS)*, pages 1223–1228. IEEE, 2023.

[KSA22] Lisa Kaati, Amendra Shrestha, and Nazar Akrami. A machine learning approach to identify toxic language in the online space. In *2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 396–402, 2022.

[Lia19] Shannon Liao. After the porn ban, tumblr users have ditched the platform as promised. *The Verge*, 3 2019.

[LOG+19] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[LTT+22] Alyssa Lees, Vinh Q Tran, Yi Tay, Jeffrey Sorensen, Jai Gupta, Donald Metzler, and Lucy Vasserman. A new generation of perspective api: Efficient multilingual character-level transformers. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3197–3207, 2022.

[MCK+22] Courtney Miller, Sophie Cohen, Daniel Klug, Bogdan Vasilescu, and Christian KaUstner. "did you miss my comment or what?": Understanding toxicity in open source discussions. In *Proceedings of the 44th International Conference on Software Engineering*, ICSE ’22, page 710–722, New York, NY, USA, 2022. Association for Computing Machinery.

[MJB+15] J Nathan Matias, Amy Johnson, Whitney Erin Boesel, Brian Keegan, Jaclyn Friedman, and Charlie DeTar. Reporting, reviewing, and responding to harassment on twitter. *arXiv preprint arXiv:1505.03359*, 2015.

[MJM00] Ron C Mittelhammer, George G Judge, and Douglas J Miller. *Econometric foundations pack with CD-ROM*. Cambridge University Press, 2000.

[MS22] Mayukh Mukhopadhyay and Sangeeta Sahney. Effect of toxic review content on overall product sentiment. *arXiv preprint arXiv:2201.02857*, 2022.

[NL22] Ehsan Noei and Kelly Lyons. A study of gender in user reviews on the google play store. *Empirical Software Engineering*, 27(2):34, 2022.

[Noe18] David Noever. Machine learning suites for online toxicity detection. *arXiv preprint arXiv:1810.01869*, 2018.

[Ope22] OpenAI. ChatGPT, 2022. (Last visited on 05.08.2023).

[Ous21] Nedjma Djouhra Ousidhoum. *On the Importance and Challenges of the Experimental Design of Multilingual Toxic Content Detection*. Hong Kong University of Science and Technology (Hong Kong), 2021.

[Pea00] Karl Pearson. X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175, 1900.

[Pet22] Peter M. Stahl. Lingua Language Detector 1.3.2. PyPI Repository, 2022.

[PM13] Dennis Pagano and Walid Maalej. User feedback in the appstore: An empirical study. In *2013 21st IEEE International Requirements Engineering Conference (RE)*, pages 125–134, 2013.

[PMA17] John Pavlopoulos, Prodromos Malakasiotis, and Ion Androutsopoulos. Deeper attention to abusive user content moderation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1125–1135, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[PPS21] Jainil Viren Parikh, YR Prajwal, and Rajashree Shettar. Toxic text classification for globalisation of software products. 2021.

[PRW$^+$20] Dany Pratmanto, Rousyati Rousyati, Fanny Fatma Wati, Andrian Eko Widodo, Suleman Suleman, and Ragil Wijianto. App review sentiment analysis shopee application in google play store using naive bayes algorithm. In *Journal of Physics: Conference Series*, volume 1641, page 012043. IOP Publishing, 2020.

[PSD$^+$20] John Pavlopoulos, Jeffrey Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos. Toxicity detection: Does context really matter? *arXiv preprint arXiv:2006.00998*, 2020.

[PTDA19] John Pavlopoulos, Nithum Thain, Lucas Dixon, and Ion Androutsopoulos. Convai at semeval-2019 task 6: Offensive language identification and categorization with perspective and bert. pages 571–576, 01 2019.

[RB22] Tyler T Reny and Matt A Barreto. Xenophobia in the time of pandemic: Othering, anti-asian attitudes, and covid-19. *Politics, Groups, and Identities*, 10(2):209–232, 2022.

[RG18] Nestor Rodriguez and Sergio Rojas Galeano. Shielding google's language toxicity model against adversarial attacks. *ArXiv*, abs/1801.01828, 2018.

[RJ$^+$12] G Rupert Jr et al. Simultaneous statistical inference. 2012.

[RJS20] Katie Rogers, Lara Jakes, and Ana Swanson. Trump defends using 'chinese virus' label, ignoring growing criticism. *New York Times*, March 2020.

[SDCW19] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[SFE$^+$19] Md Muhtasim Jawad Soumik, Syed Salvi Md Farhavi, Farzana Eva, Tonmoy Sinha, and Mohammad Shafiul Alam. Employing machine learning techniques on sentiment analysis of google play store bangla reviews. In *2019 22nd International Conference on Computer and Information Technology (ICCIT)*, pages 1–5. IEEE, 2019.

[SGK$^+$23] Punyajoy Saha, Kiran Garimella, Narla Komal Kalyan, Saurabh Kumar Pandey, Pauras Mangesh Meher, Binny Mathew, and Animesh Mukherjee. On the rise of fear speech in online social media. *Proceedings of the National Academy of Sciences*, 120(11):e2212270120, 2023.

[SHX21] Guizhe Song, Degen Huang, and Zhifeng Xiao. A study of multilingual toxic text detection approaches under imbalanced sample distribution. *Information*, 12(5):205, 2021.

[SÖO$^+$20] İbrahim Onur Siğirci, Hakan Özgür, Abdullah Oluk, Harun Uz, Emrah Çetiner, Hande Uzun Oktay, and Kaan Erdemir. Sentiment analysis of turkish reviews on google play store. In *2020 5th International Conference on Computer Science and Engineering (UBMK)*, pages 314–315. IEEE, 2020.

[VHMN12] Rajesh Vasa, Leonard Hoon, Kon Mouzakis, and Akihiro Noguchi. A preliminary analysis of mobile app user reviews. In *Proceedings of the 24th Australian Computer-Human Interaction Conference*, OzCHI '12, page 241–244, New York, NY, USA, 2012. Association for Computing Machinery.

[VKV20] Swathi Venkatakrishnan, Abhishek Kaushik, and Jitendra Kumar Verma. Sentiment analysis on google play store data using deep learning. *Applications of Machine Learning*, pages 15–30, 2020.

[WDWW17] Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84, Vancouver, BC, Canada, August 2017. Association for Computational Linguistics.

[Wik15] Wikipedia. Harassment Survey 2015 - Results Report. Online report, 2015.

[Wil18] Williams Pelegrin. Google removed 77 percent of Play Store reviews for Game Dev Tycoon, February 14 2018.

[WK15] Elizabeth Whittaker and Robin M Kowalski. Cyberbullying via social media. *Journal of school violence*, 14(1):11–29, 2015.

[WWL$^+$22] Liu Wang, Haoyu Wang, Xiapu Luo, Tao Zhang, Shangguang Wang, and Xuanzhe Liu. Demystifying “removed reviews” in iOS app store. In *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pages 1489–1499, 2022.

[XZL$^+$20] Yan Xia, Haiyi Zhu, Tun Lu, Peng Zhang, and Ning Gu. Exploring antecedents and consequences of toxicity in online discussions: A case study on reddit. *Proceedings of the ACM on Human-computer Interaction*, 4(CSCW2):1–23, 2020.

[ZBQ14] Marc Ziegele, Timo Breiner, and Oliver Quiring. What creates interactivity in online news discussions? an exploratory analysis of discussion factors in user comments on news items. *Journal of Communication*, 64(6):1111–1138, 2014.

Appendix

I. GitLab repository

All data and code files are available in a private repository at: `https://gitlab.ut.ee/triin.pohla/toxicity_in_gps` (access granted upon request)

II. Licence

Non-exclusive licence to reproduce thesis and make thesis public

I, **Triin Pohla**,

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to

   reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright,

   Toxicity in Google Play Store Reviews: What, Where and Why?,

   supervised by Vigneshwaran Shankaran and Rajesh Sharma.

2. I grant the University of Tartu a permit to make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 3.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in p. 1 and 2.

4. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Triin Pohla
11/08/2023