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

Dariya Nagashibayeva

Understanding Gender Related Discussions in Android Mobile Applications Through Reddit

Master's Thesis (30 ECTS)

Supervisor(s): Tahira Iqbal, MSc
Kuldar Taveter, PhD

Tartu 2024

Understanding Gender Related Discussions in Android Mobile Applications Through Reddit

Abstract:
A data-driven approach to the development of software applications aims to enhance the user experience and quality of life for broader groups of users. Taking into consideration the problems and needs of diverse communities is essential for creating safe and inclusive software. For this reason, understanding the discussions of such topics as gender in the software communities and identifying the inclusivity violations in software products through data analysis is crucial for improving the software design. The purpose of this thesis is to investigate the degree of contentment with popular Android mobile applications among the users of the social networking platform Reddit in terms of gender inclusiveness, explore the possibility of automated detection of gender discussions based on the preprocessed data, and evaluate the findings for suitability for improving software requirements. The research work presented in this thesis employed the quantitative and qualitative analysis of source data that included the data collection and manual annotation of keywords and Reddit posts. Application of the aforementioned methods resulted in a new textual dataset on the topic of gender ready to use for data analysis. In addition, the thesis comprises experiments of using the dataset for the automated classification of Reddit posts and suggests which machine learning models, including state-of-the-art deep learning models, can be used for the detection of gender inclusiveness violations in software applications. The thesis also includes recommendations on how to deal with the limitations of such an automated classification approach in the future.

Keywords:
Reddit data analysis, Gender discussions, Mobile applications, Machine learning

CERCS: P175 Informatics, systems theory

Androidi mobiilirakenduste soolise puutumuse arutelude analüüs sotsiaalvõrgustikus Reddit

Lühikokkuvõte:
Andmepõhise lähenemise eesmärgiks tarkvararakenduste arendamisele on laiemate kasutajarühmade kasutajakogemuse ja elukvaliteedi parandamine. Mitmekesiste kogukondade probleemide ja vajaduste arvestamine on turvalise ja kaasava tarkvara loomiseks olemuslik. Seetõttu on tarkvaratoodete disaini parendamiseks võtmetähtsusega vastavate teemade avamine andmeanalüüsi abil nagu näiteks sooline puutumus tarkvarakogukondades ja võrdõiguslikkuse rikkumiste avastamine tarkvaratoodetes. Käesoleva magistritöö eesmärgiks on uurida sotsiaalvõrgustiku Reddit kasutajate rahulolu määra soolise võrdõiguslikkusega populaarsetes Androidi mobiilirakendustes, uurida võimalusi soolise puutumuse arutelude automaatseks avastamiseks eeltöödeldud andmete alusel ning hinnata tulemusi nende sobivuse osas tarkvarale esitatavate nõuete parendamiseks. Käesolevas magistritöös esitatud uurimistöö rakendas lähteandmete kvantitatiivset ja kvalitatatiivet analüüsi, mis sisaldas lähteandmete kogumist ning võtmesõnade ja Redditi postituste käsitsi annoteerimist. Eelpoolmainitud meetodite rakendamise üheks tulemuseks oli andmeanalüüsiks kasutusvalmis soolise puutumuse tekstiandmete kogu. Lisaks sellele hõlmab magistritöö selle tekstiandmete kogu kasutamist Redditi postituste automaatseks klassifitseerimiseks ja annab soovitusi selles osas, milliseid masinõppe mudeleid, kaasa arvatud süvaõppe mudelid, saab kasutada soolise võrdõiguslikkuse rikkumiste avastamiseks tarkvararakendustes. Magistritöö sisaldab ka soovitusi selle kohta, kuidas tulevikus üle saada niisuguse automaatse klassifitseerimise lähenemisviisi piirangutest.

Võtmesõnad:
Redditi andmeanalüüs, Soolise puutumuse diskussioonid, Mobiilirakendused, Masinõpe

CERCS: P175 Informaatika, süsteemiteooria

Contents

1 Introduction

Mobile applications are created with the main purpose of increasing the quality of life and convenience of everyday users. Many software developers became more attentive to the user feedback left on online platforms and communities that have been highlighting the relevant problems, and consequently improved the standards of their products. One such problem is addressing gender inclusiveness when creating software applications. It is of critical importance to ensure gender inclusivity in mobile applications in order to guarantee accessibility, provide a positive user experience, comply with legal requirements, and gain a competitive advantage in the market.

Among the related previous research, several unifying trends in research can be identified. In theoretical terms, the most popular trend is the study of violations of human values in mobile and desktop applications, while not much attention has been paid to the discussion of gender in particular. In terms of methodologies, the most common source of user opinions is the Android Play Store [43], [22] and social platforms such as Twitter [29], [14]. The use of Reddit as a data source for user opinion analysis is less common among researchers. Moreover, it allows us to evaluate the gender issues from a user perspective. For these reasons, we decided to dedicate our project to investigating the topic of gender in mobile apps using data taken from Reddit. Our thesis aims to provide new insights into the gender discourse in online communities such as Reddit and to contribute to the automated classification of gender issues in software. Additionally, the results of our work can be utilized for requirements engineering. Based on these objectives, the research questions were formulated:

1. **RQ1:** Does Reddit data provide meaningful information about gender-related discussions in mobile applications?

   We aim to mine Reddit data from mobile application-related communities that could potentially give us relevant information. We plan to conduct a detailed quantitative and qualitative analysis of such data to better understand user feedback with a focus on gender discussions. Quantitative analysis includes exploring the popularity of mobile applications, the availability and relevance of corresponding Reddit communities, and finally mining the data. As for qualitative analysis, we aim to analyze the content of the data and manually annotate it to prepare a gender dataset.

2. **RQ2:** Can we further categorize the gender-related discussions in Reddit posts?

   Our objective in answering this question is to carefully examine data that was labeled as related to gender discussions to understand if it can be divided into different categories.

3. **RQ3:** How do different ML models used for the automated classification of Reddit data for gender discussions in mobile applications compare with each other?

As a final step of our study, we aim to apply and compare different machine learning and deep learning algorithms to understand what model and configurations are best suited for the automated classification of gender-related data.

The rest of the thesis is organized in the following order. Chapter 2 provides a thorough research of the existing works in the fields of mining data for software requirements and gender issues in software engineering. It gives the overview of the work by Shahin et. al [43] on a similar topic and defines the direction of our research. Chapter 3 is dedicated to the methodology and research design, where each process step is exhaustively explained. The explorative data analysis and statistics of Reddit metadata support the methodology. In Chapter 4, the automated classification experiments were conducted, and the best-performing ML and DL classifiers were defined. Chapter 5 continues the qualitative data analysis and proposes the categorization of the obtained gender-related discussions. Chapter 6 covers the discussions of the implications of this thesis, where the research questions are answered, and threads to validity are mentioned together with the recommendations for future work.

2 Background and related works

This section covers the review of the existing research and projects, related to gender discussions in mobile applications and using Reddit data as a source for requirements engineering. Subsection 2.1. gives a critical overview of methodology and patterns used in similar research and points to potential gaps in the research on which this thesis is built. Additionally, the chapter highlights the challenges of extracting data from Reddit due to policy regulations.

2.1 Related work

To have a broad understanding of the problem and current research state, we conducted a thorough literature analysis. We started our research by investigating the scope of the problem with gender in software engineering. Once we gained an understanding of the problem and formed a motivation for our research, we started exploring the existing methodological approaches on how to collect necessary data, perform both quantitative and qualitative data analysis, and use the obtained data for potential software requirements. Hence, the Related work subsection is grouped in two subsections - 2.2.1 Gender issues in software engineering and 2.2.2 Mining data for software requirements.

2.1.1 Mining data for software requirements

In recent years, there has been a growing interest among researchers in using user feedback and app reviews as a source for requirements engineering. One of the most popular sources is the Google Play Store. Obie et al. have done extensive work to identify human value violations in app reviews [30]. They collected 22,607 reviews of the 12 most popular mobile apps, selecting, where possible, the most recent 2,000 reviews for each app. The authors engaged two analysts to create a human values dictionary consisting of 50 semantically and contextually related value terms according to the Schwartz model [42]. They then manually annotated the labels in a new Truthset dataset of 709 reviews. Pilot labelling included 143 reviews for each analyst, which they then cross-checked with each other. The main study was conducted on a Truthset with 709 reviews, which was shared between analysts. Data preprocessing included the following: using the "Autocorrect" spell-checking library to find erroneous words, removing stop words, and stemming using the NLTK library. To extract application characteristics, the authors used a rule-based method called SAFE [18]. For sentiment analysis, they used a rule-based model called VADER [16], which is a low-resource model tuned specifically for detecting sentiments in user reviews. The results of this research proved that user reviews are a valuable source for the detection of human value violations in software applications.

Similarly, in [22], Conghui et. al. collected 46 applications from the Google Play Store covering 10 categories from the Schwartz's model. The manual annotation of value violations in apps was performed by one author, and validated by another. The authors discussed the differences together and finalized annotations that formed a truth set for the evaluation of algorithms. However, the authors did not apply machine learning in this research. Instead, they applied Android Application Package (APK) decomposition and Java Abstract Syntax Tree (AST) analysis to prepare data, and designed algorithms in Java to detect human values violations.

The methodology of the paper Collins et. al. [24] introduced a keyword filtering step in the preparation of data. The entire process included data collection from 24 mobile applications, filtering via a semi-automated keyword-based tool, manual labeling of 8 human-centric categories, data pre-processing, and applying a machine learning model consisting of binary relevance transformation method and base classifier of Support Vector Machine, SVM. The authors also extended the classification to implementing a user interface (UI) for this tool. The manual labeling was performed by 5 authors, in which they detected 1315 samples out of 8965, where human-centric issues were discussed. The researchers showed how applying filtering based on keywords helped to exclude unrelated user reviews and increase the effectiveness of the study.

Another project, where researchers utilized keywords before labeling, was done by Fazzini et. al. [11]. For data collection, authors selected 57 official apps related to COVID-19 contact tracing from 30 countries. They downloaded 33029 reviews from the Apple Store and 189321 reviews from the Google Play Store. Before annotating app reviews, the authors selected reviews by utilizing a keyword-based filtering tool, which performed preprocessing steps (correcting misspelled words, removing stop words, and stemming) prior to that. The resulting dataset contained 72566 reviews (15500 from the Apple Store and 57066 from the Google Play Store). For qualitative analysis, three authors used inductive and axial coding and created a codebook with nine codes to categorize human aspects in app reviews. This process took them 2 person-months to complete. Then, the same authors analyzed and coded a sample of reviews for each app for a total of 2611 reviews. They used a negotiated agreement method to resolve the reliability of coding. As a result, they detected 716 human-aspect-related reviews and divided them into 9 categories: Age, Disability, Emotion, Gender, Language, Location, Privacy, Socioeconomic, and Miscellaneous. Gender was one of the least represented categories in COVID-19-related apps.

Shahin et. al. [43] were the first to investigate gender discussions in mobile application reviews. They collected 7 million reviews from 70 Android apps, approximately 100 hundred reviews for each app. The authors also utilized the keywords method to filter gender-related reviews, however, they applied it in two steps. First, they introduced the initial set of keywords, consisting of 45 words, which they used for filtering the whole dataset and got 3,222 gender-related reviews. Second, the researchers applied

the keyword extraction technique KeyBERT [13], which helped them to extract 396 gender-related keywords. They used the expanded list of keywords to filter the whole dataset the second time and obtained 12368 reviews. As in the previously mentioned papers, the authors also manually annotated the data to screen out false positive values. The pilot study was performed in two steps, where they randomly selected potentially gender-related reviews with the same or half the amount of gender-non-related reviews. The pilot study revealed the tendency of uni-gram keywords (keywords consisting of a single word) to give many false positive values, which led to excluding such keywords and performing the filtering step again. In the main study, the researchers repeated the same steps with approximately one-fourth of the filtered data and formed a balanced gender dataset.

Apart from Google Play Store, Apple App Store, and Twitter have long established themselves as valuable sources for mining user opinions for potential functional and non-functional requirements. Maalej et.al demonstrated the value of App Store data by achieving high-performance results in the automation of the classification of user reviews into bug reports, feature requests, reviews related to user experience or app ratings [23]. The authors of [29], [14], and [48] utilized Twitter data for mining user opinions on software products.

Stack Overflow has been an alternative source of data to gain feedback from software developers' perspectives. The authors of [49] and [36] researched Stack Overflow forums for the possibility of generating high-quality software documentation.

Another valuable source for extracting user feedback is the Reddit social platform. In the exploratory study [17], Iqbal et. al. presented Reddit as a potential source for requirements engineering. The authors evaluated the content of Reddit posts as well as their metadata and concluded that approximately 54 percent of the data is valuable for software development as feature requests or bug reports. They selected the top 10 mobile and desktop applications and searched for the dedicated subreddits for each application. In total, the authors extracted 25794 posts and 891,066 linked comments from 20 subreddits. The subreddit selection criteria were based on the number of members, frequency of posts, and whether a subreddit was official or was created specifically for an app. However, their further analysis showed that the high membership size did not necessarily guarantee a high volume of relevant posts.

Parsons et. al. utilized Reddit data to investigate privacy concerns and the influence of privacy regulations among software engineers [31] and extracted 437,317 threads from three subreddits dedicated to web development, Android, and iOS programming. The authors applied NLP models to assess the sentiment analysis and concluded that the majority of observations shared either a positive or neutral sentiment.

2.1.2 Gender issues in software engineering

Gender issues in software engineering have been topics of discussion for many years. They started early from the imbalanced representation of students of female and other gender minorities, as well as the majority of employees working in ICT (Information and Communication Technology) sector being men. Many researchers have been concerned by the underrepresentation of women and other marginalized groups as students and later employees in such fields as Science, Technology, Engineering, and Mathematics [8].

The whole new area of technology that has been affected by gender bias is artificial intelligence and software products that use machine learning algorithms. Cho et. al. gave an extensive review of how modern NLP (Natural Language Processing) algorithms support gender bias by assigning certain gender pronouns to stereotypical professions [7]. Shrestha et. al., however, claimed that nearly seamless bias in ML/AI algorithms can be even more harmful, especially when it comes to the recommendation or decision-making systems that have been trained on binary gender classification [44]. In recent years, there has been a growing interest in developing gender-neutral software engineering programs as a tool to achieve gender balance in the field. Kovaleva et. al. suggest focusing on gender-neutral software design rather than building it for the concerns of a certain gender. Authors claim that both male and female genders suffer from gender stereotypes related to particular software or products [20]. There are many researchers, who have also analyzed to understand the topic of diversity discussed in software engineering. Rodriguez-Perez et.al. prepared an overview of methods and tools that help detect diversity issues in software engineering [37]. Such tools included GenderMag (Gender Inclusiveness Magnifier) - a tool developed by Burnett et. al. [2], whose purpose is to detect and address potential gender biases in user interfaces, technology workflows and corporate practices. The GenderMag method is based on faceted personas that represent five facets of gender difference research. A generalization of GenderMag - InclusiveMag [26] was designed to provide systematic inclusiveness methods for specific facets of diversity so that this approach could be used along the different diversity dimensions.

3 Methodology

This chapter describes the stages of the process that were required to answer the research questions and achieve the declared goals. The chapter starts with data collection and quantitative analysis and continues with manual annotation and qualitative analysis. Additionally, it includes the data preparation for the experimental part of the research. The chapter concludes with the limitations of the chosen methodology and emphasizes its alignment with the research questions.

3.1 Research design

The research structure of this thesis adopts a combination of quantitative and qualitative methodologies to get an in-depth understanding of the selected data sources and their value in the exploration of gender discussions related to Android mobile software. The study is based entirely on primary data, putting an emphasis on the collection and analysis of raw data for the extraction of meaningful insights. By choosing to work with primary data, we sought to ensure the relevance of our results to the selected topic. The main steps of the research design include the selection of popular mobile applications, searching for the corresponding subreddits, web scrapping of the qualitative data, extraction of the themed keywords for data filtration, and the preparation of the gender dataset.

3.2 Data collection and quantitative analysis

3.2.1 Selection of applications

In selecting apps for our study, a conscious decision was made to focus on the Android platform, given its widespread adoption and diverse user base. Given its global reach, Android is a prominent platform on which to examine the nuances of gender discourse within the mobile app community. By choosing Android, we aim to cover a wide range of user experiences and interactions, including different demographics and preferences. The decision to include the top 50 apps across all categories was driven by a desire to ensure that our analysis was representative and inclusive. By including a variety of apps across different genres and functionalities in our analysis, we aim to gain a full understanding of gender dynamics within the broader context of the Android app debate. This approach ensures that our results reflect the diverse interests and interactions within the Android user community, contributing to a more nuanced exploration of gender themes related to different mobile app usage. Notably, the top 50 apps in the Google Play Store are, by default, targeted at the US market. The choice of the US as the focus of our study is due to the recognition that the US app market is one of the most universal ones, with developers often adapting their apps for a global audience. This strategic choice ensures that our analysis is based on a market that not only represents a wide range of app genres

but also reflects the international orientation of many app developers, contributing to a comprehensive examination of gender dynamics within the Android user community.

3.2.2 Reddit as a source of data

Reddit represents a valuable source of data for research scientists and software developers, as it stores a significant amount of content generated by users on various themes. This content provides insights into how users perceive certain topics. Reddit is comprised of thousands of distinct communities, or subreddits, which are dedicated to specific themes. These include subreddits that focus on gender-related discussions. A significant number of popular software products have their subreddits moderated by representatives of the software products or regular users. This feature enables data engineers to gather topic-specific data and utilize it as a source for future software requirements. A total of 40,000 characters may be used for text posts on Reddit. For post titles, a limit of 3,000 characters is in effect. Finally, for comments, a total of 100,000 characters may be used. Comments are organized in a tree-like structure. This means the original post is located at the top, with comments or replies linked below it. Both posts and comments can be upvoted or downvoted. This allows users to prioritize their relevance to the discussion thread. This structure results in a larger volume of user-generated content compared to other social media [1] and makes it a rich source of data for understanding gender discussions. In addition, Reddit users often use pseudonyms or remain anonymous, which encourages open and honest discussion of sensitive topics such as gender. This can lead to more explicit and diverse content. Besides the above-mentioned advantages, Reddit data can be easily accessed and collected through its API. However, there are some limitations to consider when using its data. One of them is the query limitation - the Reddit API allows to read no more than 1000 posts per subreddit, which may result in a relatively small dataset. Another limitation is associated with the fact that the discussions on Reddit can vary from well-thought-out and constructive ones to offensive and inappropriate ones. Separating valuable information from the noise can be challenging and time-consuming.

3.2.3 Data collection

To address the objective of this study, it was decided to select the top fifty Android mobile apps in general categories and for the US market. The list was requested on October 25, 2023, from the Apptopia [2], the platform that provides aggregated statistics for mobile applications. Since the list from Apptopia included apps from the "Games" category, it was supplemented with missing apps from the list of top 45 apps on Google Play [3]. For

[1] https://support.vistasocial.com/hc/en-us/articles/4409607590427-Character-limits-for-each-social-network

[2] https://apptopia.com/

[3] https://play.google.com/store/apps

each application in the list, the top three subreddits were manually searched. The reason for choosing the top three subreddits instead of one is justified by diverse perspectives, user segmentation and demographics, content variation, community dynamics, and overall inclusive representation of opinions. We believe that this approach can contribute to the depth and validity of our study findings. In this step, the size, status, and daily and weekly frequency of each subreddit's posts were also analyzed. The main criteria for selecting subreddits were: the subreddit had to be either official or related to the app; the subreddit had to be open so that posts and comments could be read from it; and it had to have at least 1,000 participants. If a subreddit did not meet at least one of these criteria, it was removed from the list. A subreddit was considered official unless otherwise specified. The 1,000-participant threshold was chosen because initial manual analysis showed that subreddits with fewer participants were either abandoned or had very few daily posts. Such subreddits did not provide sufficient data for the study. To scrape the posts and comments, different tools were considered, such as Python Reddit API Wrapper - PRAW [4], Pushshift API [5], and Pullpush API [6]. However, due to Reddit's new policy [7] both Pushshift API and Pullpush API were no longer being supported and Academic Torrents collected by Pullpush API contained data only up to December 2022. Thus, it was decided to use the PRAW library with its limitations.

3.2.4 Results of data collection

After applying the defined criteria to subreddits, 10 apps were excluded entirely due to either the absence of the relevant subreddits or all of their subreddits did not meet our requirements. The list of apps and corresponding subreddits are described in Table 1. As was stated in the previous chapter, because of the limitation of the Reddit API, it is possible to access up to 1000 posts maximum, either the most popular posts or the newest posts. Thus, for 122 subreddits only 97345 posts were fetched. Each instance of data included the attributes described in Table 2.

Table 1. Android applications and their subreddits.

Begin of Table

Application	Category	Subreddit	Members (in thousands)
Temu	Shopping	temu_ads	1.4
		TemuCodesUSA	1.4
		Polska	532
SHEIN	Shopping	Shein	10.2
		SHEIN_	18.9
TikTok	Social	TikTok	107
		Tiktok_hotgirls	6.5
		TikTok_Tits	416

[4]https://praw.readthedocs.io/en/stable/

[5]https://pushshift.io

[6]www.pullpush.io

[7]https://github.com/reddit-archive/reddit/wiki/API

Continuation of Table

Application	Category	Subreddit	Members (in thousands)
WhatsApp Messenger	Communication	whatsapp	46.3
Instagram	Social	Instagram	483
		InstagramMarketing	56.6
		morrasInstagram	32.5
Cash App	Finance	CashApp	70
		CashappBlessing	34.9
Telegram	Communication	Telegram	146
		TelegramGroups	33.8
		TelegramR4R	5.5
HBO Max	Entertainment	HBOMAX	79.7
		hbo	1000
		HBOmaxLegendary	5.1
Snapchat	Social	Snapchat	348
		SnapchatCodes	1.7
		SnapchatHelp	17.7
Messenger	Communication	facebookmessenger	3.8
		privacy	1300
		PrivacyGuides	63.5
CapCut	Video Players & Editors	CapCut	7.6
		editing	17.8
		VideoEditing	367
Facebook	Social	facebook	98.7
		FacebookAds	51.5
		insanepeoplefacebook	2100
Tubi	Entertainment	TubiTV	5.1
		badMovies	129
Walmart	Shopping	walmart	237
		WalmartEmployees	6.7
		WalmartSellers	3.6
Etsy	Shopping	Etsy	196
		EtsySellers	108
		EtsyCommunity	8.3
McDonald's	Food & Drink	McDonalds	56.8
		McDonaldsEmployees	44
		mcdonaldsfreakout	5.1
Microsoft Authenticator	Business	microsoft	309
		sysadmin	827
		windowsphone	54.8
Peacock TV	Entertainment	peacock	12.4
		television	1700
		cordcutters	534
NewsBreak	News & Magazines	techsupport	2000
		androidapps	307
		AndroidQuestions	122
Spotify	Music & Audio	spotify	1200
		SpotifyPlaylists	220
		truespotify	51.6
Netflix	Entertainment	netflix	1600
		NetflixByProxy	8.6
		NetflixAustralia	5.4
TextNow	Communication	TextNow	2.3
		NoContract	62
		ScamNumbers	36.9
Pluto TV	Entertainment	PlutoTV	1.3
		Pluto_TV	2.1
		htpc	62.6
Youtube TV	Entertainment	youtubetv	88.3
		youtube	1000
		AndroidTV	120

Continuation of Table

Application	Category	Subreddit	Members (in thousands)
Disney+	Entertainment	DisneyPlus	684
		DisneyPlusHotstar	26
		DisneyPlusVPN	5.1
ChatGPT	Productivity	ChatGPT	3400
		OpenAI	680
		ChatGPTPro	167
Zoom	Business	Zoom	30.3
		ZoomCourt	20.1
BP Tracker	Health & Fitness	bloodpressure	10.2
		gadgets	2180
		widgy	50.6
Pinterest	Lifestyle	Pinterest	29.1
		ReversePinterest	32.6
		csMajors	206
Talkie	Entertainment	moddedandroidapps	76.4
DoorDash	Food & Drink	doordash	332
		doordash_drivers	256
		DoorDashDrivers	4.5
FOX Sports	Sports	DirectvStream	6.2
		fireTV	60.3
		4kTV	76.3
Amazon Prime Video	Entertainment	AmazonPrimeVideo	319
		amazonprime	78.2
		technology	1510
Amazon Shopping	Shopping	amazon	214
		Frugal	3600
		ShoppingDealsOnline	6.9
PDF Pro	Tools	pdf	3.6
		pdfism	1.7
		software	200
Bing Chat	Tools	bing	70.1
		BingAi	1.4
Paypal	Finance	paypal	50.2
		PaypalDonations	2.2
		Free_Paypal_Money	3.8
Chime	Finance	chimefinancial	29.9
		fintech	31.3
		personalfinance	1830
Uber	Maps & Navigation	uber	45.2
		uberdrivers	363
		UberEATS	127
ESPN	Sports	ESPN	19.5
MONOPOLY GO!	Games	Monopoly_GO	46.7
		MonopolyGoTrading	46.3
		MonopolyGoCommunity	4
Roblox	Games	roblox	970
		RobloxAvatars	21.2
		RobloxHelp	10.1
Block Blast	Games	puzzles	243
		Tetris99	14.3
Gacha Life 2	Games	GachaLifeCringe	350
		GachaClub	53.5
Geometry Dash Lite	Games	geometrydash	151

Table 2. Reddit data attributes.

Attribute	Description	Sample value
Subreddit name	Name of a subreddit	microsoft
Post date	Date of submission in YYYY-MM-DD format	2023-10-04
Created UTC	Submission creation time in Unix format	1696437550
Post ID	Unique identifier of a submission	16zs0kz
Title	Title of a submission	Microsoft Start News Bots?
Post body	Submission text body	I think Microsoft needs to do something about Bots on MSN. One person if it's a topic on LGBTQ or Abortion rights will just flood the comments, and when I mean flood, I mean the person will post a comment per minute, for around 24 hours and it's the same response all the time, which makes me think the person is a bot.
Author	Author of a submission, Reddit username	CooperHChurch427
Score	The number of upvotes for a post	2
Upvote ratio	The percentage of upvotes from all votes on a submission	0.76
Number of comments	The number of comments on a submission	0
URL	The URL a submission links to, or the permalink if a selfpost.	https://www.reddit.com/r/microsoft/comments/16zs0kz/microsoft_start_news_bots/

3.3 Annotation

Once the data was successfully collected from selected subreddits, it was processed further for the creation of the gender dataset. This subchapter describes the process of finding the appropriate amount of keywords related to the topic of gender, data filtration with selected keywords, and manual annotation of data.

3.3.1 Keywords extraction and filtering

Keywords extraction. The first step after the data was collected, was to apply the first round of filtering to our 97345 posts using the keywords related to gender discussions from the paper [43]. The list contained 45 uni-gram or two-gram keywords: "sexism", "gender bias", "gender discrimination", "sexual discrimination", "male chauvinism", "antifeminism", "favouritism", "discrimination", "gender disparity", "gender difference", "gender inequality", "gender inequity", "gender imbalance", "gender", "feminism", "patriarchy", "misogyny", "misandry", "lgbtq", "egalitarianism", "masculine", "manly", "manful", "mannish", "manlike", "womanly", "womanlike", "womanish", "femalelike", "unfeminine", "paternal", "maternal", "lgb", "lgbt", "transgender", "gay", "lesbian", "bisexual", "homosexual", "genderfluid", "no-binary", "nonbinary", "non-binary", "in-

tersex", "agender". The filtering resulted in 102 posts potentially related to gender discussions.

The next step was to expand the gender-related keywords to explore the possibility of deriving more potentially gender-related data. There are several techniques for keyword generation, among which are RAKE (Rapid Automatic Keyword Extraction) [39], TF-IDF (Term Frequency - Inverse Document Frequency) [47], and YAKE (Yet Another Keyword Extractor) [3]. After exploring the above-mentioned methods, we chose KeyBERT model [13] for its simplicity and robustness against noise. KeyBERT is a technique that utilizes BERT embeddings to extract meaningful keywords that are most similar to the semantics of a document. As shown in Figure 1, KeyBERT starts by embedding a chunk of text into a fixed-size vector that represents the semantics of the document. It then extracts keywords from the document by using the count vectorizer. After the extraction, the model embeds each keyword similarly to embedding the document and produces the list of keyword embeddings. Finally, the model computes a similarity measure between the keyword embeddings and document embeddings and outputs the array of similarity scores, sorted in a decreasing order. The cosine similarity is calculated as follows:

$$\cos(\mathbf{A}, \mathbf{B}) = \frac{\mathbf{AB}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} \mathbf{A}_i \mathbf{B}_i}{\sqrt{\sum_{i=1}^{n} (\mathbf{A}_i)^2} \sqrt{\sum_{i=1}^{n} (\mathbf{B}_i)^2}} \tag{1}$$

where $\mathbf{A}_i$ and $\mathbf{B}_i$ are the ith components of vectors $\mathbf{A}$ and $\mathbf{B}$, respectively [34].

We chose the sentence model 'all-mpnet-base-v2' from Huggingface [8], which maps sentences and paragraphs into a dense vector space that captures the semantic meaning and can be utilized for sentences similarity tasks. Our goal was to generate as many different variations of similar keywords as possible while balancing the redundancy of the resulting keywords and key phrases. To achieve this, we chose a very low diversity hyperparameter (diversity = 0.1) and the Maximal Marginal Relevance (MMR) method for diversification (use_mmr = True) [13]. The MMR approach was introduced in the paper [5] as a method to reduce redundancy while keeping the query relevance of the resulting documents or phrases. MMR is calculated as follows:

$$MMR = \arg \max_{D_i \in R \setminus S} (\lambda \cdot (Sim(D_i, Q) - (1 - \lambda) \cdot \max_{D_j \in S} Sim(D_i, D_j))) \tag{2}$$

where MMR is the Maximal Marginal Relevance score, D_i represents a document collection, R is a ranked list of documents, S is the set of already selected documents in R, $(R \setminus S)$ is a set of documents yet to be selected in R, Q is the query, $Sim(D_i, Q)$ is the similarity score between document D_i and the query Q, and λ is a parameter controlling the trade-off between relevance and diversity [5].

S is the subset of documents in R already selected; R§is the set difference, i.e, the set of as yet unselected documents in R

[8]https://huggingface.co/sentence-transformers/all-mpnet-base-v2

Similarly to [43], we utilized existing gender-related datasets as documents to pass into KeyBERT model. The model then used those documents to extract semantically similar words, i.e. keywords. In total, two datasets were used: the EXIST [38] and the Sexism datasets [40]. EXIST dataset is a collection of posts from social platforms (Twitter and Gab) in Spanish and English languages organized into 11000 annotated texts (Sexist and Non-sexist). While the dataset has been further divided into different categories of sexism, we used only texts annotated as Sexist, and only in English language. The Sexism dataset has been derived from Twitter posts and psychological survey results, as well as from their synthetic variations. Analogous to the EXIST dataset, the Sexist dataset has been further divided into 3 categories, however, we ignored this categorization and used all the Sexist texts for our task.

Based on the datasets, we generated 1, 2, and 3-gram keywords and key phrases separately. A sample output of 3-gram key phrases is shown in Figure 2.

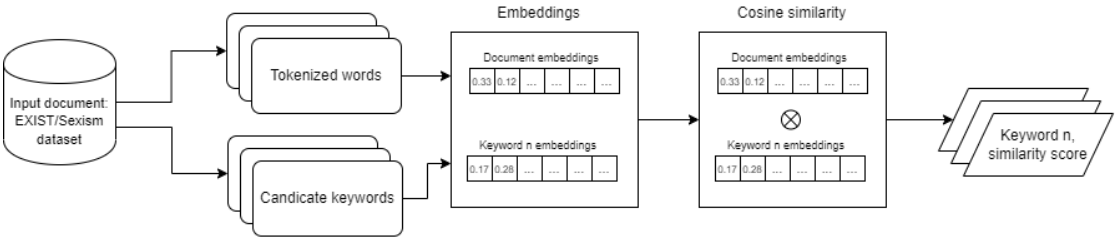

Figure 1. KeyBERT flow.

```
[('why feminism keeps', 0.6194), ('man free feminist', 0.615), ('to feminism njabulodhlamin3', 0.6132), ('instincts of patriarchy', 0.5
[('about sexism amp', 0.6023), ('the new feminist', 0.5915), ('man hating feminists', 0.5845), ('sexism amp mansplaining', 0.5821), ('p
[('liberal white women', 0.5236), ('policing women and', 0.4918), ('victims of feminismfeminism', 0.4863), ('minister auspolformer libe
[('tweet from feminist', 0.6271), ('feminist account saying', 0.5995), ('feminist attempt to', 0.5745), ('misogyny through critiques',
[('mansplaining against women', 0.6136), ('backlash against feminism', 0.5902), ('feminism you call', 0.587), ('patriarchy women patria
[('crisis in masculinity', 0.5725), ('masculinity and damage', 0.5284), ('society is men', 0.5098), ('married women depends', 0.4828),
```

Figure 2. A sample output of 3-gram keywords.

Keywords selection and filtering. The KeyBERT method extracted a total of 1350 keywords and phrases. Subsequently, 177 duplicates were eliminated from the list. Before annotating and selecting keywords and key phrases, we performed filtering and discovered that 3-gram key phrases did not return any results. Thus they were excluded from the list, resulting in 693 potential keywords and key phrases. In the process of selecting keywords, two participants were assigned to independently review the remaining keywords. They were asked to categorize the keywords into three groups:

- Include - Keywords or key phrases that demonstrate a clear thematic relevance, have a high similarity score, or have been associated with at least 20 posts on Reddit within the past year.
- Remove - Keywords or key phrases lacking thematic relevance, being either too specific, too generic, not present in the vocabulary, containing spelling errors, having a low similarity score, or failing to meet the minimum threshold of 20 posts on Reddit within the past year.
- Not decided - An interim classification that is discarded until achieving consensus.

After each participant completed their review, they convened to discuss the undecided category and resolve any disagreements. To measure the inter-rater reliability, we chose Cohen's kappa coefficient method [25], as it provides a robust measure of agreement between raters for nominal variables. The Cohen's kappa coefficient is calculated as follows:

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{3}$$

where: κ is the Cohen's kappa coefficient, P_o is the observed agreement between raters, and P_e is the expected agreement by chance [25]. We got the κ coefficient value = 0.75, which according to the interpretation scale indicates that the agreement between two raters is moderate. After that, the posts were carefully examined and deliberated by the annotators to resolve any conflicts. Through negotiations and discussions, both annotators reached a consensus on the final category for each post. Combined with initial seed keywords, the finalized list consisted of 441 keywords and key phrases, which resulted in 4111 filtered posts.

3.3.2 Manual annotation of data

The manual annotation of data was similar to the keywords annotation and selection and included a thorough review of the content of each Reddit post that contained gender-related keywords to decide if a post was related to gender discussions. The workflow of this process started with the independent review of all 4111 posts by the assigned annotators, followed by a mutual discussion of the initial annotation decision, and ended with a final decision for each post. It is noteworthy, that one of the annotators assigned for keywords annotation in subsection 3.3.1 was substituted by a newly assigned one to diversify the annotators' background. The final decision was reached by negotiations and discussion between the two raters. Cohen's kappa coefficient method [25] was also used to measure the inter-rater reliability, which resulted in a coefficient value = 0.68 and was interpreted as moderate based on the interpretation scale. In total, the annotation took two weeks to finish. Before the first round of review and annotation process began, it was decided to categorize posts into the following groups:

- **Gender-related** - posts that contain clear discussions about gender inequality, sexism, and other evident mentionings of these topics.
- **Potentially gender-related** - posts that contain keywords although relatedness to the topic is uncertain.
- **Not gender-related** - posts that while containing words related to gender are not about gender discrimination or sexism.

3.3.3 Results of annotation

The results of the manual review and annotation showed that although the posts contained gender-related keywords, only a fraction of them were related to the actual discussion. Out of the 4111 posts potentially containing gender discussions, only 119 were related to the topic, and the other 3992 posts were labeled as "not gender related". The remaining 93234 posts that did not contain any gender-related keywords were considered irrelevant to the topic of the research and were excluded from the dataset. We observed that words like "looks," "woman," and "man" resulted in many false positives, and 2-gram keywords were not found in our data at all. Such keywords proved to be uninformative. The end-to-end process of the dataset preparation is shown in Figure 3.

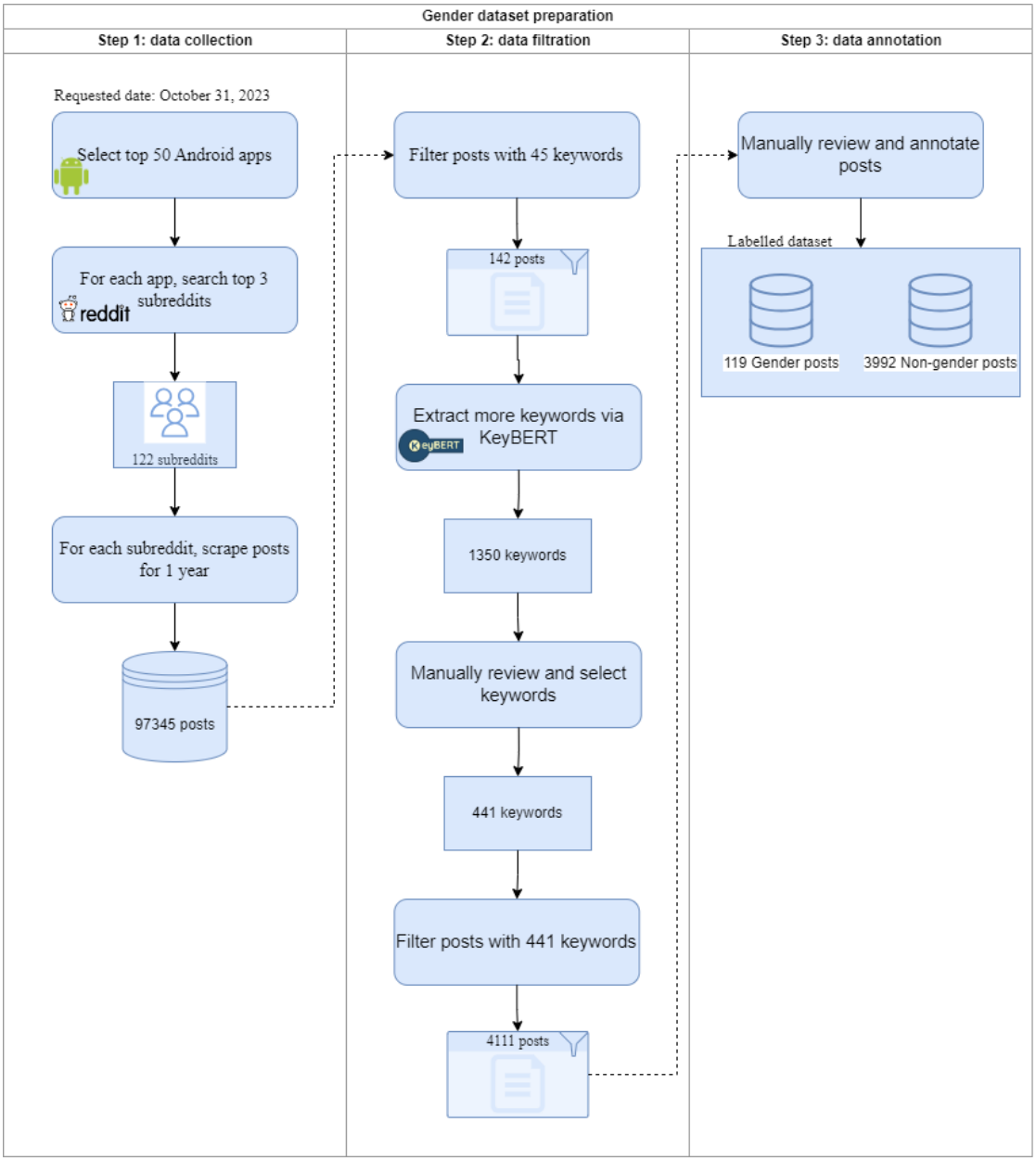

Figure 3. Dataset creation

3.4 Using generative artificial intelligence and other tools

This thesis employed the use of tools such as ChatGPT [9] and PerplexityAI [10] as a preliminary step in the investigation of novel concepts and methodologies, as well as to organize the description of the methodology, experiments, and interpretation of results. For the purpose of translating the text or identifying synonyms in the English language, the Deepl [11] Translator was utilized. Grammarly [12] was employed to proofread the grammar of the written text. Despite the extensive use of these tools throughout the thesis, the content was entirely written by the author.

[9] https://chat.openai.com/
[10] https://www.perplexity.ai/
[11] https://www.deepl.com/write
[12] https://app.grammarly.com/

4 Automated classification

This chapter describes the analysis of popular classifiers and data preparation for automated classification and demonstrates the evaluation metrics and performance results.

4.1 Data pre-processing

The necessary data cleaning and pre-processing steps were performed to prepare the raw data for automated classification. This process is important to ensure the quality and reliability of the further analysis. Data pre-processing was performed by using the Natural Language Toolkit, NLTK. [13]

Normalization Techniques. The following normalization techniques were applied:

- **Removing punctuation and emojis**: punctuation and emojis can introduce noise and distort the semantic meaning of the text. Excluding these elements helps to focus the analysis on the main content of the text.
- **Converting text to lowercase**: Converting all text to lowercase ensures consistency and prevents the model from treating uppercase and lowercase words as separate entities, which could cause potential performance issues.
- **Removing stop words**: Words that occur frequently but hold little or no meaning or information, such as "a", "the", "is", and "are", were removed from the text. Like the previous step, removing stop words helps to decrease the computational time by dimensionality reduction of the textual data and focusing the analysis on more informative features.

Tokenization and Lemmatization. The clean text was further processed by:

- **Tokenization**: The text was split into individual words or tokens, which is a crucial step for the natural language processing tasks.
- **Position Encoding**: The tokens were further assigned position encodings, which provide the model with information about the relative position of each word in the text based on the part of speech. This step was applied to increase the efficiency of lemmatization and preserve the structure and context of the language.
- **Lemmatization**: The tokens were converted into their base or dictionary form (lemmas) to generalize the semantic relationship between words and thus improve the classification results.

[13] https://www.nltk.org/

Table 3. The example of the pre-processed text.

Raw text	Pre-processed text
Your filter is extremely sexiest and discriminatory,"A man in a swimsuit" is not a blocked prompt but ""a woman in a swimsuit"" is blocked"" a man having a shower"" not blocked ""a woman having a shower"" blocked.""A womans chest"" blocked ""a man's chest"" ....? I don't know I just got blocked for the next hour lol	'filter', 'extremely', 'sexy', 'discriminatorya', 'man', 'swimsuit', 'block', 'prompt', 'woman', 'swimsuit', 'block', 'man', 'shower', 'block', 'woman', 'shower', 'blockeda', 'womans', 'chest', 'blocked', 'man', 'chest', 'dont', 'know', 'get', 'blocked', 'next', 'hour', 'lol'

Data Merging. The pre-processed posts ("Post_body") and post titles ("Title") were merged into a single column, "Text", to simplify the classification task. This approach combines the relevant textual information into a single data item, which can improve the model's ability to capture the overall context and semantics of the data.

The example of the raw text taken from one of the Reddit posts is given in Table 3 to illustrate the outcome of the data pre-processing step.

4.2 Feature extraction

After completing the cleaning and pre-processing steps of textual data, it was necessary to convert it into a numerical format to prepare the data for machine learning models. This process is commonly referred to as feature extraction or feature selection. In the context of this experiment, two different feature extraction methods have been used: the Bag of Words (BoW) model and the Term Frequency-Inverse Document Frequency (TF-IDF) approach. Feature selection was performed using the Machine Learning Python library - scikit-learn 1.4.2 [14].

The **Bag of Words (BoW)** model [50] is an essential method of natural language processing that involves converting textual data into a numerical matrix representation. This model focuses entirely on the occurrence of specific words in the text, ignoring the order of the words in the sentences. While BoW is considered as a simple and efficient method, it has limitations, such as the inability to recognize the degree of importance of every word in the document.

On the contrary, the **Term Frequency-Inverse Document Frequency (TF-IDF)** [35] represents a more advanced approach that circumvents the limitations of BoW. TF-IDF determines the significance of each word within a text by considering its frequency and the number of documents in which it appears. This method allocates a greater value to those words that are exclusive to a specific document, thereby offering a more sophisticated portrayal of the text. TF-IDF is calculated by the following formula [35]:

[14]https://scikit-learn.org/stable/index.html

$$w_{i,j} = \text{tf}_{i,j} \times \log\left(\frac{N}{\text{df}_i}\right) \tag{4}$$

tf_i, j = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents

After a thorough comparison of both methods in practice, the TF-IDF approach was deemed to be more effective for this research, as it provides a more accurate representation of the text's semantic content. Therefore, the TF-IDF method has been utilized for the feature extraction phase of this study.

4.3 Addressing the class imbalance in Reddit data

Data level. The dataset collected and manually annotated during this study resulted in a highly imbalanced distribution of the target class (gender-related data), due to the small number of posts containing gender discussions. This imbalance can produce a potential problem for automated classification by machine learning models as they would have difficulty in recognizing the minority class and ultimately lead to poor prediction accuracy. Figure 4 shows the class distribution in our dataset.

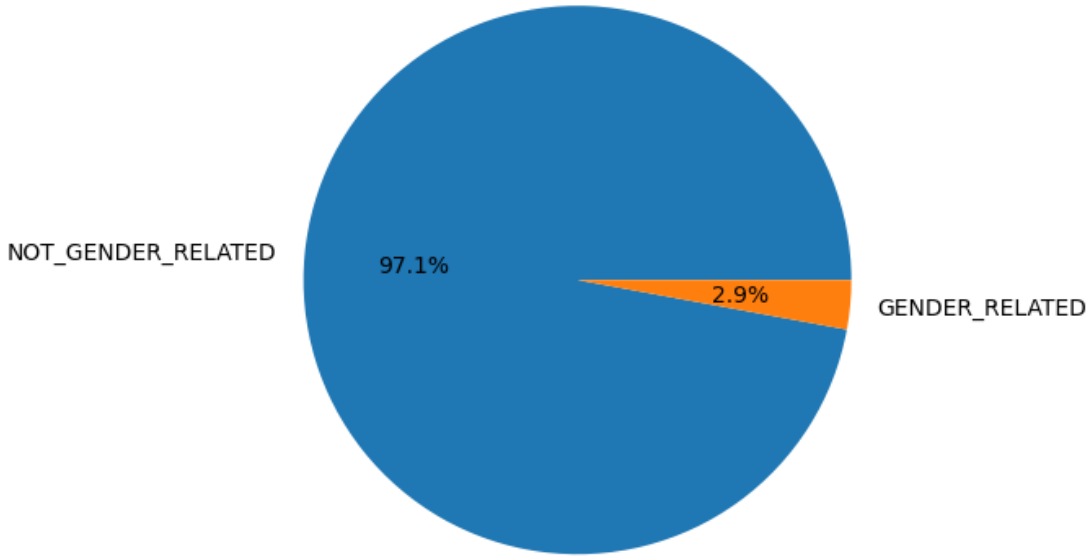

Figure 4. Imbalanced data.

To address this issue, various statistical techniques have been developed to enhance the representation of minority and majority classes in textual data. These techniques

include over-sampling the minority class and under-sampling the majority class. The efficacy of different approaches varies depending on the size of the dataset and the task at hand. Given the relatively modest size of our dataset (4,111 samples), selecting the under-sampling approach and eliminating the samples from the majority class, thereby further reducing the dataset, may not be an optimal strategy for model training. The Synthetic Minority Over-sampling Technique (SMOTE) generates synthetic examples of the rare class by selecting and joining the k nearest neighbors from the minority class, where k is selected based on the number of required new samples [6]. The SMOTE (Synthetic Minority Over-sampling Technique) method has the advantage that it does not simply copy existing original samples, but rather generates new samples that are similar to the original ones. Consequently, it can help to prevent the model from becoming overly fitted. Figure 5 illustrates the data resampling process using the SMOTE technique. The default sampling strategy was employed to achieve the desired 1:1 ratio.

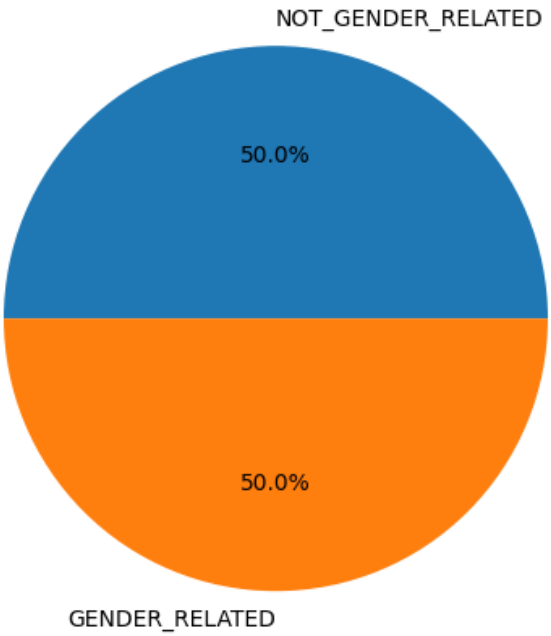

Figure 5. Resampled data.

4.4 Classification models

We have decided to work with both the traditional supervised machine learning models and the unsupervised neural networks for automated classification experiments to demonstrate how different classifiers would perform on our data. The primary objective was to investigate the effectiveness of various classification models in order to identify the most

appropriate approach for the data in question. While the potential for further optimization was acknowledged, the focus was on experimentation rather than fine-tuning for optimal results, which was beyond the scope of this research question. Similarly to the feature selection, the conventional machine learning classification tasks were performed using the scikit-learn 1.4.2 Python library [15].

The classifiers were selected based on the following criteria: (1) The classifiers must have been previously used in similar research and have demonstrated positive results. (2) The classifiers should contain both the traditional machine learning models and the state-of-the-art neural networks. The final list of classifiers included: Support Vector Machine [9], Logistic Regression [10], Decision Tree [33], Naive Bayes [21], DistilRoBERTa [41], and Meta Llama 3 [1]. These classifiers either showed good performance in the previous studies [28], [17], [4], or were ranked on of the best in the classification task [16], [17].

Support Vector Machine (SVM). The Support Vector Machine (SVM) [9] is a supervised machine learning algorithm that identifies the optimal hyperplane to separate two classes of data. This is achieved through the use of a kernel function, which transforms the data into a high-dimensional space. The linear kernel function was selected as the optimal choice for this data set, as it allows for linear separation of the data. It is calculated as follows [9]:

$$Kx, y = x^T \cdot y \tag{5}$$

where x and y are the input vectors, and x^T denotes the transpose of x. The linear kernel function computes the input vectors' dot product, which measures their similarity.

Logistic Regression. Logistic Regression [10] is a supervised machine learning algorithm that belongs to the discriminative probabilistic classifiers. It uses a logistic function to compute the probabilities of the sample values and map them between 0 and 1 based on the input features. The logistic function is based on the sigmoid function and is calculated as follows [10]:

$$f(x) = \frac{1}{1 + e^{-x}} \tag{6}$$

where x is a real value and $f(x)$ is a calculated probability.

Decision Tree. The Decision Tree classifier [33] is a supervised machine learning algorithm that can solve both classification and regression problems. The decision trees are constructed by applying partitioning conditions at each node that divide the training

[15] https://scikit-learn.org/stable/

[16] https://huggingface.co/distilbert/distilroberta-base

[17] https://huggingface.co/meta-llama/Meta-Llama-3-8B

instances into subsets with the same outcome class. The aim at each node is to find an attribute and a split condition on that attribute that minimizes the confusion of class labels, resulting in close to clean subsets.

Naive Bayes. Naive Bayes [21] is a supervised machine learning algorithm that belongs to the generative probabilistic classifiers. It is based on the Bayes theorem and the assumption that the words or input features in the document are independent given the output class. For our task, we utilized the Gaussian Naive Bayes classifier [19], which is based on Bayes' theorem and assumes that the features are continuous and follow a Gaussian distribution. The formula for Gaussian Naive Bayes can be expressed as follows [19]:

$$P(y|x_1, x_2, ..., x_n) = \frac{P(y) \prod_{i=1}^{n} P(x_i|y)}{P(x_1, x_2, ..., x_n)}$$

where: - $P(y|x_1, x_2, ..., x_n)$ is the probability of class y given features $x_1, x_2, ..., x_n$, $P(y)$ is the prior probability of class y, $P(x_i|y)$ is the conditional probability of feature x_i given class y, and $P(x_1, x_2, ..., x_n)$ is the probability of observing features $x_1, x_2, ..., x_n$.

DistilRoBERTa DistilRoBERTa is a pre-trained large language model that is based on a distilled version of the RoBERTa model. It follows the same training process as the DistilBERT model [41]. It is based on the BERT architecture and utilizes a distilled training approach to achieve a smaller size without compromising accuracy. The DistilRoBERTa English language model was trained using supervised learning techniques with the objective of replicating the capabilities of the Roberta-base model, which was itself trained with a substantial volume of data from the OpenAI WebText dataset. The DistilRoBERTa model was developed with a training data set comprising only approximately 4 times less data than the RoBERTa model used as a reference point. The model comprises six layers, with a dimension of 768 and 12 heads, resulting in a total of 82M parameters (in comparison to 125M parameters for RoBERTa-base). In general, DistilRoBERTa is observed to perform at a rate approximately twice that observed for Roberta-base [41]. For our experiment, we utilized the "distilroberta-base" pre-trained model [18] and fine-tuned it on our labeled dataset.

Llama-3-8B Llama-3-8B is the latest of LLaMA large language models introduced by Meta AI which was pre-trained and instruction-fine-tuned with 8 billion parameters and can be used for different tasks including sentence sequence classification [1]. While similarly to the previous generations of LLaMa models, Llama-3-8B is openly available and was trained on solely publicly available datasets, the training dataset of Llama-3-8B model is 7 times larger compared to the previous model [45], [46]. The model architecture

[18] https://huggingface.co/distilbert/distilroberta-base

repeats the Llama 2 and utilizes its own tokenizer of 128 000 tokens. Since the fine-tuning of large language models like Llama requires the entire list of parameters, it is computationally costly to train such models on new data. Given the number of parameters of LLama 3, we applied the Parameter-Efficient Fine-Tuning (PEFT) technique called Quantized Low-rank Adaptation (QLoRA) [19], which allows to fine-tune only a small number of extra weights in the model while freezing most of the parameters of the pre-trained network.

4.5 Evaluation metrics.

The selection of evaluation metrics is dictated by the class imbalance of the obtained dataset. Similarly to [43], the following metrics were used to assess the performance of the models: accuracy, precision, recall, F1-score, and AUC.

Accuracy. In a binary classification task, accuracy [27] is used as a statistical measure to calculate how many observations out of a total number of observation were classified correctly. Its formula is as follows [27]:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

where TP is True positive, TN is True negative, FP is False positive, and FN is False negative.

Precision. Precision [32] is the proportion of true positives (correctly predicted positive instances) out of all positive predictions made. It is a useful metric when the cost of a false positive is high. It is calculated as [32]:

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

Recall. Recall [32] is the proportion of true positives out of all actual positive instances. It is a useful metric when the cost of a false negative is high. It is calculated as [32]:

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

F1-score. The F1 score [32] is the harmonic mean of precision and recall and is a useful metric when both false positives and false negatives are costly. Its formula is as follows:

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{10}$$

[19] https://github.com/artidoro/qlora

AUC. AUC (Area Under the Curve) [15] is a measure that refers to the Receiver Operating Characteristic (ROC) curve. A ROC curve is a 2-dimensional graph that depicts the performance of a binary classification model at different threshold settings. The AUC metric is employed to calculate the area under the ROC curve. The area under the ROC curve (AUC) is calculated using the AUC metric. The AUC metric is employed to calculate the area under the ROC curve. AUC values closer to 1 indicate a superior model performance, whereas values approaching 0.5 suggest a model with random outcomes.

4.6 Results of automated classification

The results of the automated classification contain the comparison of the performance between the models mentioned in subsection 4.4 per class with and without resampling with SMOTE, and with different feature extraction approaches. Each configuration was evaluated with the metrics described in the previous subsection. The final values are collected in Table 4, where: P stands for Precision, R is Recall, F1 is F1-score, ACC is Accuracy, and AOC is Area Under the Curve. Configurations consist of different combinations of models, feature extraction methods, and oversampling techniques.

In our study, the classifiers consistently demonstrated better performance in predicting the non-gender-related majority class than the gender-related class. As it can be seen from Table 4, all models achieved evaluation metrics above 90 percent in the majority class compared to the minority class. It was anticipated that the outcomes of the two classes would diverge significantly given that the dataset was inherently imbalanced and that the gender-related class was likely to have fewer instances to train, resulting in lower precision and recall.

Further analysis determined that our most successful configuration was the Support Vector Machine (SVM) classifier using TFIDF features together with SMOTE oversampling technique. It is noteworthy that this configuration reached a precision of 91 percent for the minority class, indicating its effectiveness in handling unbalanced data. In addition, it achieved an impressive 99 percent accuracy for the majority class, demonstrating its robustness across both classes.

On the other hand, the Gaussian Naive Bayes classifier produced the least convincing results in our experiments. In particular, its precision for the minority class remained plateaued at only 11 to 13 percent.

Additionally, the deep learning models DistilRoBERTa and Llama-3-8B demonstrated overall good performance. We deliberately did not apply oversampling of the minority class of the dataset to test the performance of neural networks on our data as it is. While both models demonstrated high precision, with the values of recall and F1-score close to 1 in the majority class, Llama-3-8B outperformed DistilRoBERTa in the minority class with the precision of 0.75 compared to 0.40, recall of 0.43 against 0.30 and F1-score of 0.55 in comparison to 0.34. When put together with the results of the conventional

machine learning classifiers, Llama 3 can compete with our best configuration of TFIDF-SVM-SMOTE in performance for both classes. On the other hand, DistilRoBERTa was second to the configurations of the Decision Tree classifier with slightly lower precision and recall.

Table 4. Evaluation results per configuration

Configuration	NOT GENDER RELATED					GENDER RELATED				
	P	**R**	**F1**	**ACC**	**AUC**	**P**	**R**	**F1**	**ACC**	**AUC**
BOW + SVM	0.99	0.99	0.99	0.98	0.91	0.78	0.64	0.70	0.98	0.91
BOW + LR	0.99	1.00	0.99	0.99	0.92	0.88	0.56	0.69	0.99	0.92
BOW + DT	0.98	0.99	0.99	0.98	0.75	0.61	0.49	0.54	0.98	0.75
BOW + NB	0.98	0.95	0.96	0.93	0.57	0.11	0.21	0.14	0.93	0.57
TFIDF + SVM	0.98	1.00	0.99	0.98	0.95	1.00	0.33	0.50	0.98	0.95
TFIDF + LR	0.97	1.00	0.99	0.97	0.97	1.00	0.03	0.05	0.97	0.97
TFIDF + DT	0.99	0.98	0.98	0.97	0.76	0.48	0.54	0.51	0.97	0.76
TFIDF + NB	0.98	0.95	0.96	0.93	0.59	0.12	0.23	0.16	0.93	0.59
BOW + SVM + SMOTE	0.98	0.93	0.95	0.91	0.55	0.11	0.31	0.16	0.91	0.55
BOW + LR + SMOTE	0.98	0.95	0.96	0.93	0.64	0.16	0.31	0.21	0.93	0.64
BOW + DT + SMOTE	0.98	0.95	0.96	0.93	0.70	0.11	0.21	0.14	0.93	0.70
BOW + NB + SMOTE	0.98	0.95	0.96	0.93	0.57	0.11	0.21	0.14	0.93	0.57
TFIDF + SVM + SMOTE	0.99	1.00	0.99	0.99	0.95	0.91	0.54	0.68	0.99	0.95
TFIDF + LR + SMOTE	0.99	0.99	0.99	0.98	0.97	0.64	0.60	0.62	0.98	0.97
TFIDF + DT + SMOTE	0.99	0.96	0.97	0.95	0.76	0.30	0.54	0.39	0.95	0.76
TFIDF + NB + SMOTE	0.98	0.95	0.97	0.93	0.59	0.13	0.23	0.17	0.93	0.59
DistilRoBERTa	0.98	0.98	0.98	0.96	0.65	0.40	0.30	0.34	0.96	0.65
Llama-3-8B	0.99	1.00	0.99	0.99	0.75	0.75	0.43	0.55	0.99	0.75

To better illustrate the AOC metric, the ROC curves for each configuration are depicted in Figures 6, 7, 8, 9, 10, and 11. The ROC plots illustrate the manner in which the selected classifiers reflected the overall performance. The SVM and Logistic Regression classifiers exhibited the most favorable outcomes in nearly all configurations, with the exception of the combination of BoW feature selection and SMOTE oversampling, where the curve was nearly parallel to the diagonal with scores of 0.55 and 0.64, respectively. The highest area under the curve score was achieved when the TFIDF method was applied, with values of 0.95 for the SVM and 0.97 for the Logistic Regression. In contrast, the Naive Bayes and Decision Tree classifiers produced curves that resembled diagonal lines, with AUC scores between 0.57 and 0.59 for Naive Bayes and between 0.70 and 0.76 for Decision Tree. These results indicate that the predictions of these classifiers were as good as a random choice. The DL models exhibited moderate results, with DistilRoBERTa achieving an AUC score of 0.64 and Llama-3-8B achieving an AUC score of 0.71.

Additionally, the confusion matrices for each configuration are depicted in Figures 12, 13, 14, 15, 16, and 17. There is a clear trend in the plots showing stronger results for the true prediction of the majority class of "Non gender-related" samples compared to the number of correctly predicted "Gender-related" samples.

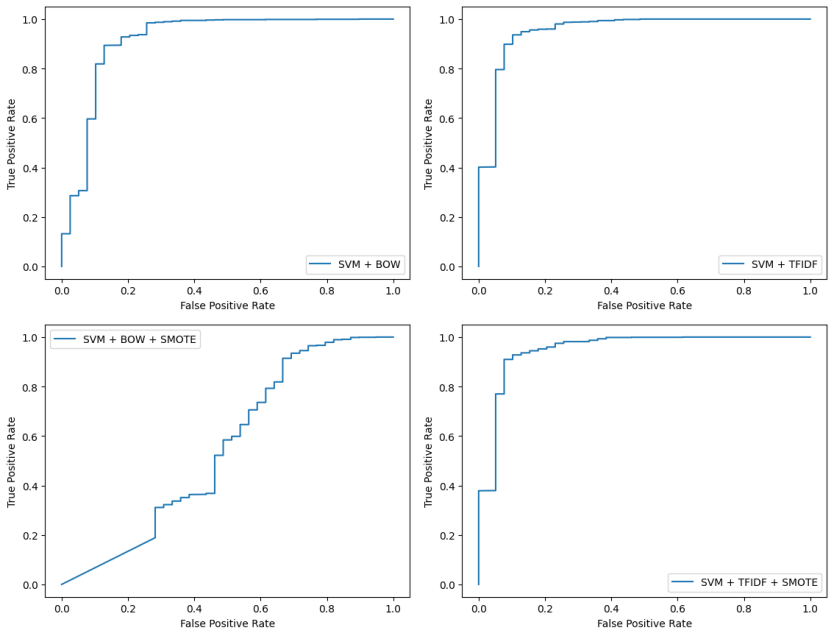

Figure 6. ROC curves for Support Vector Machine configuration models

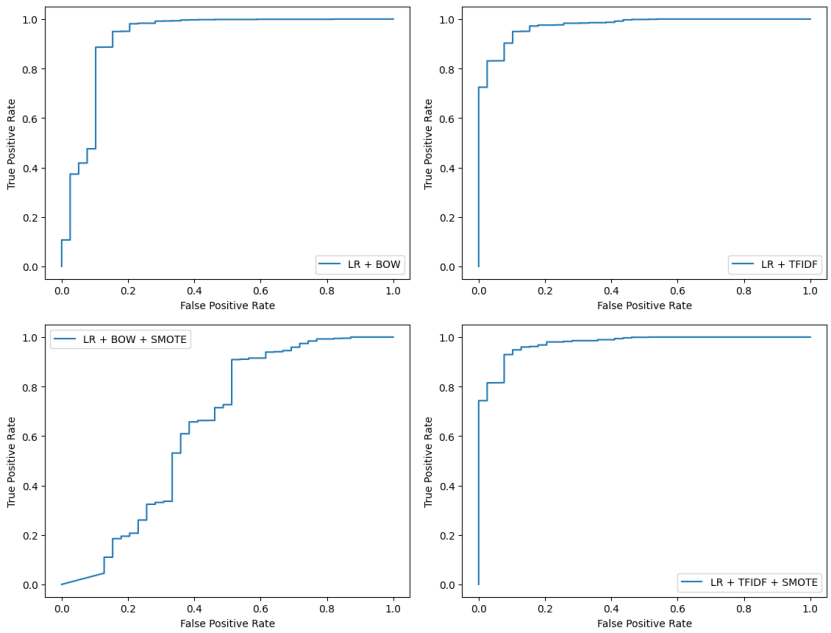

Figure 7. ROC curves for Logistic Regression configuration models

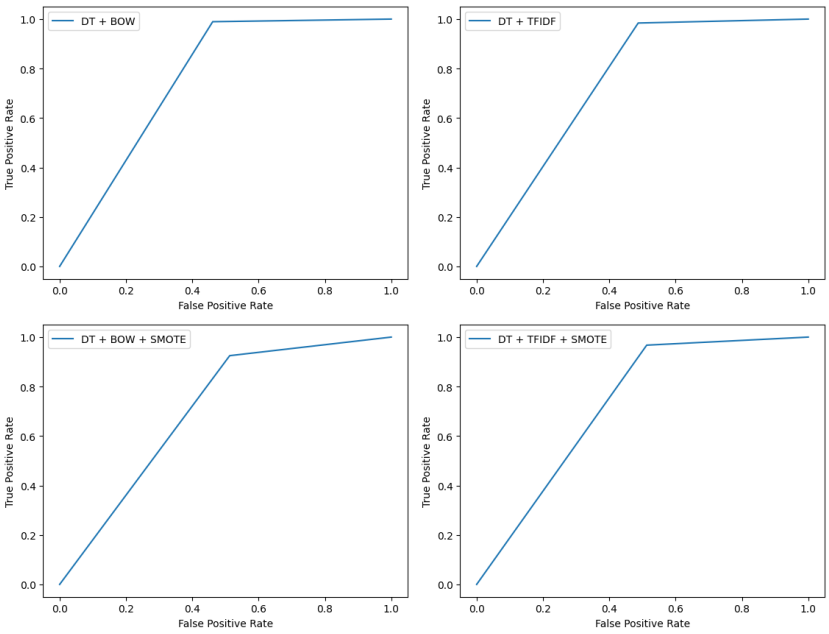

Figure 8. ROC curves for Decision Tree configuration models

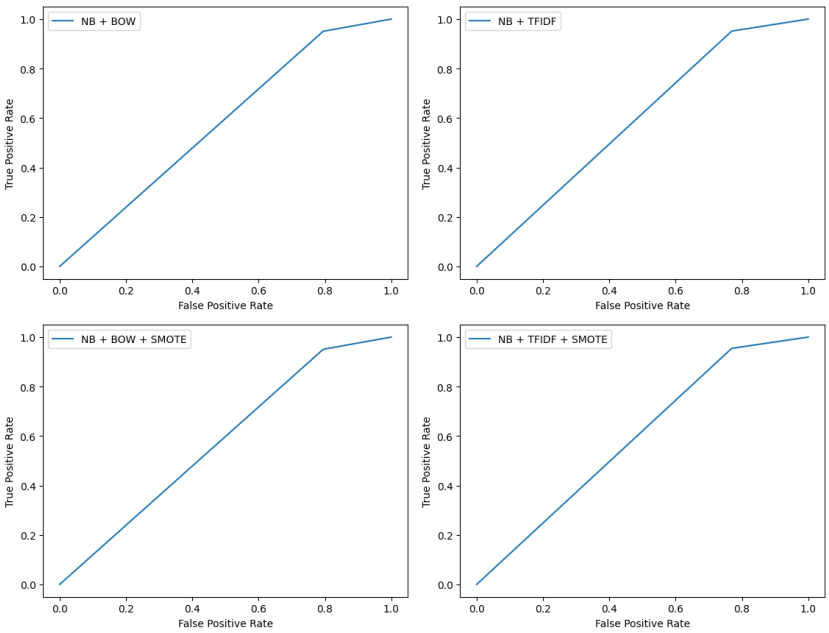

Figure 9. ROC curves for Naive Bayes configuration models

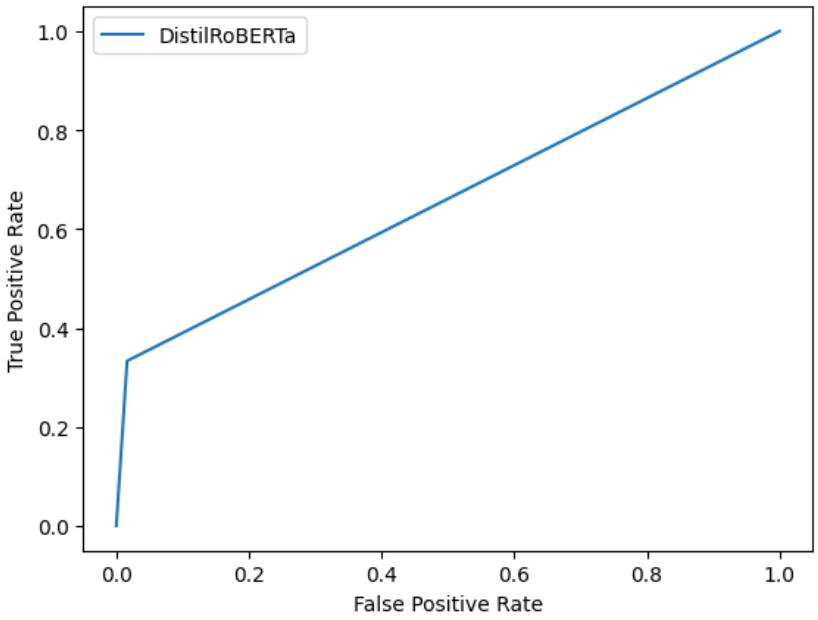

Figure 10. ROC curve for DistilRoBERTa model

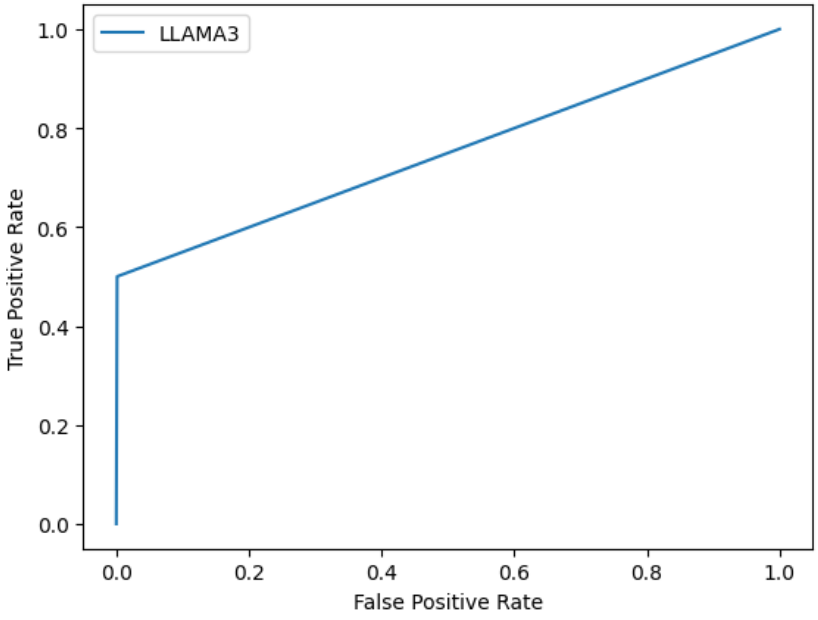

Figure 11. ROC curve for Llama-3-8B model

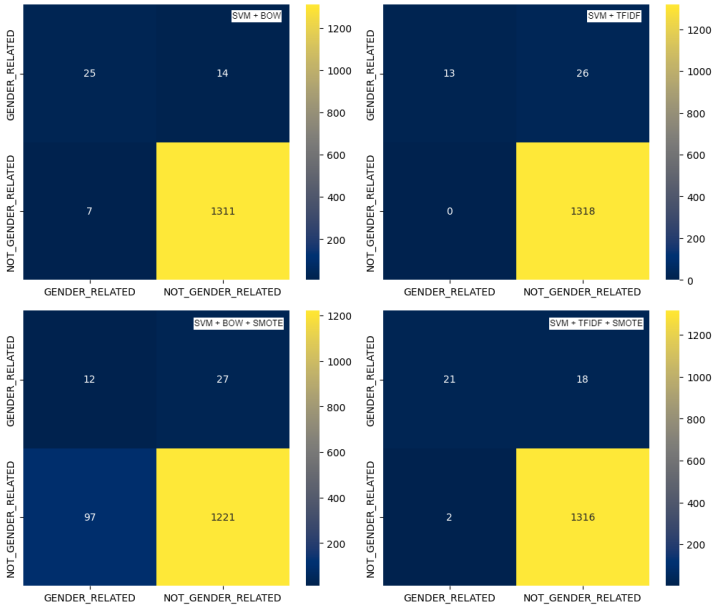

Figure 12. Confusion matrix for Support Vector Machine configuration models

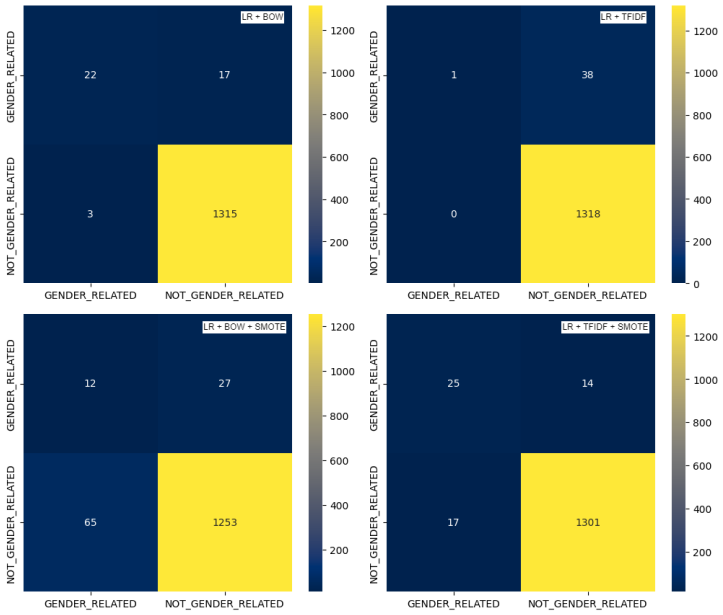

Figure 13. Confusion matrix for Logistic Regression configuration models

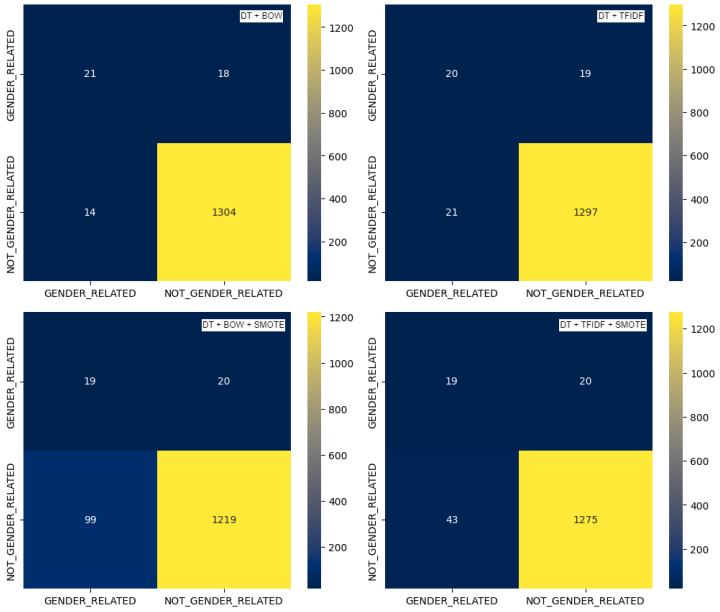

Figure 14. Confusion matrix for Decision Tree configuration models

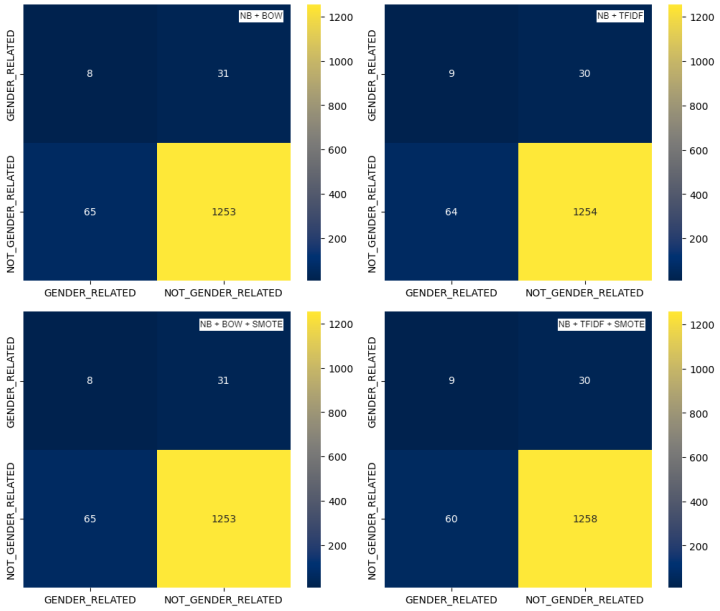

Figure 15. Confusion matrix for Naive Bayes configuration models

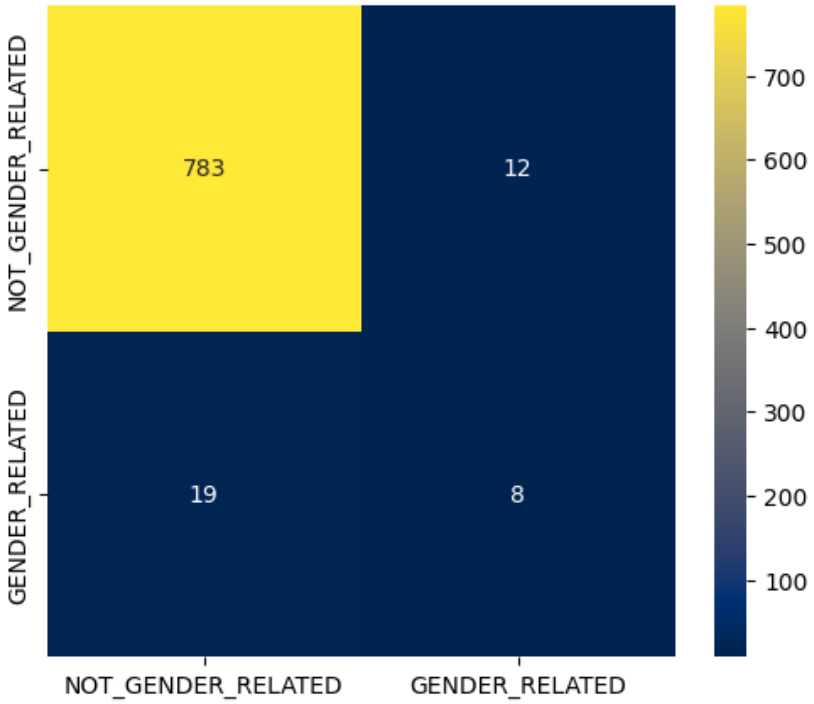

Figure 16. Confusion matrix for DistilRoBERTa model

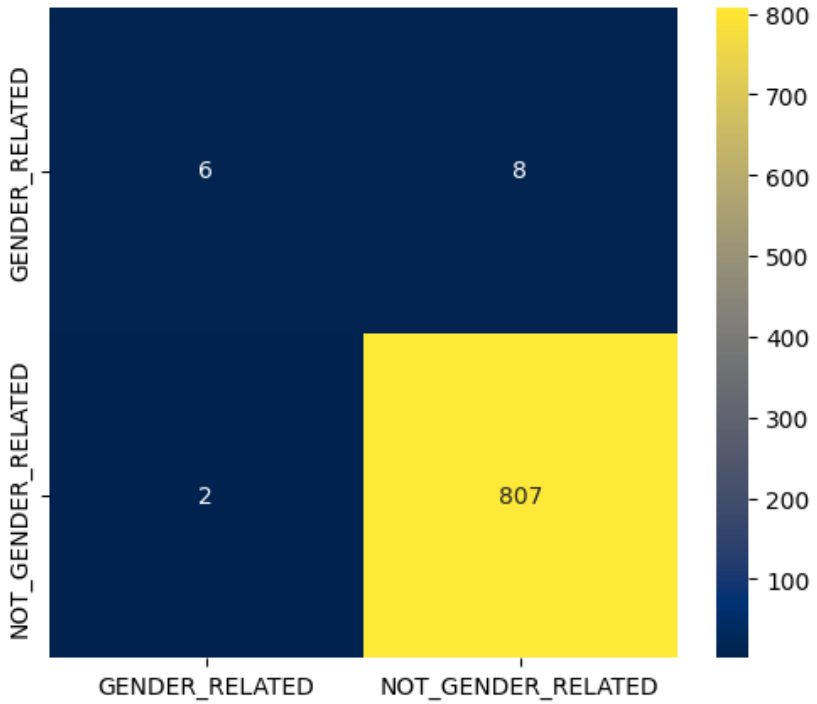

Figure 17. Confusion matrix for Llama-3-8B model

5 Categories of gender discussions

To answer the research question **RQ2**, "Can we further categorize the gender-related discussions in Reddit posts?", we explored the data that was collected and annotated to find trends in gender discussions. This section describes the process of derivation of categories inside the gender-related data. Each category choice is explained and supported by examples from the analyzed data.

5.1 Categorization approach

Similarly to the study [43], we consulted the open coding method from the grounded theory, as outlined in [12]. Grounded theory is a qualitative research method that allows the extraction of new theories from iteratively collected real-world data. The open coding method begins with the decomposition of the transcripts of data into excerpts, which are then grouped into codes. This process is repeated by collecting more data, which is then reviewed to ascertain its alignment with the codes. In the course of our research, we reviewed the initial set of previously collected Reddit posts that had been annotated as "gender-related." We then extracted the principal topics of these posts and grouped them into common categories. We added the next set of data and reviewed it based on our previously established codes. This process was repeated four times and took two days in total.

5.2 Results of categorization

The categorization of gender-related data led to the derivation of the six categories and associated concepts presented in Table 5. We will explain these categories below.

Advertisement. The category "Advertisement" contains posts that provide complaints about different issues with advertisements in mobile apps. We identified the following topics: gender representation and identity on digital platforms, inappropriate and sexually explicit advertising in online content, and algorithmic bias and content curation on online platforms. For example, in the next two posts from the subreddit "FacebookAds", the users complain that there is no option to configure ads for the selected gender groups and as a result, their ads attract the wrong audiences:

"Detail targeting....Did Meta just get rid of everything? I don't have the option for Gender targeting anymore, and interest."

"Algorithm Showing Ads to old men on FB. My website is clothing aimed towards 18 to 41. Did an interest stack with the help of a fb marketing expert from Meta for VC. No sales and demo breakdown shows ad was show to old men on fb. Hardly any ads show on insta. Data is absolutely rubbish can't even run retargeting or LA."

Table 5. Categories of gender discussions.

Category	Topics
Advertisement	Gender representation and identity in digital platforms Inappropriate and sexually explicit advertising in online content Algorithmic bias and content curation on online platforms
AI	Gender bias and misinterpretation in language models Gender representation and identity in digital platforms LGBTQIA+ discrimination Navigating discussions on gender identity and online moderation Sexual objectification of women
App features	Account settings Video and conference filters Search algorithms Privacy and safety measures Emojis
Community	Hate speech and toxicity in online communities
Company Policy and Censorship	Gender-biased content moderation on social media platforms Censorship and content restrictions in online research platforms Accountability of online platforms in addressing sexual misconduct
Content	Algorithmic bias and content curation on online platforms Sexist, inappropriate or fraudulent content

AI. All user posts that addressed discussions about AI-based applications and algorithms used in mobile apps were included in this category. A notable trend emerged in the discourse surrounding gender bias in chatbots, where the representation of women was often sexualized, stereotyped, or entirely blocked. As an example, the next two posts from the "Bing" and the "OpenAI" subreddits tell:

"No matter what I prompt, bing is not able to generate images of women that aren't fitting classic beauty standards. But will just fine for men. I am working on a project and I need images of people who don't fit into the standard beauty aesthetics. Eg. People with larger bmis, non symmetrical faces and such. However bing image create will generate this fine for men but whatever prompt I try for women it just won't do it. They always fit classic beauty standards or just no image at all! Can anyone help suggest a prompt to get around this. I've tried wording it many different ways! Maybe I'm just being an idiot?"

"Your filter is extremely sexiest and discriminatory. "A man in a swimsuit" is not a blocked prompt but "a woman in a swimsuit" is blocked " a man having a shower" not blocked "a woman having a shower" blocked. "A womans chest" blocked "a man's chest" ....? I don't know I just got blocked for the next hour lol"

There was another trend, where chatbots either rejected LGBTQIA+ prompts or failed to provide meaningful results. The following post and comment from the "ChatGPTPro" subreddit tell:

"As a gay man, I actually feel discriminated against with bings moderation... Not for lesbian couples! My friend can't seem to stop getting them, even with nudity that isn't asked for! It's insane."

"Lesbian kisses violate content policy but straight and gay ones do not. I asked it to

generate a picture of two women kissing, and one of a man and a woman kissing. It only generated the straight couple. For some reason, it was willing to generate a gay male kiss too - but lesbians violate the content policy apparently. Disappointing."

Another post from the "Bing" subreddit shows how the moderation and censorship is biased towards women:

"Bing might've alleviated the censors but... Do they actually hate women??? It's honestly ridiculous how completely safe prompts get so heavily blocked. This is what I'm trying to create: A serious woman, with long flowy white and light blue hair, light orange eyes, an ethereal white princess dress with a Chinese-inspired light blue pattern, white heels, render, anime style, facing the camera, single subject, stiff pose, full body shot, white background It's totally sfw, right? I'm not interested in making it lewd at all, it's not my intention. I just want to create a godly princess and I can't. The result is the image I showed (a ton of blocked results) I want to create a full body image, it doesn't work, it never shows the legs too much and even less the shoes. I added "white heels" on purpose to see if it did but it doesn't, and it gets blocked like 9 times out of 10 anyway. I want to make her barefoot because it adds an interesting flair, it gets blocked. I don't understand why the censors are so strict when making women, with men it allows for a lot more creativity."

App features. In the category "App features", we put the posts that talk about different functionalities of apps, such as search and filters, account settings, emojis, and privacy and safety measures. In the following post, a user is concerned with how changing their name will affect certain processes after recent gender transition:

"Will changing my name to my preferred name mess with the end of year tax documents? Hello, Recent transgender here. I am looking to change to my preferred name on DoorDash. I know I can do this in the app. However, will the tax documents have my legal name or preferred name? I have not changed my legal name yet so my taxes needs to be done with my legal name. Wondering if anyone has experience with this before I make the change. Thanks!"

Another post from the "Whatsapp" subreddit discusses how gender related emojis are not supported on certain platforms:

"Emoji genders confusion. Whatsapp is being weird for me. This isn't an invitation to discuss genders. Male friends are using emojis representing their gender and whatsapp is showing me them as female (preserving the skin colour though, at least). What is happening, how do I remove the ghost from the machine?"

Community. We observed some posts from the subreddits dedicated to the social networking apps, which had users talking in different forms about hate speech and gender discrimination. Both of the following posts have been taken from the "Instagram" subreddit:

"The Reels comment section makes me lose faith in humanity. Considering how Insta is a lot of people's main online presence, there is an insane amount of bigotry in the comment section of almost all the of reels I watch on there. The video will be completely unrelated to anything and there will be people finding a way to be terrible. And if it's not bigotry, it's just people being rude to whoever is in the vid It's definitely more than TikTok and even YouTube shorts at this point (which is also pretty bad) I can't tell if it's an algorithm or censorship thing."

"Awful comments. This poor person got run off the internet because of the hate! The kind of content they made was furry based and yes a little cringey but the internet is for people to share their interests. These people could have moved on without commenting death threats and patting themselves on the back for them shutting down their accounts. All I see in the insta comments are awful people commenting awful things, especially on videos about a woman or have a woman in them. They're filled with sexist people and it's making me mad. I blurred out their names including the account name for privacy."

Company policy and censorship. This category contains posts describing numerous issues with how certain apps resolve the moderation policies and censor inappropriate content. However, in many cases, these policies turned out to be biased based on gender and sexual orientation. The following example post from the "Pinterest" subreddit tells that the censorship for the app is biased toward images of gay people:

*"Why does this keep happening? For the past few months now I've been having pins removed left right and centre mainly for **"SEXUAL CONTENT AND NUDITY"** When I look at the images I find myself rather confused. I recently had an image of two men kissing removed, no nudity, no sexual content just two men kissing, a few months back I had another pin falsely flagged even though the pin was an image of two army men asleep on a train with some art underneath it of them waking up and one blushing as they make eye contact. Has this been happening to anyone else? I've contacted the Help team and currently waiting to see what they will say."*

In the following post, we observed an issue with one of the Microsoft products, where gender-related topics trigger spamming with hateful comments that are not moderated accordingly:

"Microsoft Start News Bots? I think Microsoft needs to do something about Bots on MSN. One person if it's a topic on LGBTQ or Abortion rights will just flood the comments, and when I mean flood, I mean the person will post a comment per minute, for around 24 hours and it's the same response all the time, which makes me think the person is a bot."

Content. The last category that we identified in our data was concerned with content-related issues. In such posts, users complained about the excessive amount of certain gender-related inappropriate content in their social media feeds or generally in the apps.

For instance, the following example post from the "Youtube" subreddit tells:

"Is there any way to stop YouTube constantly pushing right wing/hateful content on me in shorts? I am mostly watch baking, lgbt, funny/comedy and cute animal videos but recently any time I go on shorts it's one or two videos I like and then it's non stop anti trans/lgbt videos, racist videos, pro abortion videos, shit like that. It's crazy. Every video like it I tell YouTube not to recommend the channel anymore but I still see the same faces over and over pushing the same hateful shit. I don't interact with the videos in any other way. I don't look at the comments, I don't thumb them down, I just block and swipe. I used to really like the cozy vibes I had going in shorts and I want that back =/."

The following is an example post about inappropriate content in the Facebook app:

"Why is my new account being bombarded with sexual content even though I have not subscribed to any pages nor have I indicated any of my hobbies to associate to such content? Everytime I scroll down I keep getting recommended ass pictures galore? What's up with this Facebook. I indicated my interests are soccer and Carpentry. Wetin be this?"

6 Discussion

This chapter presents the findings of the research and offers an analysis of the implications of the research for further study. Additionally, it states the limitations of the study and lists the recommendations for further research.

6.1 Answering the research questions

The methodology described in Chapter 3 helped to answer the research question **RQ1**, *"Does Reddit data provide meaningful information about gender-related discussions in mobile applications?"*. The research started in Section 2 with an examination of the existing communities on the Reddit platform, specifically those dedicated to the top 50 Android mobile apps. It was observed that numerous popular subreddits existed where users engaged in discussions pertaining to the apps and provided feedback. The number of members in these communities ranged from one thousand to more than three million, yet this did not correlate with the frequency of their posts. In particular, the qualitative analysis of the raw Reddit data in subsection 3.3 revealed that while there was a small fraction of data that contained gender discussions, these were still valuable for software developers to consider when engineering the requirements for mobile applications. In summary, we can answer the research question RQ1 as follows: Yes, Reddit data provides meaningful information about gender-related discussions in mobile applications.

In response to the research question **RQ2**, *"Can we further categorize the gender-related discussions in Reddit posts?"*, our approach involved a manual examination of each gender-related post in subsection 3.3. Through this close examination, identifiable patterns emerged, suggesting the presence of distinct themes within these discussions that hold a promise for deeper analysis to understand the broader issues and develop potential solutions. Our research revealed six identifiable categories of discussions: Advertisement, AI, App Features, Company Policy and Censorship, Community, and Content. Examining these discussions yielded valuable insights, highlighting that the prevailing topics related to gender revolved around company policy and censorship, as well as biases inherent in AI-driven mechanisms. This discovery not only sheds light on the common concerns surrounding gender-related conversations but also provides pathways for further investigation of gender biases embedded in AI-driven solutions and their corresponding moderation rules. Overall, the research question RQ2 can be answered as follows: Yes, we can further categorize the gender-related discussions in Reddit posts.

To answer the research question **RQ3**, *"How do different ML models used for the automated classification of Reddit data for gender discussions in mobile applications compare with each other?"*, we evaluated the following four different ML and 2 DL classifiers: Support Vector Machine (SVM), Logistic Regression, Decision Tree, Gaussian Naive Bayes, DistilRoBERTa, and Llama 3. We used two widely used feature extraction

techniques, Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF), and compared these classifiers and feature extraction techniques in different configurations to determine their effectiveness.

We faced a considerable challenge in our exploration, caused by the unbalanced nature of the dataset, which impeded the achievement of sufficient results during training. In response, we employed the Synthetic Minority Over-sampling Technique (SMOTE) to overcome this problem by augmenting the training dataset and thereby improving model performance. Additionally, we expanded the gender-related class of the dataset by capturing all related comments alongside the Reddit posts to strengthen the representation of the minority class.

We found that among the classifiers we tested, Linear SVM performed the best, achieving a significant value of the F1 score of 68 percent for the minority class and 99 percent for the majority class in the configuration with TF-IDF and SMOTE. Logistic regression was closely in the second position, with the F1 scores of 0.62 and 0.99 for the gender-related and non gender-related classes, respectively.

By contrast, the Decision Tree and Gaussian Naive Bayes classifiers performed poorly, with the F1 scores of 0.39 and 0.17, respectively, demonstrating their low effectiveness, particularly in predicting the minority class. Being the weakest classifier in our experiments, Gaussian Naive Bayes's prediction results can be explained by the classifier's naive assumption of feature independence, which although it saves computational resources, fails to consider the intricate contextual relationships between features typical in linguistic data. In other words, the presence of keywords in the texts did not necessarily indicate the belonging to the target class.

The DistilRoBERTa classifier demonstrated results that were statistically significant, yet not as good as we expected with the evaluation results of the minority class being drastically lower than the predictions of the majority class. Compared to DistilRoBERTa, the Llama-3-8B classifier exhibited better prediction results in the minority class while maintaining strong results in the majority class.

These results highlighted the importance of employing robust techniques, such as SMOTE, and selecting appropriate classifiers to overcome the challenges posed by skewed datasets, ultimately improving the accuracy and reliability of gender classification in online discourse.

6.2 Main contributions and implications

One of the primary objectives of this research was to examine the potential value of Reddit data in the study of gender inclusivity violations in mobile apps. This data could be utilized as a source for software requirements. To validate the data, a comprehensive qualitative analysis of Reddit communities related to mobile apps was conducted. Each subreddit was reviewed and validated based on the number of members, the frequency of posts, and the relevance of the content to the topic of our research. Consequently,

122 subreddits were selected for analysis, and 97,345 Reddit posts were retrieved. The utilization of keywords for the prior filtering of the data served to narrow the analysis and focus only on those posts that were potentially gender-related. To facilitate more comprehensive data filtering, the KeyBERT keywords extraction method was employed as a keyword extraction tool, resulting in an expanded keyword list of 441 items and enhanced filtering accuracy. The keyword filtering approach indicated that 4.2 percent of the collected data potentially referenced gender-related topics. Given that Reddit is a social media platform and does not focus entirely on mobile applications' user feedback, it can be concluded that Reddit is a potentially valuable source for requirements related to gender inclusivity.

In order to gain an accurate understanding of gender relations, a quantitative and qualitative analysis of primary data from Reddit posts was conducted. This analysis resulted in a gender-related dataset, which consisted of two distinct classes, and included a total of 4,111 instances of data. By making this dataset publicly available [20], we aim to foster progress in the field of gender-inclusive software development, and thus contribute to the more comprehensive discussion of gender diversity and representation in digital spaces. In addition to mapping gender-related data, our research efforts included a thorough examination of the gender discourse within app reviews. Our comprehensive study enabled a more differentiated categorization of gender-related discussions, thereby providing a more detailed examination of issues of inclusivity, and identifying specific challenges inherent in software products.

Moreover, our study provided valuable insights into the predominant issues surrounding gender inclusivity, with notable findings highlighting instances of bias within recommendation algorithms and instances of skewed censorship practices within digital platforms. The qualitative analysis of the collected data revealed a dilemma regarding the approach to censorship of inappropriate and abusive content and freedom of speech in online communities. For example, one group of users expressed concern about an excessive amount of content related to certain themes as in the following post from the YouTube app subreddit:

Is there any way to stop YouTube constantly pushing right wing/hateful content on me in shorts? I am mostly watch baking, lgbt, funny/comedy and cute animal videos but recently any time I go on shorts it's one or two videos I like and then it's non stop anti trans/lgbt videos, racist videos, pro abortion videos, shit like that. It's crazy. Every video like it I tell YouTube not to recommend the channel anymore but I still see the same faces over and over pushing the same hateful shit. I don't interact with the videos in any other way. I don't look at the comments, I don't thumb them down, I just block and swipe. I used to really like the cozy vibes I had going in shorts and I want that back =/.

At the same time, others argued that blocking such content may potentially harm the representation of those gender groups, as was expressed in this post from the subreddit

[20]https://github.com/Dariya-Nagashi/master-thesis-ut

for the Bing AI chatbot:

"I've also had content restriction after content restriction for the following:

A female zombie, black and white photoshoot Female zombie, black and white photo op Zombie lady, posing for camera Zombie lady

Literally just trying to make a framed picture for my video game. I search for Male zombie or "Zombie" in general and it's fine.

I don't know what Bing are doing but immediately putting up content warning for innocuous content featuring women is borderline erasure of female identities, if not just really freaking weird."

These revelations highlight not only the critical need for preventative measures to address such biases but also the vital role of research in exposing and tackling systemic disparities within technological environments.

As a final step of our study, the experiments with automated classification have been conducted using the dataset that has been constructed, and the performance of each configuration was analyzed. By classifying gender-related discussions within Reddit posts using machine learning techniques, our research contributes to a deeper understanding of the delicate nature of gender discourse in mobile applications. Our experiments with different machine learning classifiers reveal the effectiveness of different models in precisely classifying gender-related content. By identifying the Linear SVM classifier as the most proficient in our task, we provide useful guidance to researchers and practitioners seeking optimal approaches to automated classification of gender-related issues in software products. The overall training of the models revealed the importance of the balanced dataset to obtain precise prediction results in both classes.

6.3 Limitations and future work

This subsection is dedicated to the potential limitations of this thesis and the recommended steps to mitigate them. One of the possible limitations is related to the unanticipated obstacles during the data collection process, which was constrained by the Reddit API, which limits the maximum number of posts to be retrieved from a subreddit to 1000. This may have resulted in an insufficient amount of data, potentially leading to inaccurate or incomplete insights into gender discussions in Android apps. To mitigate this risk, we extended the list of subreddits from one per app to the top three associated with the app. This significantly increased the dataset size. As a recommendation for future work, data from third-party provider services might be purchased to utilize larger volumes of data.

Another possible limitation is that the dataset was manually annotated. The cultural background and beliefs of annotators may have influenced their perception of gender, potentially leading to biased annotations. To address this issue, two annotators with diverse backgrounds, genders, and sexual orientations were engaged to annotate and discuss the data from their perspectives. Furthermore, inter-rater reliability was assessed using Cohen's kappa coefficient. Another potential risk is inconsistency and annotation

fatigue, which may compromise the accuracy and overall quality of data labeling. To avoid these potential issues, we divided the process into four iterations and agreed upon the criteria in advance. Nevertheless, as the future improvement steps, we propose the involvement of additional raters with diverse backgrounds to mitigate sociocultural bias and enhance the accuracy of labeling.

A third potential limitation pertains to the automated classification of gender-related data and involves model training on an imbalanced dataset, which may introduce bias into the prediction of the minority class. As this thesis does not address the issue of data imbalance, further investigation on that issue was not conducted. However, in order to support our experiments with automated classification, we applied the oversampling technique SMOTE. In addition, other random sampling techniques can be employed, and cost-effective machine learning models can be trained.

The fourth limitation might be concerned with the relatively small ratio of 4.2 percent of potentially gender-related Reddit posts achieved by keywords filtration and 0.12 percent of posts labeled as gender-related during the manual annotation. However, other researchers such as [43] have achieved a comparable 0.07 percent of potentially gender-related reviews out of the total number of retrieved App Store reviews. Furthermore, it is important to note that the noise level is higher in Reddit posts due to the nature of the platform as a social networking service.

Nonetheless, the results obtained in this thesis are valid and have led to the answers to our research questions. One possible future direction for research could be the exploration of categories of gender-related discussions and the preparation of a multi-label gender dataset. This thesis proposed a possible categorization of topics and issues in gender-related discussions, which could serve as a foundation for future research in this area.

7 Conclusions

The main question that this thesis aimed to address was the possibility of using Reddit data to understand the gender discourse in modern mobile applications. Based on the quantitative and qualitative analysis of the collected data, it can be concluded that Reddit data provides valuable information for investigating gender-related issues in mobile applications and can be classified and categorized into separate groups of gender inclusivity violations in software products. The results of our research provide a solid basis for engineering functional, quality, and emotional requirements for software products aiming to design more inclusive applications for end users.

Based on a thorough analysis of previous research on similar topics, we chose the direction of our study and utilized a methodology that included data collection and manual annotation of the data with prior keyword extraction and filtering. The choice of this particular methodology was guided by the absence of a gender-related dataset based on Reddit data. Moreover, utilizing the keywords extraction and filtering was expected to increase the speed of the manual dataset creation by narrowing down the potential data search and finding more nuanced insights into the gender discussion topic. The objective of this study was to identify insightful information related to gender inclusivity in the features of mobile applications and to prepare a dataset to be used to train machine learning models for future automated classification of gender-related feedback. The results of our study successfully met our expectations and unveiled new insights into the current trends and gender-related issues with software products. This thesis clearly illustrates the urgency of gender inclusiveness in mobile applications and the relevance of such discussions in the Reddit communities. However, it also raises the issue of challenges of precise automated detection of gender discussions due to the subtlety of the topic. During the attempt to provide an automated classification of gender-related discussions, impediments such as a skewed class distribution of data were met. As a suggestion to tackle this problem, practitioners should consider the sparse nature of Reddit data and mine more Reddit posts to broaden the dataset for model training. Furthermore, the prediction results can be potentially improved by fine-tuning the applied models and investigating the state-of-the-art deep learning classifiers.

To better understand the implications of the qualitative analysis of this thesis, future studies could address the taxonomy of gender issues in the context of requirements engineering. The categorization proposed in this work can be utilized as a starting point for this kind of research. As another direction of study, potential mitigation strategies can be analyzed to address the issues related to gender-inclusive mobile applications that have been discovered in this thesis.

References

[1] Introducing meta llama 3: The most capable openly available llm to date. url:https://ai.meta.com/blog/meta-llama-3/ Accessed: 01.05.2024.

[2] Margaret Burnett, Simone Stumpf, Jamie Macbeth, Stephann Makri, Laura Beckwith, Irwin Kwan, Anicia Peters, and William Jernigan. GenderMag: A Method for Evaluating Software's Gender Inclusiveness. *Interacting with Computers*, 28(6):760–787, 10 2016.

[3] Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. Yake! keyword extraction from single documents using multiple local features. *Information Sciences*, 509:257–289, 2020.

[4] Edna Dias Canedo and Bruno Cordeiro Mendes. Software requirements classification using machine learning algorithms. *Entropy*, 22(9):1057, 9 2020.

[5] Jaime Carbonell and Jade Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, page 335–336, New York, NY, USA, 1998. Association for Computing Machinery.

[6] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

[7] Won Ik Cho, Jiwon Kim, Jaeyeong Yang, and Nam Soo Kim. Towards Cross-Lingual Generalization of Translation Gender Bias. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 449–457, New York, NY, USA, 2021. Association for Computing Machinery.

[8] Michael Christie, Maureen O'Neill, Kerry Rutter, Graham Young, and A. Medland. Understanding why women are under-represented in Science, Technology, Engineering and Mathematics (STEM) within Higher Education: a regional case study. *Production*, 27(spe), 1 2017.

[9] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 9 1995.

[10] Stephan Dreiseitl and Lucila Ohno-Machado. Logistic regression and artificial neural network classification models: a methodology review. *Journal of biomedical informatics*, 35(5-6):352–359, 10 2002.

[11] Mattia Fazzini, Hourieh Khalajzadeh, Omar Haggag, Zhaoqing Li, Humphrey Obie, Chetan Arora, Waqar Hussain, and John Grundy. Characterizing human aspects in reviews of COVID-19 apps. In *Proceedings of the 9th IEEE/ACM International Conference on Mobile Software Engineering and Systems*, MOBILESoft '22, page 38–49, New York, NY, USA, 2022. Association for Computing Machinery.

[12] B. Glaser and A. Strauss. *The discovery of grounded theory*. New York: Routledge, 2017.

[13] Maarten Grootendorst. KeyBERT: Minimal keyword extraction with BERT, 2021. `https://doi.org/10.5281/zenodo.4461265`. Accessed: 2023.10.15.

[14] Emitza Guzman, Mohamed Ibrahim, and Martin Glinz. A little bird told me: Mining tweets for requirements and software evolution. In *IEEE 25th International Requirements Engineering Conference (RE)*, pages 11–20, 2017.

[15] James A. Hanley and Barbara J. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36, 4 1982.

[16] C. Hutto and Eric Gilbert. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1):216–225, 2014.

[17] Tahira Iqbal, Moniba Khan, Kuldar Taveter, and Norbert Seyff. Mining reddit as a new source for software requirements. In *2021 IEEE 29th International Requirements Engineering Conference (RE)*, pages 128–138, 2021.

[18] Timo Johann, Christoph Stanik, Alireza M. Alizadeh B., and Walid Maalej. SAFE: A Simple Approach for Feature Extraction from App Descriptions and App Reviews. In *IEEE 25th International Requirements Engineering Conference (RE)*, pages 21–30, 2017.

[19] Zeng jun Bi, Yao quan Han, Cai quan Huang, and Min Wang. Gaussian naive bayesian data classification model based on clustering algorithm. In *Proceedings of the 2019 International Conference on Modeling, Analysis, Simulation Technologies and Applications (MASTA 2019)*, pages 396–400. Atlantis Press, 2019/07.

[20] Yekaterina Kovaleva, Ari Happonen, and Eneli Kindsiko. Designing gender-neutral software engineering program. stereotypes, social pressure, and current attitudes based on recent studies. In *Proceedings of the Third Workshop on Gender Equality, Diversity, and Inclusion in Software Engineering*, GE@ICSE '22, page 43–50, New York, NY, USA, 2022. Association for Computing Machinery.

[21] David D. Lewis. Naive (Bayes) at forty: The independence assumption in information retrieval. In *Machine Learning: ECML-98*, pages 4–15, Berlin, Heidelberg, 1998. Springer Berlin Heidelberg.

[22] Conghui Li, Humphrey O. Obie, and Hourieh Khalajzadeh. A First Step Towards Detecting Values-violating Defects in Android APIs, 2021. arXiv:2109.14359 [cs.SE].

[23] Walid Maalej and Hadeer Nabil. Bug Report, Feature Request, or Simply Praise? On Automatically Classifying App Reviews. In *IEEE 23rd International Requirements Engineering Conference (RE)*, pages 116–125, 2015.

[24] Collins Mathews., Kenny Ye., Jake Grozdanovski., Marcus Marinelli., Kai Zhong., Hourieh Khalajzadeh., Humphrey Obie., and John Grundy. AH-CID: A Tool to Automatically Detect Human-Centric Issues in App Reviews. In *Proceedings of the 16th International Conference on Software Technologies - ICSOFT*, pages 386–397. INSTICC, SciTePress, 2021.

[25] Mary L. McHugh. Interrater reliability: the kappa statistic. biochemia medica, 22(3), 276–282., 2012. `https://pubmed.ncbi.nlm.nih.gov/23092060/`. Accessed: 2024.03.10.

[26] Christopher Mendez, Lara Letaw, Margaret Burnett, Simone Stumpf, Anita Sarma, and Claudia Hilderbrand. From gendermag to inclusivemag: An inclusive design meta-method. In *IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, pages 97–106, 2019.

[27] Charles E. Metz. Basic principles of ROC analysis. *Seminars in nuclear medicine*, 8(4):283–298, 10 1978.

[28] Ali Rezaei Nasab, Maedeh Dashti, Mojtaba Shahin, Mansooreh Zahedi, Hourieh Khalajzadeh, Chetan Arora, and Peng Liang. A study of fairness concerns in AI-based mobile app reviews. *arXiv (Cornell University)*, 1 2024.

[29] Maleknaz Nayebi, Henry Cho, and Guenther Ruhe. App store mining is not enough for app improvement. *Empirical software engineering*, 23(5):2764–2794, 2 2018.

[30] Humphrey O. Obie, Waqar Hussain, Xin Xia, John Grundy, Li Li, Burak Turhan, Jon Whittle, and Mojtaba Shahin. A First Look at Human Values-Violation in App Reviews. In *IEEE/ACM 43rd International Conference on Software Engineering: Software Engineering in Society (ICSE-SEIS)*, pages 29–38, 2021.

[31] J. Thomas Parsons, Michael Schrider, Oyebanjo Ogunlela, and Sepideh Ghanavati. Understanding developers privacy concerns through Reddit thread analysis. *arXiv (Cornell University)*, 1 2023.

[32] David M. W. Powers. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation, 2020. arXiv:2010.16061 [cs.LG].

[33] Priyanka and Dharmender Kumar. Decision tree classifier: a detailed survey. *Int. J. Inf. Decis. Sci.*, 12:246–269, 2020.

[34] Faisal Rahutomo, Teruaki Kitasuka, and Masayoshi Aritsugi. Semantic cosine similarity. volume 4, page 1, 10 2012.

[35] Juan Ramos. Using tf-idf to determine word relevance in document queries, 2003. `https://api.semanticscholar.org/CorpusID:14638345`. Accessed: 2024.03.10.

[36] Martin P. Robillard, Andrian Marcus, Christoph Treude, Gabriele Bavota, Oscar Chaparro, Neil Ernst, Marco Aurélio Gerosa, Michael Godfrey, Michele Lanza, Mario Linares-Vásquez, Gail C. Murphy, Laura Moreno, David Shepherd, and Edmund Wong. On-demand Developer Documentation. In *IEEE International Conference on Software Maintenance and Evolution (ICSME)*, pages 479–483, 2017.

[37] Gema Rodríguez-Pérez, Reza Nadri, and Meiyappan Nagappan. Perceived diversity in software engineering: a systematic literature review. *Empirical software engineering*, 26(5), 7 2021.

[38] Francisco Rodríguez-Sánchez, Jorge Carrillo-De-Albornoz, Laura Plaza, Julio Gonzalo, Paolo Rosso, Miriam Comet, and Trinidad Donoso. Overview of EXIST 2021: SEXism Identification in Social NETworks, 9 2021.

[39] Stuart Rose, Dave Engel, Nick Cramer, and Wendy Cowley. *Automatic Keyword Extraction from Individual Documents*, pages 1 – 20. 03 2010.

[40] Mattia Samory, Indira Sen, Julian Kohne, Fabian Floeck, and Claudia Wagner. "Call me sexist, but...": Revisiting Sexism Detection Using Psychological Scales and Adversarial Samples. *arXiv (Cornell University)*, 1 2020.

[41] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108, 2019.

[42] Shalom H. Schwartz. An overview of the Schwartz Theory of basic values. *Online readings in psychology and culture*, 2(1), 12 2012.

[43] Mojtaba Shahin, Mansooreh Zahedi, Hourieh Khalajzadeh, and Ali Rezaei Nasab. A study of gender discussions in mobile apps. *arXiv (Cornell University)*, 1 2023.

[44] Sunny Shrestha and Sanchari Das. Exploring gender biases in ml and ai academic research through systematic literature review. *Frontiers in Artificial Intelligence*, 5, 2022.

[45] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

[46] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023.

[47] Xinyun Wang and Hongyun Ning. Tf-idf keyword extraction method combining context and semantic classification. In *Proceedings of the 3rd International Conference on Data Science and Information Technology*, page 123–128, New York, NY, USA, 2020. Association for Computing Machinery.

[48] Grant Williams and Anas Mahmoud. Mining Twitter Feeds for Software User Requirements. In *2017 IEEE 25th International Requirements Engineering Conference (RE)*, pages 1–10, 2017.

[49] Xin Xia, David Lo, Xinyu Wang, and Bo Zhou. Tag recommendation in software information sites. In *2013 10th Working Conference on Mining Software Repositories (MSR)*, pages 287–296, 2013.

[50] Yin Zhang, Rong Jin, and Zhi Zhou. Understanding bag-of-words model: a statistical framework. *International journal of machine learning and cybernetics*, 1(1-4):43–52, 8 2010.

Appendix

I. Code

The source code of this thesis, the resulting dataset as well as the other supporting documents are publicly available at https://github.com/Dariya-Nagashi/master-thesis-ut.

II. Licence

Non-exclusive licence to reproduce thesis and make thesis public

I, **Dariya Nagashibayeva**,

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,

   Understanding Gender Related Discussions in Android Mobile Applications Through Reddit,

   supervised by Tahira Iqbal and Kuldar Taveter.

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.

Dariya Nagashibayeva
15/05/2024