2023 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)

Auditing Gender Analyzers on Text Data

Siddharth D Jaiswal
Dept. of CSE
IIT Kharagpur, India
siddsjaiswal@kgpian.iitkgp.ac.in

Ankit Verma
Dept. of CSE
IIT Kharagpur, India
ankitverma@kgpian.iitkgp.ac.in

Animesh Mukherjee
Dept. of CSE
IIT Kharagpur, India
animeshm@cse.iitkgp.ac.in

***Abstract*—AI models have become extremely popular and accessible to the general public. However, they are continuously under the scanner due to their demonstrable biases toward various sections of the society like people of color and non-binary people. In this study, we audit three existing gender analyzers – uClassify, Readable and HackerFactor, for biases against non-binary individuals. These tools are designed to predict only the cisgender binary labels, which leads to discrimination against non-binary members of the society. We curate two datasets – Reddit comments (660k) and, Tumblr posts (2.05M) and our experimental evaluation shows that the tools are highly inaccurate with the overall accuracy being $\approx 50\%$ on all platforms. Predictions for non-binary comments on all platforms are mostly *female*, thus propagating the societal bias that non-binary individuals are effeminate. To address this, we fine-tune a BERT multi-label classifier on the two datasets in multiple combinations, observe an overall performance of $\approx 77\%$ on the most realistically deployable setting and a surprisingly higher performance of 90% for the non-binary class. We also audit ChatGPT using zero-shot prompts on a small dataset (due to high pricing) and observe an average accuracy of 58% for Reddit and Tumblr combined (with overall better results for Reddit).**

Thus, we show that existing systems, including highly advanced ones like ChatGPT are biased, and need better audits and moderation and, that such societal biases can be addressed and alleviated through simple off-the-shelf models like BERT trained on more gender inclusive datasets.

***Index Terms*—Gender Analyzer, bias, social media**

I. Introduction

Gender, a characteristic defined by the socio-cultural structures that a person lives in, plays a huge role in how an individual is perceived in society and the kind of societal facilities that are made available to them. While there is a growing acceptance of the non-binary gender [1] and acknowledgement that it is different from the sex assigned at birth, most societies still treat gender as a binary attribute with only two identities- 'male' and 'female' [2][1]. This prejudiced practice that has been continuing for thousands of years has discriminated against various non-binary genders, and denied them access to basic public facilities like public toilets [3], [4], employment opportunities [5] and civil rights, leading to legal discrimination [6].

Platform	Comments	UC	RD	HF	BT
Reddit	You're allowed to use he/him AND they/them ... I use she/her and he/him but not they/them ... You're allowed to present however you want but still identify as genderqueer, genderfluid, non-binary, etc. ...	F	F	WF	NB
Tumblr	"There is no queer community" is not and will never be a true statement. I am queer and I am part of a queer community. We are real, we exist in real life, we are here. ... Because that is what you are.	F	F	WF	NB

TABLE I: Some example comments by non-binary authors from Reddit and Tumblr dataset and the predicted labels by the gender analyzers. M: male, F: female, N: neutral, NB: non-binary, WF: weak female. UC: uClassify, RD: Readable, HF: HackerFactor, BT: BERT.

Impact of discrimination: Societal discrimination can be dealt with through legal recourse, but overcoming these issues is arduous when there are non-human elements in the loop, like AI-based software, which are often *unexplainable* black boxes. Significant research has been done on discrimination against cisgender individuals in tasks like face recognition [7], [8], automated hiring [9], and image search [10], but there has been a very limited research [11]–[13] investigating discrimination against non-binary individuals. A possible reason is the non-availability of 'non-binary' as a gender label on many AI services like face recognition softwares [14]–[17] and gender analyzers from text [18]–[20]. It becomes difficult to study the outputs for non-binary individuals under real deployment scenarios and erodes their existence from societal conscience. Questions can be raised on the morality of having a non-binary gender label but as these services are used at scale and deployed around the world, their impact on non-binary individuals is high. For example, a non-binary individual may be incorrectly classified as a male or female by a face recognition service and thus may be denied access to public toilets.

Audit of gender analyzers: In this work, we audit tools that predict the gender of the author from a piece of text. These are used for safety on social media [21] and, advertising [22], but only work for the two binary genders - female and male, thus omitting non-binary authors. Amazon's resume shortlister was scrapped [23] as it was using gender-indicative words to shortlist only male candidates. A more acute problem is when such analyzers are used for law enforcement operations like crime detection. AI tools are being used to determine user gender and link it with the content produced by them to

[1] In this work, we refer to cisgender men and women as male and female, respectively.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ASONAM '23, November 6-9, 2023, Kusadasi, Turkey

© 2023 Copyright is held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 979-8-4007-0409-3/23/11…$15.00
http://doi.org/10.1145/3625007.3627324

understand how such content is expressed to aid crime detection and investigations. In cases of cybercrime, for instance, the identification of the gender from the content posted and linking these to the user's identity can narrow down the number of potential suspects [24]. However, since such AI tools are not trained to predict the non-binary class, such users are always going to be mispredicted thus placing them in a very risky position. Thus gender analyzers are important AI tools but come with their share of discrimination. Economical pricing, free-tier subscriptions, and easy to use web-interfaces have led to wide-scale adoption of these tools for both personal and professional use. One such tool, READABLE, for instance, has Shopify, Netflix, Adobe, and NASA as clients. Here we audit [25] three very common existing tools that predict the gender of the author from a piece of text: two commercial tools – UCLASSIFY [18] and READABLE [19] and a free open-source tool – HACKERFACTOR [20]. For a small subset of textual inputs, we also audit ChatGPT [26], an LLM that generates human like text based on input prompts.

Designing a fair gender analyzer: Next, we finetune a pretrained BERT [27] based multi-label classifier which predicts three gender labels – 'male', 'female', and 'non-binary'. Our experiments using this off-the-shelf model show that while there are *simple* ways to address existing societal biases, ML developers remain uninformed of these possible solutions.

In this study, we curate two datasets from the Reddit [28] (660k comments) and Tumblr [29] (2.05M posts) platforms, from various self-identified male, female and non-binary individuals (more details in Section III). The gender label for all posts on Tumblr and more than 20% comments on Reddit are self-annotated by the author of the text. The rest have been annotated by two annotators based on the subreddits the post is collected from[2]. Some example comments by non-binary authors along with their associated predictions by the different platforms are available in Table I. The columns are labeled in the following order – UCLASSIFY, READABLE, HACKERFACTOR, and our BERT model. We see that the BERT-based model correctly predicts the labels whereas the other platforms misclassify the statements. Note that since the commercial classification models are black boxes and do not release the weights and training data, their outputs are not explainable. HACKERFACTOR, though open-source, has fixed weights assigned to various *gender-centric* words, which is used in a fixed formula to classify a given sentence. Finally, this investigation also provides us an opportunity to study various forms of machine learning biases [30] like *representation, user interaction* and *emergent bias*.

Research questions: Here, we state the key research questions that we address as part of this study.

RQ1: What are the gender predictions by the four platforms - UCLASSIFY, READABLE, HACKERFACTOR and BERT, for the comments from the male and female gender groups and whether the accuracy of the three existing tools is as high as they claim and how the finetuned BERT model performs in comparison to them. Through this question, we seek to verify if the tools are indeed accurate in identifying the gender of the binary authors through various standard measures like precision, recall, and F1 score.

RQ2: What is the predicted gender for the non-binary comments on all the platforms? Through this question, we attempt to gain insight into how these platforms perceive text written by non-binary authors and what kind of social implications this might have. Since the BERT model is trained to predict non-binary labels too, the accuracy values will give us an indication of how a gender-fair model may be designed and deployed.

RQ3: How do the audited platforms and more importantly, the BERT classifier, perform on the comments from two distinctly different social media platforms – Reddit and Tumblr? Through this question, we attempt to understand the generalization capability of the BERT model for different textual datasets as well as compare the performance of the audited platforms on the two datasets.

RQ4: How do the advanced LLMs like CHATGPT perform when suitably prompted to classify gender based on the user posts.

Our contributions: In this paper, we audit three existing gender analyzers that classify the gender of the author based on input text and also fine-tune and test a BERT-based multilabel classifier for Reddit and Tumblr comments from binary and non-binary authors. The existing systems – UCLASSIFY, READABLE, HACKERFACTOR report an accuracy of $\approx 50\%$ on both Reddit and Tumblr for the binary authors' texts, lower than the minimum 70% claimed by both READABLE and HACKERFACTOR. One of our BERT-based fine-tuned models has the highest overall accuracy of 83% on Reddit and 66% on Tumblr. For the non-binary authors' comments, on Tumblr, all the existing systems – UCLASSIFY, READABLE and HACKERFACTOR predict female as the author's gender for a majority of the comments, whereas on Reddit, only UCLASSIFY and READABLE report similar observations. Finally, in a transfer learning setup, our *best* BERT-based model (fine-tuned on one dataset and then provided with few-shot examples from another dataset) has an accuracy of 90% for non-binary authors, with an overall accuracy of 77%, which demonstrates that even a simple off-the-shelf model can be accurately used to design a gender-inclusive system. Surprisingly, when we prompt CHATGPT in a zero-shot setting to perform gender classification based on the user comments from across the two datasets, the average accuracy is only 58% with overall results being better for Reddit. This indicates that even such a powerful model is not very suitable as a general purpose gender analyzer.

II. RELATED WORK

On social media platforms, the task of gender identification is used for the purpose of security [31], advertising [22], online safety [21] and opinion mining amongst others. Due to the different features of these social media websites, the task of gender identification is done based on the analysis of various

[2] Data available on request– https://forms.gle/VVSrTxoYKPW16vCYA

user generated content like visual posts (display pictures [32]), text (blog posts, comments, tweets [33]–[38]) etc. In this work, we focus on auditing commercial and open-source tools that are used for author gender identification through textual data on Reddit and Tumblr.

The early works in author gender identification were by Cheng et al. [21] to identify gender in short, content-free text, and, by Deitrick et al. [39] who used stylometric and word count features to classify email authors. Bartle and Zheng [40] proposed a Windowed RCNN (WRCNN) that achieved an accuracy of 86% on a blog dataset. Mukherjee and Bala [22] compared different machine learning models against commercial softwares for the binary gender task prediction and achieved a higher accuracy than the baseline. Vicente et al. [32] used textual analysis of English and Portuguese tweets as part of their larger gender identification pipeline. Fatemeh et al. [31] used an ANN based classifier coupled with whale optimization to identify the gender of email authors and report an accuracy of 98%. Recently, there has been significant work in developing gender analyzers for English text [33]–[35]. Vasilev's [41] thesis deals with inferring gender of Reddit users, but like all existing works in this domain, including ones described above, it treats gender as a binary attribute.

It is apparent that there has been considerable research in this domain but none of them have focused on addressing the problem for non-binary gender groups, thereby allowing the existing discrimination to fester. We also notice that while all of the previous works have studied design and development of open-source models or compared against commercial products, none of them have focused on auditing any of these models. We address both these problems in this paper by auditing commercial and open-source third-party softwares for multiple gender groups, and fine-tuning a pretrained model that is able to make predictions for the non-binary gender, thus presenting a gender-inclusive alternative to the existing softwares.

III. Dataset & Methodology

A. Dataset curation

In this study, we create social media comments datasets from two platforms– Reddit and Tumblr, in the English language.

- Reddit: Top-level comments are collected from various subreddits belonging to male, female and non-binary interests. We assign a subreddit as **X**-interest (where **X** is either male, female or non-binary) if the subreddit name or description refers to any of the corresponding gender categories.
- Tumblr: Blogs (unique per user) are collected and segregated based on the presence of the gender label in the blog bio. This is followed by the collection of self-posts and answers given by the bloggers, per blog. Each such post or answer (referred to as comment henceforth) is assigned an **X**-gender (where **X** is either male, female, or non-binary) based on the pronoun or gender mentioned by the blogger in their profile description.

More details from the two datasets are present in Table II.

	Reddit						Tumblr				
G	**#S**	**#C**	L_{avg}	# A_{unq}	# P_{unq}	# $(C/P)_\mu$	**#B**	**#C**	L_μ	# A_{unq}	# $(C/B)_\mu$
M	3	240k	99	62k	33k	7	230	343k	99	230	1491
F	2	240k	99	67k	29k	8	688	704k	141	688	1023
NB	7	180k	80	44k	65k	3	1670	1.01M	132	1670	605
T	12	660k	93	173k	127k	6	2588	2.05M	124	2588	1040

TABLE II: Details of the subreddits/blogs, comments, authors and posts in the Reddit & Tumblr comment dataset. On Tumblr, each blog post is available with a unique UUID, without an author username, hence the number of unique authors is the same as the number of unique blogs. For Reddit, #S is subreddit count and #P_{unq} is count of unique posts. On both platforms, #C is comment count, L_μ is the avg. length of a comment, # A_{unq} is count of unique authors and # $(C/P)_\mu$ ($(C/B)_\mu$) is the avg. number of comments per post (blog). B: blog, M: male, F: female, NB: non-binary.

Gender annotation strategy: For Reddit, over 20% of the data is self-annotated for gender by the commenters themselves. We annotate the rest as either male, female, or non-binary based on the subreddit it is collected from. We believe our annotation strategy is effective because all of these subreddits serve the interests of the corresponding gender group exclusively and thus the posts and comments are only by individuals who identify with that particular gender[3]. On Tumblr, all posts are self annotated by the bloggers.

Going by this annotation scheme, we note a few more interesting details about the dataset in Table II. The Reddit dataset contains approximately 173k unique authors, of which male and female comments are made by about $\approx 65k$ unique authors each and non-binary comments are by $\approx 44k$ unique authors. Female and male subreddits have about 32k unique posts whereas non-binary subreddits have $\approx 65k$ unique posts. Finally, male and female posts have an average of 8 comments per post where as non-binary posts have an average of 3 comments per post. Similarly for Tumblr, the dataset contains 2588 unique bloggers, of which there are 230 male, 688 female, and 1670 non-binary bloggers. Male and female blogs have more than 1K posts per blog whereas non-binary blogs have ≈ 600 posts per blog.

B. Methodology

Post collection: We collect the posts/blogs from both platforms using Pushshift IO[4] and Tumblr API[5] respectively. The data for Reddit is collected for posts between Jan 2017 and Aug 2022, and for Tumblr, blogs posted between May 2008 and Nov 2022. On Reddit, we collect only the top-level comments for each post. Similarly, for Tumblr, we collect all the posts by the blogs filtered for the gender group. The three platforms are now described in brief, followed by a short description of the BERT based model.

uClassify: This is a ML web service that provides over 120 publicly available text-based classifiers that perform tasks like sentiment, gender, age and even Myer-Briggs analysis for a given piece of text. The service is accessible through a web interface, REST APIs and SDKs with multiple programming language bindings and for multiple languages like English, Spanish, French and Swedish.

[3] https://tinyurl.com/4mkkud2j

[4] https://api.pushshift.io

[5] tumblr.com/docs/en/api/v2

READABLE: This is a platform that scores textual data to improve readability and utilizes multiple formulae like Gunning Fog index and SMOG index to evaluate the input text. The set of free classifiers available on the website can be used for gender analysis, profanity and buzzword detection, etc. The service is accessible either through a web interface or REST API for the English language.

HACKERFACTOR: This platform offers an open source software called Gender Guesser that determines an author's gender by using vocabulary statistics. While the software is not ML based, it is highly popular and freely available to use through a web interface. Users may also copy the code and run it locally.

BERT base model: We finetune a pretrained BERT base-uncased model to predict three classes – male, female and non-binary. This is a transformer model that has been pretrained on a large corpus of English data. We finetune the model for multiple settings on both datasets. On both datasets, the train-validation-test split of the comments per gender group is $70 : 10 : 20$. The batch size was set to 64, learning rate to $2.0e^{-5}$, seed value to 4, dropout to 0.1 and optimizer to ADAM.

CHATGPT: This is a large language model that generates human-like text based on prompts. We prepare appropriate prompts that allow it to simulate a text-based gender analyzer and predict the gender for the input text's author.

C. Steps for experimentation

The steps for experimentation on the three platforms are described in brief as follows -

- UCLASSIFY: A Python script issues POST requests to the **genderanalyzer_v5** REST API endpoint for every comment and stores the responses (either male or female), which are then finally analyzed against the ground truth data to calculate the metrics.
- READABLE: The comments are submitted to the web interface of the *gender analyzer* tool through Selenium web automation which also collects the responses (either male, female or neutral) and finally calculates the accuracy metrics.
- HACKERFACTOR: All the comments are processed locally by the freely available source code downloaded from the Gender Guesser tool of HACKERFACTOR. The code evaluates the text for both formal and informal types, and we capture the predicted gender for the informal type.
- CHATGPT: Due to budget constraints, only 1500 comments (500 from each gender group) per platform are audited here. We generate appropriate prompts that simulate a text-based gender analyzer.

For the BERT base model, we experiment with multiple finetuning settings using the datasets from Reddit ($\mathbb{R}$) and Tumblr ($\mathbb{T}$) to address our research questions from Section I.

- Finetuning with comments from $\mathbb{R}$ and $\mathbb{T}$ individually.
- Finetuning by combining equal percentage of comments from $\mathbb{R}$ and $\mathbb{T}$.
- Finetuning first with comments from one platform and then again finetuning for a second time in a few-shot setting using comments from the other platform.

Platform	Training data	Claimed Acc.	Predicted classes
UCLASSIFY	Blogs	N.A.	M, F
READABLE	N.A	70%	M, F, N
HACKERFACTOR	N.A.	70%	M, F, UNK, WM, WF, WU
BERT base	WIKI & book-corpus	N.A.	M, F, NB

TABLE III: Training dataset and claimed accuracy for the four platforms. The BERT model is fine-tuned on the two comment datasets for evaluation. M: male, F: female, N: neutral, NB: non-binary, UNK: unknown, WM: weak male, WF: weak female, WU: weak unknown.

IV. OBSERVATIONS & RESULTS

A. Performance on male and female comments (RQ1, RQ3)

We first discuss the results for the three baseline platforms –UCLASSIFY, READABLE and HACKERFACTOR for the cis-gender binary comments. UCLASSIFY is trained on a balanced dataset of 11,000 blogs (divided equally between males and females), but no baseline accuracy is provided. READABLE and HACKERFACTOR platforms mention a baseline accuracy of 70% for the gender analyzer but do not provide any information on the training data or previous evaluation metrics. More information regarding the classes predicted by each platform is present in Table III.

Observations. The table in Figure 1 presents the results for the accuracy of prediction for comments by male and female individuals on the REDDIT (480k comments) and TUMBLR datasets (1.04M comments). READABLE reports the lowest accuracy for both datasets – 47.8% and 41.6% respectively. The best performing platform for REDDIT is HACKERFACTOR ($\approx 57\%$) and for TUMBLR is UCLASSIFY (52%). The highest precision on both datasets is reported by UCLASSIFY – 63% and 54.3% respectively. For recall and F1, we note that UCLASSIFY is the best for TUMBLR and HACKERFACTOR is the best for REDDIT.

The accuracy for male comments is lower than for females on all platforms. The lowest value is reported for UCLASSIFY – 13.4% and 15.5% respectively and the highest on HACKERFACTOR – 55% and 43% respectively. Thus, it is clear that the existing platforms are not very accurate in predicting the gender of comments by male authors. Similar observations can be made regarding the recall and F1. Thus, the existing models are overcompensating and misclassifying the male comments as female. It is difficult to know whether this is due to the training data, the weights of the gendered words, or the model inference mechanism. Interestingly, the precision for males is higher than for females for Reddit comments. There is a $> 70\%$ gap between the female and male accuracies on UCLASSIFY, independent of the dataset while this difference is less stark on the other platforms. The accuracy for the non-binary gender group is not shown as none of these platforms predict a non-binary gender.

B. Performance on non-binary comments (RQ2, RQ3)

Next, we study the responses of the three platforms for the non-binary dataset. Since these platforms do not predict a non-

Platform	Gender	Accuracy		Precision		Recall		F1	
		$\mathbb{R}$	$\mathbb{T}$	$\mathbb{R}$	$\mathbb{T}$	$\mathbb{R}$	$\mathbb{T}$	$\mathbb{R}$	$\mathbb{T}$
UCLF	Overall	54.2%	52.2%	63%	54.3%	54.2%	52.2%	45.2%	49.8%
	Male	13.4%	15.5%	73.5%	40.3%	13.4%	15.5%	22.8%	22.4%
	Female	95.1%	88.8%	52.4%	68.3%	95.1%	88.8%	67.6%	77.2%
RDBL	Overall	47.8%	41.6%	56.7%	50%	47.8%	41.6%	51.5%	44.8%
	Male	41.0%	36.9%	57.5%	32.9%	41.0%	36.9%	47.8%	34.8%
	Female	54.7%	46.2%	55.9%	67.1%	54.7%	46.2%	55.3%	54.7%
HCFTR	Overall	56.8%	48.6%	57.6%	49.4%	56.8%	48.6%	57.2%	48.3%
	Male	55.2%	42.8%	57.9%	32.1%	55.2%	42.8%	56.5%	36.7%
	Female	58.5%	54.4%	57.3%	66.6%	58.5%	54.4%	57.0%	59.9%

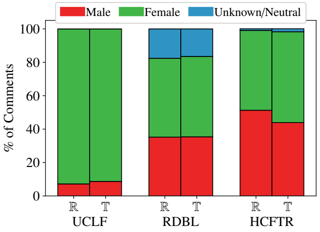


Fig. 1: Accuracy of uCLASSIFY (UCLF), READABLE (RDBL) and HACKERFACTOR (HCFTR) on comments by male, female (in Table) & non-binary (in Figure) authors from the two datasets – Reddit ($\mathbb{R}$) and Tumblr ($\mathbb{T}$). From the table, we see the platforms have low accuracy, independent of the dataset. Male comments have lower accuracy than females on all platforms. From the figure, we see all platforms predict female as the gender most often for non-binary authors, independent of the dataset. READABLE is the most 'fair' platform.

binary class, we cannot measure the accuracy. We thus study only the predictions for all comments on these platforms.

Observations. In the bar plots in Figure 1, we show the distribution of predictions for the non-binary author comments by the three gender analyzer platforms. We can see that for uCLASSIFY ($> 91\%$) and READABLE ($> 47\%$), the majority prediction is female, independent of the dataset. On HACKERFACTOR, the majority prediction for Reddit is male ($\approx 51\%$), but on Tumblr, the majority is again female ($\approx 54\%$). On READABLE, $\approx 18\%$ of Reddit comments and 16.5% Tumblr comments are predicted as neutral. HACKERFACTOR predicts the 'unknown' label for less than 2% of the inputs.

C. Performance of the BERT classifier (RQ1, RQ2, RQ3)

In this final segment, we test a BERT-base classifier under different fine-tuning settings described in Section III to verify if a simple classifier, finetuned on social media data from our two datasets can be used to mitigate some of the existing biases observed in the audit above. For these experiments, we choose 660k comments from each dataset (240k comments from male and female authors each, and 180k comments from non-binary authors) unless otherwise stated.

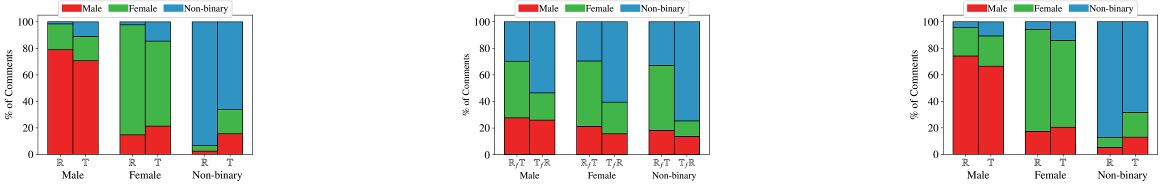
*1) **Fine-tuning entirely on one dataset:*** In this experiment, we check the model performance by fine-tuning entirely on comments from one dataset and then testing on comments from the same dataset (as baseline) and the other dataset (for generalizability). Figure 2a presents the results for the first scenario and Figure 2b presents the results for the second scenario.

Observations. From Figure 2a we note that the model performs very well on the comments from Reddit, with the overall accuracy being 84.4%. In contrast, the overall accuracy for Tumblr is only 67.1%. This shows that the model learns better from the comments on Reddit than on Tumblr. On looking at the results for each gender group, it is evident that the BERT model can identify non-binary authors' comments particularly well in the Reddit dataset, but the accuracy for cisgender binary authors, especially males (79%) is comparatively low. On the Tumblr dataset, we observe that the best performance is reported for males ($\approx 71\%$), and only 66% of the non-binary authors' comments are classified correctly. Compared to the existing platforms – uCLASSIFY, READABLE and HACKERFACTOR, the BERT models report higher accuracy, on both datasets. This shows that even a simple baseline model can not only improve performance (for cisgender binary authors) but also be fair toward sensitive groups (non-binary authors) within the target population.

Next, in Figure 2b, we evaluate the generalizability across datasets by testing a Reddit fine-tuned model on the Tumblr dataset ($\mathbb{R}_f\mathbb{T}$) and vice-versa ($\mathbb{T}_f\mathbb{R}$). We see a drastic drop in performance – accuracies of 37% for $\mathbb{R}_f\mathbb{T}$ and $\approx 38\%$ for $\mathbb{T}_f\mathbb{R}$. This indicates that there is a stark difference in the language dynamics of the two datasets and the model cannot generalize the learning from one to another. Interestingly, while in $\mathbb{R}_f\mathbb{T}$, the highest accuracy is reported for females – 49%, non-binary authors' accuracy is 32.8%. On the other hand, in $\mathbb{T}_f\mathbb{R}$ we see that the lowest accuracy is reported for female authors – $\approx 24\%$ and the accuracy for non-binary authors is 75%. Hence, a model trained on Tumblr can identify comments from non-binary authors on Reddit better than on Tumblr itself. This is possibly because the model trained on Tumblr sees more nuanced examples thus constructing a more complex decision boundary that is able to efficiently classify the non-binary comments from Reddit that are typically more distinct in nature.

*2) **Fine-tuning using a mixed dataset:*** To understand how a generalizable model may be designed, we fine-tune BERT using comments from both Reddit and Tumblr. We randomly sample 50% comments for each gender group from the two datasets to create a mixed fine-tuning dataset.

Observations. In Figure 2c, we report the results for this experiment. We can immediately observe gains from this setup as the overall accuracies improve to 79% ($\mathbb{R}$) and 67% ($\mathbb{T}$) compared to the case of fine-tuning on one and testing on the other dataset. For individual gender groups, we see that this mixed setup performs poorly on the Reddit dataset as compared to the original setup where the model was trained completely on Reddit. All gender groups report a drop in accuracy by $\approx 5\%$. For Tumblr, the results are different – only the accuracy of male authors reduces by 4%, with there being a slight improvement in accuracy for both female authors (1.3%) and non-binary authors (2%). Thus, while such a setup does not give any significant improvement for Tumblr and in fact reports lower accuracy for Reddit, it is more generalizable than fine-tuning on only one dataset. For the cisgender binary groups, this model is better than both READABLE and HACKERFACTOR, and for the male authors on uCLASSIFY.



(a) Fine-tune and test on the same dataset. (b) Fine-tune and test on different datasets. (c) Fine-tune by mixing the two datasets equally.

Fig. 2: Accuracy for the BERT model under different fine-tuning settings when testing on the Reddit ($\mathbb{R}$) and Tumblr ($\mathbb{T}$) datasets. $\mathbb{R}_f\mathbb{T}$ refers to the model fine-tuned on Reddit and tested on Tumblr and $\mathbb{T}_f\mathbb{R}$ refers to the model fine-tuned on Tumblr and tested on Reddit. Each bar shows how many comments for each gender group are classified as 'male', 'female', and 'non-binary'.

*3) **Few-shot learning with limited examples:*** The final setup we test for is also the most realistic scenario that is observed in real-world social networks. It is possible that none of the existing gender analyzers have trained their models on non-binary data because large-scale gender-labeled data is extremely difficult to collect, especially for the non-binary gender group. Thus, we envision a scenario where the BERT model has been fine-tuned on data from one dataset (for example – Reddit or Tumblr in our work) and is available for use by other platforms. The other platforms may have a very small number of labeled examples of all gender groups available through voluntary self-disclosure which can then be used to perform a second round of fine-tuning, similar to a few-shot learning scenario, and the model can then be deployed for use on the said platform.

We experimentally verify this scenario for the following values of k (number of shots) – $\{0, 1, 40, 200, 800, 2000, 5000\}$. The model is first fine-tuned using the entire training data (70% of the total) from one of the platforms. This is followed by a k-shot learning and testing using data from the second dataset. Thus, $\mathbb{T}_f\mathbb{R}_{200}$ implies that the model is first fine-tuned using Tumblr's data and then 200 examples from each gender group in Reddit's dataset are used to perform a second round of learning. The results are noted in Figure 3. We perform 5-fold cross validation and report the average accuracy.

Observations. From Figure 3a we see that the model fine-tuned on Tumblr and provided with k examples from Reddit always performs better than the opposite scenario (average accuracy difference of 18%). In the 0- and 1-shot setting, the accuracy of both models is low – 38.5% and 37.5% for $\mathbb{T}_f\mathbb{R}_0$ & $\mathbb{R}_f\mathbb{T}_0$ respectively and, 40.6% and 33% for $\mathbb{T}_f\mathbb{R}_1$ & $\mathbb{R}_f\mathbb{T}_1$ respectively. For $\mathbb{T}_f\mathbb{R}_{5000}$ & $\mathbb{R}_f\mathbb{T}_{5000}$, the overall accuracies are 77% and 52% respectively. Thus it is clear that $\mathbb{T}_f\mathbb{R}_k$ is the better model and generalizes well with very few examples from Reddit.

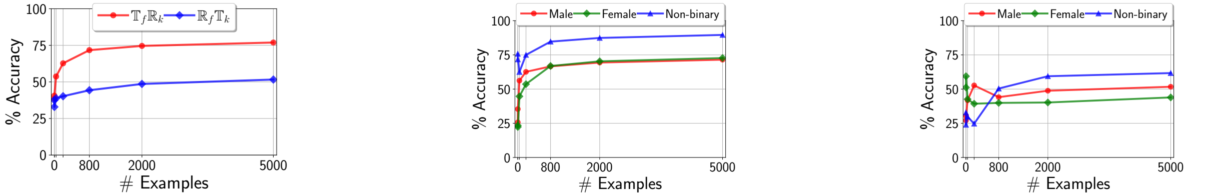
On looking at the gender group-wise accuracy for both models in Figures 3b and 3c, we see that with an increasing number of examples, the accuracy for non-binary gender group is higher than the cisgender group. From Fig. 3b, we see that $\mathbb{T}_f\mathbb{R}_k$ has more than 40% difference in accuracy between the non-binary and cisgender binary authors for the 0- vs 1-shot learning scenario. This difference is 25% on average, across all k-shots. The model achieves 90% accuracy for non-binary authors and 72% for the cisgender binary authors after training with only 5000 examples from Reddit. For $\mathbb{R}_f\mathbb{T}_k$ (Fig. 3c), we see that not only is the overall accuracy low, but the gender-wise accuracies are also low. While the accuracies do increase with an increasing number of examples, the growth is not very significant. The maximum accuracy for males and non-binary authors is observed for $\mathbb{R}_f\mathbb{T}_{5000}$ – 51.7% for males and 61.7% for non-binary. The accuracy for females is 42%, with the highest being observed in the 1-shot scenario at 59.4%.

These results indicate that few-shot learning works when only a limited set of examples is available and the generalizability from Tumblr to Reddit is higher than from Reddit to Tumblr.

Overall takeaways. The following conclusions can be drawn from the above subsections.

- The prediction accuracy for all existing platforms is $\approx 53\%$ for comments from Reddit and $\approx 47.5\%$ for Tumblr. Thus, the platforms perform poorly on both datasets, with an accuracy that is worse than a coin-toss for Tumblr. The accuracy is higher for females on all platforms. This indicates possible *representation* bias, whereby the sampling of the training dataset on the platforms may have idiosyncrasies.
- None of the platforms show good performance on either dataset, indicating a lack of generalizability.
- The commercial third-party platforms predict the majority of the non-binary comments to be female, with uCLASSIFY's predictions being overwhelmingly female. HACKERFACTOR predicts majority female for Reddit and majority male for Tumblr. As noted earlier, if cybercrime investigations would resort to such tools for gender identification then most of the non-binary individuals would be at risk of gender misprediction and may be exposed to unnecessary harassment given that there is a recent surge in the number of cybercrimes committed by women[6].
- While both READABLE and HACKERFACTOR have a third prediction class – 'neutral' and 'unknown' respectively, only READABLE has a non-negligible prediction for this class. Hence, it is slightly fairer among the two platforms.
- A fine-tuned BERT-based multilabel classifier works better than all the existing platforms that we have audited for the task of gender classification. It can be trained to predict non-binary gender labels as well, with sufficiently high accuracy.
- A BERT model fine-tuned on a mixed dataset from both Reddit and Tumblr generalizes better than the one fine-tuned on only one dataset. It should be noted that such a

[6] https://www.trendmicro.com/vinfo/us/security/news/cybercrime-and-digital-threats/gender-in-cybercrime



(a) Overall accuracy for $\mathbb{T}_f\mathbb{R}_k$ & $\mathbb{R}_f\mathbb{T}_k$

(b) Per gender accuracy for the $\mathbb{T}_f\mathbb{R}_k$ setup.

(c) Per gender accuracy for $\mathbb{R}_f\mathbb{T}_k$ setup.

Fig. 3: Accuracy for the few-shot learning scenario. Two models are evaluated– fine-tuning on Tumblr with few-shot examples from Reddit ($\mathbb{T}_f\mathbb{R}_k$) and vice versa ($\mathbb{R}_f\mathbb{T}_k$). k signifies the number of examples under consideration, ranging from 0 to 5000. ($\mathbb{T}_f\mathbb{R}_k$) has the overall best performance as well as the best performance for non-binary data with 5000 examples from Reddit.

Platforms	$\mathbb{R}$				$\mathbb{T}$			
	M	**F**	**NB**	**Overall**	**M**	**F**	**NB**	**Overall**
ChatGPT	**80.8**	58.4	87.8	75.7	26.2	44.6	**48.2**	39.7
BERT ($\mathbb{T}_f\mathbb{R}_{5000}$)	75.8	73	**88.4**	**79.1**	**55.8**	60.8	40.6	**52.4**
uClassify	14.8	**95.4**	X	55.1	16.8	**88.8**	X	35.2
Readable	41	51.4	X	30.8	30.8	48.4	X	26.4
HackerFactor	55.4	57.2	X	56.3	43.8	57.8	X	50.8

TABLE IV: Accuracy on 1500 comments (500 from each gender group) from each dataset for all platforms. On the Reddit dataset, ChatGPT's performance is comparable to our best finetuned BERT model and far superior to the other platforms. On the Tumblr dataset, ChatGPT's performance is best for the non-binary comments. For male and female comments, ChatGPT performs worse than the other platforms. Maximum values are in bold. M: male, F: female, NB: non-binary.

model is still comparable to the few-shot setting.

D. *Performance comparison with* ChatGPT *(RQ4)*

LLMs are being adopted on a wide-scale for general purpose and domain specific tasks these days. We simulate a text-based gender analyzer on ChatGPT, one of the most sophisticated LLMs available presently and audit it for a small sample of comments (500 from each gender group) from the two datasets. This allows us to understand its performance in a zero-shot scenario for a domain-specific task. We also compare it against the other available tools as well as our best finetuned model. The results are presented in Table IV. We see that its performance is comparable to our BERT model for the Reddit dataset, with the best performance reported for male comments and almost as good as the BERT model for non-binary comments. Conversely, on the Tumblr dataset it performs worse than all other platforms for the male/female comments but performs better than BERT for the non-binary comments. Overall, even such a powerful model is not able to report a sustained performance across all the three classes and the two datasets (albeit a small sample due to budget constraints) and therefore one needs to use it with appropriate caution (and possibly after improvement) for the simulation of the gender analysis task.

V. Discussion and Conclusion

In this paper, we audit various gender analyzer platforms viz. uClassify, Readable and HackerFactor. Overall results indicate a prediction accuracy of less than 60% on both the Reddit and Tumblr datasets. The platforms perform better for text by female authors than by male authors, especially on uClassify. Thus, both uClassify and Readable seem to be over-compensating and are predisposed to predict female for the input data. HackerFactor has a more balanced output but is still not very accurate for either gender. Thus to answer **RQ1** – the existing tools are not accurate at identifying the gender of binary authors, reporting lower accuracy than what they claim. The performance on the two datasets (from the table in Figure 1) answers **RQ3** – the reported accuracies are similar, albeit poor.

Next, we audit these tools for non-binary authors' text and observe that all platforms are *female-leaning*, thus reinforcing the societal belief about non-binary individuals being considered more effeminate. uClassify doesn't predict any third label at all; Readable predicts a 'neutral' class for $> 16\%$ of the text by non-binary authors and is the most 'fair' amongst the three platforms. As these softwares are easily available through their web platforms, bias propagation, and reinforcement are very easy. Our study has shown that these platforms are highly biased against non-binary individuals, and any use in downstream tasks without proper acknowledgement of these shortcomings may lead to discriminatory practices against these minority groups. Even if the platforms acknowledge their low accuracies for datasets like social media comments, they fail to identify the bigger social issue of a missing class - viz. non-binary authors. This answers **RQ2**. Here we see that the performance on the two datasets is similar (from the bar plots in Figure 1), answering **RQ3**.

Finally, we fine-tune a simple BERT-based multilabel classifier to address the shortcomings identified in our audit study. We assessed multiple fine-tuning settings using one or both datasets and tested on the same and the other dataset. Our results showed that the model when fine-tuned on Reddit and tested on the same gave an overall accuracy of 84%. The non-binary authors' comments were accurately identified 93% of the time (Figure 2a). The performance on the Tumblr dataset on the other hand is not as high – 67%. We also note that the models do not generalize well unless they have been fine-tuned with data from both platforms in some way. If a model fine-tuned on dataset A is tested for dataset B, the accuracy is lower than 40% (Figure 2b). Fine-tuning using a mixture of the two datasets leads to improved performance (Figure 2c). Our final approach is geared toward understanding how such models can perform under realistic scenarios where gender-annotated data from multiple social media platforms is hard to come by. We fine-tune using the data from one of the datasets and then use few-shot learning (between 0 and 5000 examples for each gender group from the other dataset). The model fine-tuned on Tumblr data and with 5000 Reddit examples (per gender group) performs the best with 90% accuracy for the non-binary class and overall accuracy of 77% on the test set from Reddit (Figure 3). The BERT model is better than the existing platforms for the binary gender prediction as seen in Figure 2 (**RQ1**) and is highly accurate for predicting the non-binary gender both in the standard fine-tuning setting and the few-shot learning scenario, as seen in Figs. 2a and 3 (**RQ2**). Finally, the model generalizes well when it observes data from both datasets in some combination, but the individual performance on Tumblr dataset is not noteworthy, whereas the platform learns better from the Reddit dataset (**RQ3**). We attribute this to the language style prevalent in each of these platforms.

On auditing a small sample of comments on CHATGPT, we observe that performance on Reddit is comparable to our best finetuned BERT model whereas the accuracy on Tumblr is $\approx 13\%$ less (Table IV). This indicates that even a highly sophisticated language model does not identify the gender markers on the Tumblr dataset well and further analysis is needed for more generalizable conclusions (**RQ4**).

Finally, as part of our future work, we plan to increase the scope of this audit to include more open-source and commercial gender analyzers and test their prediction accuracies on more diverse datasets from other sources like Twitter, Facebook, Internet message boards, and other niche platforms[7]. Our experiments using the BERT base model has shown merit in using transformer models for this application and we plan to extend this to other neural architectures which may exhibit equivalent or better performance but with a shorter inference time. This would go a long way in reducing the societal bias against non-binary individuals and also reinstating their voice and position in the online sphere of influence.

References

[1] C. Richards, W. P. Bouman, L. Seal, M. J. Barker, T. O. Nieder, and G. T'Sjoen, "Non-binary or genderqueer genders," *International Review of Psychiatry*, pp. 95–102, 2016.

[2] UK, "Gender recognition act 2004," 2004. Accessed: 2023-01-31.

[3] L. S. Weinhardt, P. Stevens, H. Xie, L. M. Wesp, S. A. John, I. Apchemengich, D. Kioko, S. Chavez-Korell, K. M. Cochran, J. M. Watjen, *et al.*, "Transgender and gender nonconforming youths' public facilities use and psychological well-being: a mixed-method study," *Transgender health*, pp. 140–150, 2017.

[4] B. P. Bagagli, T. V. Chaves, and M. G. Zoppi Fontana, "Trans women and public restrooms: The legal discourse and its violence," *Frontiers in Sociology*, 2021.

[5] T. Bates, C. S. Thomas, and A. R. Timming, "Employment discrimination against gender diverse individuals in western australia," *Equality, Diversity and Inclusion: An International Journal*, pp. 273–289, 2021.

[6] UN, "The struggle of trans and gender-diverse persons." https://www.ohchr.org/en/special-procedures/ie-sexual-orientation-and-gender-identity/struggle-trans-and-gender-diverse-persons, 2021. Accessed: 2023-01-31.

[7] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *PMLR FAT**, 2018.

[8] S. Jaiswal, K. Duggirala, A. Dash, and A. Mukherjee, "Two-face: Adversarial audit of commercial face recognition systems," *AAAI ICWSM*, pp. 381–392, 2022.

[9] T. Sühr, S. Hilgard, and H. Lakkaraju, "Does fair ranking improve minority outcomes? understanding the interplay of human and algorithmic biases in online hiring," in *AIES*, pp. 989–999, 2021.

[10] Y. Feng and C. Shah, "Has ceo gender bias really been fixed? adversarial attacking and improving gender fairness in image search," in *AAAI*, 2022.

[11] O. Keyes, "The misgendering machines: Trans/hci implications of automatic gender recognition," *CSCW*, pp. 1–22, 2018.

[12] M. K. Scheuerman, J. M. Paul, and J. R. Brubaker, "How computers see gender: An evaluation of gender classification in commercial facial analysis services," *CSCW*, pp. 1–33, 2019.

[13] S. Jaiswal and A. Mukherjee, "Marching with the pink parade: Evaluating visual search recommendations for non-binary clothing items," CHI Extended Abstracts, 2022.

[14] Amazon, "Amazon aws rekognition." https://aws.amazon.com/rekognition/faqs/, 2022. Accessed: 2023-01-31.

[15] Face++, "Face++ detect." https://www.faceplusplus.com/face-detection/, 2022. Accessed: 2023-01-31.

[16] Clarifai, "Clarifai." https://www.clarifai.com/models/ai-face-detection, 2022. Accessed: 2023-01-31.

[17] Microsoft, "Microsoft azure face." https://azure.microsoft.com/en-in/services/cognitive-services/face/, 2022. Accessed: 2023-01-31.

[18] uClassify, "uclassify gender analyzer." https://www.uclassify.com/browse/uclassify/genderanalyzer_v5, 2022. Accessed: 2023-01-31.

[19] Readable, "Readable gender analyzer." https://app.readable.com/text/gender/, 2022. Accessed: 2023-01-31.

[20] HackerFactor, "Hackerfactor gender guesser." https://www.hackerfactor.com/GenderGuesser.php, 2022. Accessed: 2023-01-31.

[21] N. Cheng, R. Chandramouli, and K. Subbalakshmi, "Author gender identification from text," *Digital Investigation*, pp. 78–88, 2011.

[22] S. Mukherjee and P. K. Bala, "Gender classification of microblog text based on authorial style," *ISeB*, pp. 117–138, 2017.

[23] J. Dastin, "Amazon scraps secret ai recruiting tool that showed bias against women." https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G, 2018. Accessed: 2022-05-31.

[24] B. Onikoyi, N. Nnamoko, and I. Korkontzelos, "Gender prediction with descriptive textual data using a machine learning approach," *Natural Language Processing Journal*, vol. 4, p. 100018, 2023.

[25] C. Sandvig, K. Hamilton, K. Karahalios, and C. Langbort, "Auditing algorithms: Research methods for detecting discrimination on internet platforms," *Data and discrimination: converting critical concerns into productive inquiry*, 2014.

[26] OpenAI, "Chatgpt." https://chat.openai.com/, 2022. Accessed: 2023-01-31.

[27] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[28] Reddit, "Reddit." https://reddit.com. Accessed: 2023-01-31.

[29] Tumblr, "Tumblr." https://www.tumblr.com/. Accessed: 2023-01-31.

[30] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM CSUR*, pp. 1–35, 2021.

[31] F. Safara, A. S. Mohammed, M. Yousif Potrus, S. Ali, Q. T. Tho, A. Souri, F. Janenia, and M. Hosseinzadeh, "An author gender detection method using whale optimization algorithm and artificial neural network," *IEEE Access*, pp. 48428–48437, 2020.

[32] M. Vicente, F. Batista, and J. P. Carvalho, *Gender Detection of Twitter Users Based on Multiple Information Sources*, pp. 39–54. 2019.

[33] P. Vashisth and K. Meehan, "Gender classification using twitter text data," in *ISSC*, 2020.

[34] A. F. Sotelo, H. Gómez-Adorno, O. Esquivel-Flores, and G. Bel-Enguix, "Gender identification in social media using transfer learning," in *Mexican Conference on Pattern Recognition*, pp. 293–303, 2020.

[35] E. E. Abdallah, J. R. Alzghoul, and M. Alzghool, "Age and gender prediction in open domain text," *Procedia Computer Science*, pp. 563–570, 2020.

[36] A. Angeles and M. N. Quintos, "Text-based gender classification of twitter data using naive bayes and svm algorithm," in *TENCON*, 2021.

[37] H. Liu and M. Cocea, "Fuzzy rule based systems for gender classification from blog data," in *ICACI*, pp. 79–84, 2018.

[38] C. Aravantinou, V. Simaki, I. Mporas, and V. Megalooikonomou, "Gender classification of web authors using feature selection and language models," in *Speech and Computer*, pp. 226–233, 2015.

[39] W. Deitrick, Z. Miller, B. Valyou, B. Dickinson, T. Munson, and W. Hu, "Author gender prediction in an email stream using neural networks," 2012.

[40] A. Bartle and J. Zheng, "Gender classification with deep learning," *Stanfordcs, 224d Course Project Report*, pp. 1–7, 2015.

[41] E. Vasilev, "Inferring gender of reddit users," master's thesis, Universität Koblenz-Landau, Universitätsbibliothek, 2018.

[7] https://nonbinary.wiki/wiki/Websites_and_social_networks