Automated Gender Bias Evaluation in YouTube

Gizem Gezici*†
*Huawei Turkey R&D Center, Istanbul, Turkey
†Sabanci University, Istanbul, Turkey
gizem.gezici@huawei.com, gizemgezici@sabanciuniv.edu

Abstract—Students are increasingly using online materials to learn new subjects or to supplement their learning process in educational institutions. Issues regarding gender bias have been raised in the context of formal education and some measures have been proposed to mitigate them. In our previous work, we investigate the perceived gender bias in YouTube using manually annotations for detecting the narrators’ perceived gender in educational videos. In this work, our goal is to evaluate the perceived gender bias in online education by exploiting an automated annotations. The automated pipeline has already proposed in a recent paper, thus in this paper we only share our empirical results with important findings. Our results show that educational videos are biased towards the male and STEM-related videos are more biased than their NON-STEM counterparts.

I. INTRODUCTION

In our previous work [1], we achieved to show that there is ideological bias in search results of Bing and Google. It seems that both search engines are biased and they are biased towards the liberal side. Yet, we had fulfilled this study in the top-10 search engine result pages (SERPs) annotated via crowd-sourcing so these results does not help us to investigate the source of bias, i.e. if the bias comes from the dataset, or the ranking algorithm itself. Therefore, as the next step of this research study, we aimed to investigate the source of bias. For detecting the source of bias, we need to annotate the full SERPs of Bing and Google, note that we have 250 web documents for each controversial query and in total there are 57 queries which means that we need an automated model for the annotation phase.

For these reasons, we firstly extended our annotated dataset and then experiment with different models especially deep learning models to be able to annotate a given document with an acceptable level of accuracy. In the previous report, our models did not give satisfactory results which means that we could not use them for the annotation task, thereby to examine the source of bias in SERPs. Because of this, in this phase of our research we tried various approaches from literature to improve our BERT-based models as well as experimented with more traditional machine learning models such as XGBoost [2], Random Forest [3].

After we tried different approaches to improve our model, we observed a significant increase in class-wise model evaluation results. However, we believe that class-wise accuracies are still not sufficient for an automated model that will be used for bias analysis. Thus, we left the source of bias analysis part as unsolved and decided to apply our proposed bias measures for analysing gender bias in Youtube in the scope of education.

Then, in [4], we display our automated perceived gender bias pipeline to detect the source of bias without any empirical results. Finally, in this work we show our findings using the aforementioned automated pipeline with two different automated gender annotation models to obtain narrators’ gender distribution for a given video.

II. GENDER BIAS EVALUATION METHODOLOGY

This section describes the methodology for evaluating perceived gender bias without a binary gender assumption. The bias measures proposed in [5] are adapted to using male/female probability values instead of single perceived gender labels. Further, a similar evaluation procedure presented in [5] is fulfilled for identifying potential bias as well as tracking the source of bias (if applicable) associated with those adapted measures.

A. Preliminaries and Research Questions

Using an automated approach to obtain the perceived gender probability distributions of videos through voice information, this work mainly aims to following research questions. The first research question is:

RQ1: On a perceived male-female gender spectrum, does YouTube return biased YVRPs in response to various educational queries using more fine-grained measures?

The second research question is:

RQ2: Is there a significant difference in perceived gender bias that is computed in a more fine-grained manner in YVRPs returned in response to STEM vs. NON-STEM educational queries?

The third research question is:

RQ3: Do different cut-off values affect the existence of perceived bias and magnitude of bias difference by using more fine-grained measures between STEM and NON-STEM fields?

The fourth research question is:

RQ4: Do different cut-off values affect the magnitude of perceived bias of STEM and NON-STEM fields separately that is measured in a more fine-grained manner?

The last research question which is totally a new question that could not be answered in our previous work [5]. In the context of this paper, the source of bias (if exists) is investigated. If the bias measured in the full YVRPs is consistent with the top video search results, i.e. especially in top-3, top-10 since these search results attract users’ attention the most,
then it can be inferred that the bias comes from the data itself. If there are some differences between those bias results, then the ranking algorithm could also be blamed. Note that the data and ranking algorithm both could be responsible for the biased YVRPs.

The fifth research question is:

**RQ5:** What is the source of bias (if exists), does it come from the input data, or the ranking algorithm?

### B. Measures of Bias

Let \( q \in Q \) be the set of educational queries about the majors in STEM and NON-STEM fields. When a query \( q \in Q \) is issued to YouTube, YouTube returns a YVRP \( r \). The probability value associated with the perceived gender of the \( i \)-th retrieved video \( r_i \) with respect to \( q \) is defined as \( \text{prob}_{G_m}(r_i) \) for male and \( \text{prob}_{G_f}(r_i) \) for female. For reference, Table 2 shows a summary of all the symbols, functions and labels used throughout this paper.

In the scope of this paper, similar to Section 3, the main aim is to satisfy group fairness criteria of equality of outcome where male and female genders should be equally represented in YVRPs. Thus the perceived gender bias is measured as the difference between the representation of male and female genders.

Formally, the perceived gender bias in a YVRP \( r \) is measured as follows:

\[
\Delta_f(r) = f_{G_m}(r) - f_{G_f}(r)
\]  

(1)

For the function \( f(r) \), two bias measures that are proposed in 3 are adapted to using probability scores instead of single perceived gender labels in the scope of this paper. Note that the videos annotated with \( G_{not-rel} \) and \( G_{N/A} \) are initially discarded before the bias score computations. The two adapted measures of representation and exposure are denoted by \( \text{Rep}_{\text{prob}}(\@n) \) and \( \text{Exp}_{\text{prob}}(\@n) \) respectively. The first adapted measure of bias, \( \text{Rep}_{\text{prob}}(\@n) \) which computes a bias score using probability values associated with the perceived gender label of male as follows:

\[
\text{Rep}_{\text{prob}}G_m(\@n) = \frac{1}{\text{prob}_{mf}} \sum_{i=1}^{n} \text{prob}_{G_m}(r_i)
\]  

(2)

Note that \( \text{Rep}_{\text{prob}}G_f(\@n) \) is computed in the same way. The following equation by substituting Eq. (2) in Eq. (1):

\[
\Delta_{\text{Rep}_{\text{prob}}}(\@n) = \frac{1}{\text{prob}_{mf}} \sum_{i=1}^{n} (\text{prob}_{G_m}(r_i)) - \text{prob}_{G_f}(r_i))
\]  

(3)

Since the first bias measure of \( \text{Rep}_{\text{prob}}(\@n) \) has a weak sense of rank information, i.e. all the positions contribute to bias score in an equal manner, the second measure is presented by adapting \( \text{Exp}_{\text{prob}}(\@n) \) that was proposed in 3 to using probability scores. The second adapted measure of bias, \( \text{Exp}_{\text{prob}}(\@n) \) which computes a bias score in terms of exposure, using probability values associated with the perceived gender label of male as follows:

\[
\text{Exp}_{\text{prob}}G_m(\@n) = \sum_{i=1}^{n} \frac{1}{\log(i+1)} \frac{\text{prob}_{G_m}(r_i)}{\text{prob}_{G_m}(r_i) + \text{prob}_{G_f}(r_i)}
\]  

(4)

Note that \( \text{Exp}_{\text{prob}}G_f(\@n) \) is computed in the same way. The following equation is obtained by substituting Eq. (4) in Eq. (1):

\[
\Delta_{\text{Exp}_{\text{prob}}}(\@n) = \sum_{i=1}^{n} \frac{1}{\log(i+1)} \frac{\text{prob}_{G_m}(r_i) - \text{prob}_{G_f}(r_i)}{\text{prob}_{G_m}(r_i) + \text{prob}_{G_f}(r_i)}
\]  

(5)

The scores of the proposed measures are easy to interpret, for a given ranked list, the scores of two gender groups sum up to 1. If the bias scores are interpreted with respect to the equal representation using \( \text{Rep}_{\text{prob}}(\@n) \), then it can be inferred which gender group is more/less represented than the desired representation. Same holds true for the exposure measure, \( \text{Exp}_{\text{prob}}(\@n) \), which determines if a gender group is more or less exposed than the desirable situation of the equal exposure. For interpreting the results, if the value of 0.5 which is the desired case, is subtracted from the measure scores of male and female gender for a given list, then the remaining bias scores of each gender group are symmetric. Same holds for the exposure measure. Additionally, these adapted bias measures are expected to compute smoother and more realistic bias scores owing to the probability scores of the perceived gender groups instead of single labels which are too deterministic for annotating real datasets.

After the computation of adapted representation and exposure bias scores, the mean bias (MB) and mean absolute bias (MAB) of these measures can be further computed over a set of queries in the dataset to aggregate the bias results. MB score of STEM field computes a mean value over all the STEM queries’ scores for the corresponding measure, whereas MAB computes a mean value over all the absolute value of the measure scores for the STEM queries. Note that MB shows towards which perceived gender the results are biased and MAB solves the limitation of MB if different queries have bias contributions with opposite signs and cancel each other out. Thus, MB and MAB measures are complementary for aggregating the results and interpreting those results in a proper way.

Please note that in the context of this paper, the probability scores correspond to each perceived gender label, i.e. \( \text{prob}_{G_m}(r_i) \) and \( \text{prob}_{G_f}(r_i) \) for male and female respectively, of a given video is computed merely based on the narrators’ perceived gender by using voice information of the narrators. Since the probability scores are leveraged, there is no gender binary assumption throughout this paper.

### III. Experimental Setup

First, it is determined if YouTube returns biased results in terms of the narrator’s perceived gender annotated with prob-
ability scores for STEM and NON-STEM fields separately, and if so, whether YouTube’s top-200 relevant search results suffer from the same level of bias, such that the difference between STEM and NON-STEM queries/course modules is not statistically significant. Further the source of bias is tracked, if bias exists then it is investigated whether the bias comes from the input data or the ranking algorithm itself. Note that computing probability scores for each perceived gender group was fulfilled by using two different automated models, namely Feed-Forward Gender Detector and inaSpeechSegmenter models. Thus, perceived gender bias is measured for these two models separately.

### A. Feed-Forward Gender Detector Model Results

In Table I, the perceived gender bias is displayed using the adapted measures presented in [4] namely \( \text{RecProb} @ n \) and \( \text{ExpProb} @ n \) for different cut-off values of \( n = 3, 10, 20 \). All MB and MAB scores are positive for both bias measures; the one-sample t-test computed on MBs and MABs are statistically significant for the measures where p-value < .001 denoted as ***. The two-tailed paired t-test computed on MB and MABs to compare the difference in bias between the STEM and NON-STEM fields, the results indicate that their differences are statistically not significant on the bias measures of \( \text{RecProb} @ n \) and \( \text{ExpProb} @ n \) where \( n = 3, 10, 20 \). Some effect sizes that correspond to the difference of bias using MAB scores are negative which indicates that the MAB score of the perceived gender group of female is higher than male, albeit statistically not significant.

Regarding the impact of different cut-off values, in Table III using MB scores, it is observed that, both STEM and NON-STEM fields show similar scores for different cut-off values on both \( \text{RecProb} @ n \) and \( \text{ExpProb} @ n \) measures, i.e. the two-tailed paired t-test computed on MB scores are statistically not significant. On the other hand, in Table IV using MAB scores, it is observed that cut-off values might affect the perceived gender bias in STEM and NON-STEM fields. The two-tailed paired t-test computed on MABs of STEM field is statistically significant for both measures of \( \text{RecProb} @ n \) and \( \text{ExpProb} @ n \) between the following cut-off values, \( n = 3 \) vs. \( n = 10 \) and \( n = 3 \) vs. \( n = 20 \) with difference confidence levels. For the measure of \( \text{RecProb} @ n \) between \( n = 3 \) vs. \( n = 10 \), p-value which corresponds to the significance level of \( \alpha = .01 \), and \( n = 3 \) vs. \( n = 20 \), p-value = .0002 corresponds to the significance level of \( \alpha = .01 \) and p-value = 0.00001 corresponds to the significance level of \( \alpha = .005 \) respectively after Bonferroni correction was applied. For the measure of \( \text{ExpProb} @ n \) between \( n = 3 \) vs. \( n = 10 \) and \( n = 3 \) vs. \( n = 20 \), p-value = .00005 which corresponds to the significance level of \( \alpha = .01 \) and p-value = .000003 which corresponds to the significance level of \( \alpha = .001 \) respectively with Bonferroni correction. The two-tailed paired t-test computed on MABs of NON-STEM field is statistically significant for both measures of \( \text{RecProb} @ n \) and \( \text{ExpProb} @ n \) between different cut-off values. For the measure of \( \text{RecProb} @ n \) the difference is statistically significant only between \( n = 3 \) vs. \( n = 20 \), p-value = .0002 which corresponds to the significance level of \( \alpha = .01 \). For the measure of \( \text{ExpProb} @ n \) between \( n = 10 \) vs. \( n = 20 \) and \( n = 3 \) vs. \( n = 20 \), p-value = .0002 which corresponds to the significance level of \( \alpha = .05 \) and p-value = .0001 which corresponds to the significance level of \( \alpha = .01 \) respectively with Bonferroni correction.

For tracking the source of bias, only the bias scores from the measure of \( \text{RecProb} @ n \) are used since the strong sense of rank information, i.e. \( \text{ExpProb} @ n \), is not meaningful where \( n = 200 \) from the user’s perspective. In Table V the two-tailed paired t-test computed on MBs of STEM and NON-STEM fields are statistically not significant. Unlike, the two-tailed paired t-test computed on MABs of STEM field is statistically

### TABLE I: Symbols, functions, and labels used throughout this paper

| Symbols                  | Description                                                                 |
|--------------------------|-----------------------------------------------------------------------------|
| \( Q \)                  | set of queries.                                                             |
| \( q \)                  | a query \( q \in Q \).                                                      |
| \( r \)                  | a ranked list of the given YVRP (list of retrieved videos).                |
| \( r_i \)                | the video in \( r \) retrieved at rank \( i \).                            |
| \( |r| \)                  | size of \( r \) (number of videos in the ranked list).                     |
| \( n \)                  | number of videos considered in \( r \) (cut-off).                          |

| Functions                | Description                                                                 |
|--------------------------|-----------------------------------------------------------------------------|
| \( \text{prob}_{G_m}(r_i) \) | returns the probability value corresponds to male gender for \( r_i \).   |
| \( \text{prob}_{G_f}(r_i) \) | returns the probability value corresponds to female gender for \( r_i \). |
| \( f(r) \)               | an evaluation measure for YVRPs.                                           |

| Labels                   | Description                                                                 |
|--------------------------|-----------------------------------------------------------------------------|
| \( G_m \)                | perceived male gender.                                                      |
| \( G_f \)                | perceived female gender.                                                    |
| \( G_{not\_rel} \)       | not-relevant wrt a query.                                                   |
| \( G_{N/A} \)            | N/A - gender annotation is not applicable.                                  |
TABLE II: **Feed-Forward Gender Detector**: *Perceived* gender bias in YouTube for the top-20 relevant results where \( n = 3, 10, 20 \), p-values of a two-tailed independent t-test computed between the **MB** and **MAB** scores of STEM and NON-STEM fields

|        | **Rep@3** | **Rep@10** | **Rep@20** | **Exp@3** | **Exp@10** | **Exp@20** |
|--------|-----------|-----------|-----------|-----------|-----------|-----------|
| **MB** |            |           |           |           |           |           |
| STEM   | 0.2881*** | 0.2651*** | 0.2352*** | 0.3059*** | 0.2777*** | 0.2525*** |
| NON-STEM | 0.2209*** | 0.2225*** | 0.1868*** | 0.2424*** | 0.2324*** | 0.2050*** |
| p-value | .53       | .51       | .41       | .56       | .52       | .43       |
| effect size d | 0.129    | 0.133    | 0.169    | 0.119    | 0.131    | 0.159    |
| **MAB** |            |           |           |           |           |           |
| STEM   | 0.5018*** | 0.3189*** | 0.2842*** | 0.5227*** | 0.3426*** | 0.3005*** |
| NON-STEM | 0.4893*** | 0.3667*** | 0.3139*** | 0.5026*** | 0.4028*** | 0.3425*** |
| p-value | .84       | .25       | .45       | .75       | .16       | .29       |
| effect size d | 0.041    | -0.233   | -0.154   | 0.065    | -0.286   | -0.215   |

TABLE III: **Feed-Forward Gender Detector**: *Perceived* gender bias in YouTube for the top-20 relevant results where \( n = 3, 10, 20 \), p-values of a two-tailed paired t-test computed between the **MB** scores of STEM and NON-STEM fields

|        | **Rep@3** | **Rep@10** | **Rep@20** | **Exp@3** | **Exp@10** | **Exp@20** |
|--------|-----------|-----------|-----------|-----------|-----------|-----------|
| **MB** |            |           |           |           |           |           |
| STEM   | 0.2881*** | 0.2209*** | 0.2225*** | 0.2777*** | 0.2525*** |           |
| p-value | .72       | .97       | .12       | .42       | .57       |           |
| effect size d | 0.057    | -0.004   | 0.104    | 0.134    | 0.077    |           |
| **MAB** |            |           |           |           |           |           |
| STEM   | 0.3059*** | 0.2424*** | 0.2324*** | 0.3025*** | 0.2424*** |           |
| p-value | .58       | .80       | .13       | .35       | .45       |           |
| effect size d | 0.066    | 0.021    | 0.075    | 0.129    | 0.083    |           |

TABLE IV: **Feed-Forward Gender Detector**: *Perceived* gender bias in YouTube for the top-20 relevant results where \( n = 3, 10, 20 \), p-values of a two-tailed paired t-test computed between the **MAB** scores of STEM and NON-STEM fields

|        | **Rep@3** | **Rep@10** | **Rep@20** | **Exp@3** | **Exp@10** | **Exp@20** |
|--------|-----------|-----------|-----------|-----------|-----------|-----------|
| **MAB** |            |           |           |           |           |           |
| STEM   | 0.5018*** | 0.3677*** | 0.3139*** | 0.5026*** | 0.3425*** |           |
| p-value | .0002     | .0051     | .0083     | < .0001   | .0002     |           |
| effect size d | 0.709    | 0.461    | 0.257    | 0.866     | 0.681     |           |
| **MAB** |            |           |           |           |           |           |
| STEM   | 0.5227*** | 0.5026*** | 0.3426*** | 0.5026*** | 0.3425*** |           |
| p-value | < .0001   | .0026     | .0002     | < .0001   | .0001     |           |
| effect size d | 0.674    | 0.382    | 0.291    | 0.857     | 0.621     |           |

TABLE V: **Feed-Forward Gender Detector**: *Perceived* gender bias in YouTube for all the relevant results crawled, i.e. \( |r| = 200 \) where \( n = 3, 10, 20, 200 \), p-values of a two-tailed paired t-test computed between the **MB** and **MAB** scores of STEM and NON-STEM fields using the measure of **Rep@tn**.

|        | **Rep@3** | **Rep@10** | **Rep@20** | **Exp@3** | **Exp@10** | **Exp@20** |
|--------|-----------|-----------|-----------|-----------|-----------|-----------|
| **MB** |            |           |           |           |           |           |
| STEM   | 0.2881*** | 0.2209*** | 0.2225*** | 0.2352*** | 0.1879*** | 0.1564*** |
| p-value | .16       | .32       | .405      | .301      | .211      | .194      |
| effect size d | 0.250    | 0.155    | 0.301    | 0.301    | 0.194    | 0.105    |
| **MAB** |            |           |           |           |           |           |
| STEM   | 0.5018*** | 0.4893*** | 0.3677*** | 0.2842*** | 0.3139*** | 0.3425*** |
| p-value | < .0001   | < .0001   | .0781     | < .0001   | .71       | .0001     |
| effect size d | 0.954    | 0.991    | 0.253    | 0.627     | 0.056     | 0.366     |
significant only between $n = 3$ vs. $n = 200$, p-value = .000004 which corresponds to the significance level of $\alpha = .0005$ with Bonferroni correction. On the other hand, the two-tailed paired t-test computed on MABs of NON-STEM field is statistically significant between $n = 3$ vs. $n = 200$, $n = 10$ vs. $n = 200$, and $n = 20$ vs. $n = 200$, p-value = .000003 which corresponds to the significance level of $\alpha = .0005$, p-value = .000008 to the $\alpha = .0005$, and p-value = .0001 to the $\alpha = .01$ respectively with Bonferroni correction.

In Table VI, the bias scores for each major in STEM and NON-STEM fields are displayed using the measures $\Re_{prob}@3$ and $\Exp_{prob}@n$ for different cut-off values. Note that the highest/lowest bias scores are denoted as highlighted. In Figure 1a, the overall perceived gender bias scores are compared for STEM and NON-STEM fields on the MB scores of $\Exp_{prob}@10$ using the Feed-Forward Gender Detector model. It is observed that STEM is more biased towards the male than NON-STEM field. In Figure 2(a) and (b), the impact of different cut-off values is displayed on bias scores of $\Delta_{\Re_{prob}@n}$ where $n = 3$ vs. $n = 10$ for STEM and NON-STEM fields. Similarly, in Figure 3(a) and (b), the measure of $\Delta_{\Exp_{prob}@n}$ is used for the same purpose where $n = 3$ vs. $n = 10$. Note that both of these figures visualise the effect of different cut-off values on the perceived gender bias using the Feed-Forward Gender Detector model.

**B. inaSpeechSegmenter Model Results**

In Table VII the perceived gender bias is displayed using the adapted measures for different cut-off values of $n = 3$, 10, 20. All MB and MAB scores are positive for both bias measures; the one-sample t-test computed on MBs and MABs are statistically significant for the measures where p-value < .001 denoted as ***. The two-tailed paired t-test computed on MB and MABs to compare the difference in bias between the STEM and NON-STEM fields, the results indicate that their differences are statistically not significant on the bias measures of $\Re_{prob}@n$ and $\Exp_{prob}@n$ where $n = 3$, 10, 20. Since Bonferroni correction was applied, the results are not statistically significant.

Regarding the impact of different cut-off values, in Table VIII using MB scores, it is observed that, both STEM and NON-STEM fields show similar scores for different cut-off values on both $\Re_{prob}@n$ and $\Exp_{prob}@n$ measures. i.e. the two-tailed paired t-test computed on MB scores are statistically not significant. Some effect sizes that correspond to the difference of bias using MB scores are negative, i.e. the MB score of the perceived gender group of female is higher than male, albeit statistically not significant. On the other hand, in Table IX using MAB scores, it is observed that cut-off values might affect the perceived gender bias in STEM and NON-STEM fields. The two-tailed paired t-test computed on MABs of STEM field is statistically significant for both measures of $\Re_{prob}@n$ and $\Exp_{prob}@n$. For the measure
TABLE VII: \textit{inSpeechSegmenter}: Perceived gender bias in YouTube for the top-20 relevant results where $n = 3, 10, 20$, p-values of a two-tailed independent t-test computed between the MB and MAB scores of STEM and NON-STEM fields

|       | $Rep@3$ | $Rep@10$ | $Rep@20$ | $Exp@3$ | $Exp@10$ | $Exp@20$ |
|-------|---------|---------|---------|---------|---------|---------|
| **MB** |         |         |         |         |         |         |
| STEM   | 0.4669*** | 0.4033*** | 0.3850*** | 0.4738*** | 0.4254*** | 0.4049*** |
| p-value | .0109  | .0143  | .0195  | .0122   | .0073   | .0084   |
| effect size d | .525   | .505   | .480   | .516    | .555    | .545    |
| NON-STEM | 0.1818*** | 0.2303*** | 0.2323*** | 0.1869*** | 0.2228*** | 0.2288*** |
| **MAB** |         |         |         |         |         |         |
| STEM   | 0.5906*** | 0.4189*** | 0.4039*** | 0.5968*** | 0.4469*** | 0.4194*** |
| p-value | .39     | .51     | .24     | .46     | .57     | .25     |
| effect size d | .173   | .135   | .237   | .152    | .117    | .233    |

TABLE VIII: \textit{inSpeechSegmenter}: Perceived gender bias in YouTube for the top-20 relevant results where $n = 3, 10, 20$, p-values of a two-tailed paired t-test computed between the MB scores of STEM and NON-STEM fields

|       | STEM | NON-STEM | STEM | NON-STEM | STEM | NON-STEM |
|-------|------|----------|------|----------|------|----------|
| $Rep@3$ | 0.4669*** | 0.1818*** | 0.4033*** | 0.2303*** | 0.4738*** | 0.4049*** |
| $Rep@10$ | 0.4033*** | 0.2323*** | 0.3850*** | 0.2323*** | 0.1869*** | 0.2288*** |
| p-value | .162   | .162   | .42     | .42     | .46     | .22     |
| effect size d | -.095 | -.095 | -.005   | -.005   | .0209   | -.103   |
| $Exp@3$ | 0.4738*** | 0.1869*** | 0.4049*** | 0.2288*** | 0.4738*** | 0.2288*** |
| $Exp@10$ | 0.4254*** | 0.2288*** | 0.3850*** | 0.2288*** | 0.4049*** | 0.2288*** |
| p-value | .31    | .31     | .24     | .24     | .22     | .47     |
| effect size d | .119   | .068   | .076    | .015    | .173    | -.083   |

TABLE IX: \textit{inSpeechSegmenter}: Perceived gender bias in YouTube for the top-20 relevant results where $n = 3, 10, 20$, p-values of a two-tailed paired t-test computed between the MAB scores of STEM and NON-STEM fields

|       | STEM | NON-STEM | STEM | NON-STEM | STEM | NON-STEM |
|-------|------|----------|------|----------|------|----------|
| $Rep@3$ | 0.5906*** | 0.5350*** | 0.4033*** | 0.3850*** | 0.4738*** | 0.4194*** |
| $Rep@10$ | 0.4189*** | 0.3849*** | 0.3850*** | 0.3849*** | 0.4049*** | 0.4194*** |
| p-value | .010   | .0025   | .24     | .24     | .0005   | .0008   |
| effect size d | -.596 | -.519  | .063    | .063    | .656    | .658    |
| $Exp@3$ | 0.5968*** | 0.5462*** | 0.4049*** | 0.4194*** | 0.5968*** | 0.5462*** |
| $Exp@10$ | 0.4469*** | 0.4181*** | 0.4049*** | 0.4194*** | 0.4469*** | 0.4181*** |
| p-value | .005   | .0013   | .11     | .11     | .0002   | .0002   |
| effect size d | .497  | .448   | .114    | .222    | .601    | .636    |

of $Re_{p_{\text{prob}}\@n}$, between only the cut-off values, $n = 3$ vs. $n = 20$, p-value = .0005 which corresponds to the significance level of $\alpha = .05$ with Bonferroni correction. For the measure of $Ex_{p_{\text{prob}}\@n}$, between the following cut-off values, $n = 3$ vs. $n = 10$ and $n = 3$ vs. $n = 20$ with difference confidence levels, p-value = .0005 which corresponds to the significance level of $\alpha = .05$ and p-value = .0002 which corresponds to the significance level of $\alpha = .05$ respectively with Bonferroni correction.

For tracking the source of bias, again only the bias scores from the measure of $Re_{p_{\text{prob}}\@n}$ are used. In Table X the two-tailed paired t-test computed on MBs of STEM and NON-STEM fields are statistically not significant. Unlike, the two-tailed paired t-test computed on MABs of STEM field is statistically significant only between $n = 3$ vs. $n = 200$, p-value = .0002 which corresponds to the significance level of $\alpha = .05$ with Bonferroni correction. Likewise, the two-tailed paired t-test computed on MABs of NON-STEM field is statistically
TABLE X: *inaSpeechSegmenter*: Perceived gender bias in YouTube for all the relevant results crawled, i.e. \( |r| = 200 \) where \( n = 3, 10, 20, 200 \), p-values of a two-tailed paired t-test computed between the MB and MAB scores of STEM and NON-STEM fields using the measure of \( \mathcal{R}_{ep@n} \).

|        | STEM  | NON-STEM | STEM  | NON-STEM | STEM  | NON-STEM |
|--------|-------|----------|-------|----------|-------|----------|
| Rep@3  | 0.4669*** | 0.1818*** | 0.4033*** | 0.2303*** | 0.3850*** | 0.2323*** |
| Rep@200| 0.1816*** | 0.2194*** | 0.1816*** | 0.2194*** | 0.1816*** | 0.2194*** |
| p-value | .15   | .62      | .27    | .74      | .49    | .53      |
| effect size d | 0.236 | -0.080   | 0.119  | 0.030    | 0.062  | 0.038    |

|        | MB    | NON-STEM | MAB   | NON-STEM | MB    | NON-STEM |
|--------|-------|----------|-------|----------|-------|----------|
| Rep@3  | 0.5906*** | 0.5350*** | 0.4189*** | 0.3849*** | 0.4039*** | 0.3454*** |
| Rep@200| 0.4001*** | 0.3108*** | 0.4001*** | 0.3108*** | 0.4001*** | 0.3108*** |
| p-value | .0002 | < .0001 | .52    | .0124    | .89    | .08      |
| effect size d | 0.522 | 0.830   | 0.057  | 0.310    | 0.012  | 0.145    |

TABLE XI: *inaSpeechSegmenter*: Perceived gender bias for specific majors of STEM and NON-STEM fields in YouTube for the top-20 relevant results - red denotes bias towards male while blue towards female.

| Biology | Chemistry | CS | Maths | Physics |
|---------|-----------|----|-------|---------|
| Rep@3  | 0.2419    | 0.3889 | 0.4883 | 0.6779 | 0.5319 | 0.1076 | 0.4098 | 0.4098 | 0.1443 | -0.2027 |
| Rep@10 | 0.1519    | 0.2940 | 0.3712 | 0.6105 | 0.5890 | 0.0763 | 0.5319 | 0.2229 | -0.0152 | 0.0763 |
| Rep@20 | 0.1553    | 0.2414 | 0.3923 | 0.5122 | 0.6238 | 0.0642 | 0.5366 | 0.3001 | 0.2272 | 0.0335 |
| Exp@3  | 0.2928    | 0.3663 | 0.4895 | 0.7011 | 0.5192 | 0.2521 | 0.4451 | 0.4398 | 0.0269 | -0.2295 |
| Exp@10 | 0.1922    | 0.3196 | 0.4012 | 0.6457 | 0.5682 | 0.1656 | 0.5091 | 0.3740 | 0.1458 | -0.0806 |
| Exp@20 | 0.1787    | 0.2757 | 0.4039 | 0.5644 | 0.6015 | 0.1251 | 0.5228 | 0.3406 | 0.1803 | -0.0248 |

TABLE XII: Annotation Model Analysis: Perceived gender bias in YouTube for the top-20 relevant results where \( n = 3, 10, 20 \), p-values of a two-tailed paired t-test computed between the MB and MAB scores of STEM field.

|        | Rep@3 | Rep@10 | Rep@20 | Exp@3 | Exp@10 | Exp@20 |
|--------|-------|--------|--------|-------|--------|--------|
| MB     | Feed-Forward Gender Detector | 0.2881*** | 0.2651*** | 0.2352*** | 0.3059*** | 0.2777*** | 0.2525*** |
|        | inaSpeechSegmenter | 0.4669*** | 0.4033*** | 0.3850*** | 0.4738*** | 0.4254*** | 0.4049*** |
| p-value | .0001 | < .0001 | < .0001 | .0002 | < .0001 | < .0001 |
| effect size d | -0.358 | -0.516 | -0.590 | -0.326 | -0.525 | -0.601 |
| MAB    | Feed-Forward Gender Detector | 0.5018*** | 0.3189*** | 0.2842*** | 0.5227*** | 0.3426*** | 0.3005*** |
|        | inaSpeechSegmenter | 0.5906*** | 0.4189*** | 0.4039*** | 0.5968*** | 0.4469*** | 0.4194*** |
| p-value | .0261 | .0006 | < .0001 | .0335 | .0008 | .0001 |
| effect size d | -0.281 | -0.449 | -0.569 | -0.224 | -0.453 | -0.560 |

significant only between \( n = 3 \) vs. \( n = 200 \), p-value = .00008 corresponds to the \( \alpha = .005 \) with Bonferroni correction.

In Table XI the bias scores for each major in STEM and NON-STEM fields are displayed using the measures \( \mathcal{R}_{ep@prob@n} \) and \( \mathcal{E}_{ep@prob@n} \) for different cut-off values. Note that the highest/lowest bias scores are denoted as highlighted. In Figure 1B the overall perceived gender bias scores are compared for STEM and NON-STEM fields on the MB score of \( \Delta \mathcal{E}_{ep@prob@10} \) using the inaSpeechSegmenter model. Table XII displays comparison of the bias scores of the two annotation models only for the STEM field. The two-tailed paired t-test computed on MBs of the STEM field, the bias differences are statistically significant for both measures with different confidence values. On the measure of \( \mathcal{R}_{ep@prob@n} \), p-value = .0001 which corresponds to the significance level of \( \alpha = .001 \) and p-value = .000007 which corresponds to the significance level of \( \alpha = .0005 \) and p-value = .00000003 which corresponds to the significance level of \( \alpha = .0001 \) respectively for \( n = 3, \ n = 10, \) and \( n = 20 \) with Bonferroni correction. On the measure of \( \mathcal{E}_{ep@prob@n} \), p-value = .00002.
which corresponds to the significance level of $\alpha = .001$, p-value $= .000004$ which corresponds to the significance level of $\alpha = .0005$ and p-value $= .0000002$ which corresponds to the significance level of $\alpha = .0001$ respectively for $n = 3$, $n = 10$, and $n = 20$ with Bonferroni correction. STEM field is biased using both models, towards the male gender (all MBs are positive) and inaSpeechSegmenter provides higher bias scores for the STEM field. Unlike, the two-tailed paired t-test computed on MABs of the STEM field is statistically significant for both measures only for $n = 20$ with the confidence levels of $\alpha = .0001$ and $\alpha = .0005$ for $\text{Rep}_{\text{prob}@n}$ and $\text{Exp}_{\text{prob}@n}$ respectively. For the NON-STEM field, NON-STEM fields show similar level of bias irrespective of the annotation model – the two-tailed paired t-test computed on MBs/MABs of the NON-STEM field is statistically not significant. This is verified across two measures with different cut-off values.

Figure 1 displays the comparison of the bias scores in STEM and NON-STEM fields for the Feed-Forward Gender Detector and inaSpeechSegmenter models. The error bars show the standard error on the scores of the corresponding field.

In Figure 1(a) and (b), the impact of different cut-off values is displayed on bias scores of $\Delta \text{Rep}_{\text{prob}@n}$ where $n = 3$ vs. $n = 10$ for STEM and NON-STEM fields. Similarly, in Figure 5(a) and (b), the measure of $\Delta \text{Exp}_{\text{prob}@n}$ is used for the same purpose where $n = 3$ vs. $n = 10$. Note that both of these figures visualise the effect of different cut-off values on the perceived gender bias using the inaSpeechSegmenter model.

IV. DIscussion

Initially, it is verified if the YVRPs are biased using the adapted measures of $\text{Rep}_{\text{prob}@n}$ and $\text{Exp}_{\text{prob}@n}$ (RQ1). If so, then it is investigated if the YVRPs suffer from the different magnitude of bias (RQ2) by examining if the difference between the bias scores of the YVRPs of STEM and NON-STEM fields are statistically significant. In Table II and Table VII using Feed-Forward Gender Detector and inaSpeechSegmenter models respectively, regarding the RQ1 the YVRPs of STEM and NON-STEM fields are both biased – the one-sample t-test applied on MB/MAB scores to check the existence of bias is statistically significant with p-value $< .0001$. These findings suggest that both STEM and non-STEM fields are biased towards male (all MB scores are positive). On the basis of MAB scores, it is observed that both STEM and NON-STEM exhibit an absolute bias. Regarding the RQ2, STEM and NON-STEM fields show similar levels of bias – the two tailed independent t-tests applied on MB/MAB scores in Table II and Table VII. The differences of bias are statistically not significant irrespective of the measure, cut-off value RQ3 and the automated model (Feed-Forward Gender Detector or inaSpeechSegmenter). With respect to the RQ3, it is also examined whether different cut-off values affect the existence of bias, the results are in these tables indicate that both STEM and NON-STEM fields are biased regardless of the cut-off values. Note that both groups of measures in Table II and Table VII show consistent results. Nonetheless, Table VII shows higher MB scores for the STEM field which implies that the inaSpeechSegmenter model produces more probability scores towards the male.

Regarding the RQ4, it is investigated whether different cut-off values change the magnitude of bias – the two tailed paired t-tests applied on MB scores are statistically not significant regardless of the model, see Table III and Table VIII. Unlike the MB scores, the two tailed paired t-tests applied on MAB scores show some statistically significant results both for the Feed-Forward Gender Detector and inaSpeechSegmenter models. For the first model, in Table IV the bias differences of $\text{Rep}_{\text{prob}@3}$ and $\text{Rep}_{\text{prob}@10}$ and $\text{Rep}_{\text{prob}@3}$ and $\text{Rep}_{\text{prob}@20}$ are statistically significant for the STEM field and the latter shows a higher difference (effect size of 0.866). This indicates that the STEM field, using the measure of $\text{Rep}_{\text{prob}@n}$ shows higher bias in top-3 in comparison to top-10 and top-20 search results and the difference is even bigger between top-3 and top-20. For the NON-STEM field, the bias difference of only $\text{Rep}_{\text{prob}@3}$ and $\text{Rep}_{\text{prob}@20}$ is statistically significant with a lower difference in magnitude (effect size is 0.681) than the STEM field. Similarly, the bias differences of $\text{Exp}_{\text{prob}@3}$ and $\text{Exp}_{\text{prob}@10}$, $\text{Exp}_{\text{prob}@3}$ and $\text{Exp}_{\text{prob}@20}$ are statistically significant for the STEM field and the latter shows a higher difference (effect size of 0.857). Unlike the $\text{Rep}_{\text{prob}@n}$, for the NON-STEM field, the bias difference of both $\text{Exp}_{\text{prob}@10}$ and $\text{Exp}_{\text{prob}@20}$, $\text{Exp}_{\text{prob}@3}$ and $\text{Exp}_{\text{prob}@20}$ are statistically significant with lower differences in magnitude (effect sizes are 0.291 and 0.621) than the STEM field.

In addition to these, the inaSpeechSegmenter model shows similar results with respect to the RQ4. For this model, in

### TABLE XIII: Annotation Model Analysis: Perceived gender bias in YouTube for the top-20 relevant results where $n = 3$, 10, 20, p-values of a two-tailed paired t-test computed between the MB and MAB scores of NON-STEM field

|        | $\text{Rep}_{@3}$ | $\text{Rep}_{@10}$ | $\text{Rep}_{@20}$ | $\text{Exp}_{@3}$ | $\text{Exp}_{@10}$ | $\text{Exp}_{@20}$ |
|--------|------------------|--------------------|--------------------|-------------------|-------------------|-------------------|
| MB     | $0.2209^{***}$   | $0.2225^{***}$     | $0.1868^{***}$     | $0.2424^{***}$    | $0.2243^{***}$    | $0.2050^{***}$    |
| p-value | .37              | .79                | .05                | .25               | .76               | .36               |
| effect size d | 0.069           | -0.020             | -0.132             | 0.097             | 0.023             | -0.066            |
| MAB    | $0.4893^{***}$   | $0.3677^{***}$     | $0.3139^{***}$     | $0.5026^{***}$    | $0.4028^{***}$    | $0.3425^{***}$    |
| p-value | .5350^{***}      | $0.3849^{***}$     | $0.3454^{***}$     | $0.5462^{***}$    | $0.4181^{***}$    | $0.3630^{***}$    |
| effect size d | -0.148          | -0.072             | -0.136             | -0.139            | -0.066            | -0.090            |
Table IV shows the bias differences of $R_{\text{prob} @3}$ and $R_{\text{prob} @10}$ and $R_{\text{prob} @3}$ and $R_{\text{prob} @20}$ are statistically significant for the STEM field and again the latter shows a higher difference (effect size of 0.656), yet lower than Feed-Forward Gender Detector. This indicates that the STEM field shows higher bias in top-3 in comparison to top-10 and top-20 search results and the difference is even bigger between top-3 and top-20. For the NON-STEM field, the bias difference of only $R_{\text{prob} @3}$ and $R_{\text{prob} @20}$ is statistically significant with a slightly higher difference in magnitude (effect size is 0.658 instead of 0.656) than the STEM field, yet lower than the Feed-Forward Gender Detector. Similarly, the bias differences of $E_{\text{prob} @3}$ and $E_{\text{prob} @10}$, $E_{\text{prob} @3}$ and $E_{\text{prob} @20}$ are statistically significant for the STEM field and the latter shows a higher difference (effect size of 0.601), yet lower than the Feed-Forward Gender Detector (effect size of 0.857). Similar to the $R_{\text{prob} @n}$ for the NON-STEM field, the bias difference of both $E_{\text{prob} @3}$ and $E_{\text{prob} @10}$, $E_{\text{prob} @3}$ and $E_{\text{prob} @20}$ are statistically significant with comparable differences in magnitude (effect sizes are 0.448 and 0.636) than the STEM field. Unlike the Feed-Forward Gender De-
Fig. 4: $\Delta \Rep_{n}^{\text{prob}}$ measured on perceived gender probability scores of the \texttt{inaSpeechSegmenter}, where x-axis denotes $n = 3$ and y-axis $n = 10$.

Fig. 5: $\Delta \Exp_{n}^{\text{prob}}$ measured on perceived gender probability scores of the \texttt{inaSpeechSegmenter}, where x-axis denotes $n = 3$ and y-axis $n = 10$.

tector, the bias difference in $\Exp_{10}^{\text{prob}}$ and $\Exp_{20}^{\text{prob}}$ is statistically not significant using the \texttt{inaSpeechSegmenter} model.

Regarding the RQ5, the source of bias is tracked to check whether it comes from the input data or the ranking algorithm – the two tailed paired t-tests applied on MB scores are statistically not significant regardless of the model, see Table [V] and Table [X]. Unlike the MB scores, the two tailed paired t-tests applied on MAB scores show some statistically significant results both for the Feed-Forward Gender Detector and \texttt{inaSpeechSegmenter} models. For the first model, in Table [V] the bias difference of only $R_{3}^{\text{prob}}$ and $R_{200}^{\text{prob}}$ is statistically significant for the STEM field with an effect size of 0.954 (noticeable difference in terms of magnitude). Unlike the STEM, for the NON-STEM field, all the bias differences using the MABs are statistically significant and the highest difference of bias in magnitude is between $R_{3}^{\text{prob}}$ and $R_{200}^{\text{prob}}$ with an effect size of 0.991 – higher than the STEM field as well. For the second model, in Table [X] again only the bias difference of $R_{3}^{\text{prob}}$ and $R_{200}^{\text{prob}}$ is statistically significant for the STEM field with an effect size of 0.522 that is lower than Feed-
Forward Gender Detector. On the other hand, unlike the Feed-Forward Gender Detector model for the NON-STEM field, only the bias difference of $R_{ep\text{prob}@3}$ and $R_{ep\text{prob}@200}$ is statistically significant with an effect size of 0.830 that is lower. With respect to the RQ5, although both top results and the representative of the full corpus (which is the top-200 video search results) show bias, the magnitude of bias in the top results is higher than the full list. Thus, it can be inferred that the source of bias does not only come from the input data which is the indexed videos in the context of this chapter, but also from the ranking algorithm since in the top results there is a higher magnitude of perceived gender bias. In addition, the Feed-Forward Gender Detector model shows higher differences both for STEM and NON-STEM fields and for the NON-STEM field, not only the top-3 but also the results in top-10 and top-20 show high bias differences. Based on these findings, it seems that the ranking algorithm could also be blamed for the perceived gender bias results in online education using YVRPs.

Table VI and Table XI show the bias results for the STEM and NON-STEM majors. STEM majors indicate higher scores with the inaSpeechSegmenter model. The most biased STEM majors towards the male gender are Maths and Physics. The most biased NON-STEM majors towards the male gender is Politics and Psychology for the Feed-Forward Gender Detector model and only politics for the inaSpeechSegmenter. Sociology provides some negative scores for both models - biased towards the female. In Figure 1 the results show that both STEM and NON-STEM are overall biased towards the male (positive mean scores) but STEM is more biased both for the Feed-Forward Gender Detector and inaSpeechSegmenter models. Moreover, STEM shows higher bias using the inaSpeechSegmenter model and the difference between STEM and NON-STEM fields is also higher for it. This finding is also consistent with the aforementioned implication that the inaSpeechSegmenter model produces higher probability scores for the male gender. The error bar of both STEM and NON-STEM fields are higher for the Feed-Forward Gender Detector model. Also, the error bar of the NON-STEM is higher than the STEM field. Figure 2 displays the bias scores of the measure, $\Delta R_{ep\text{prob}@n}$, for the STEM and NON-STEM fields using the Feed-Forward Gender Detector model, while Figure 4 using the inaSpeechSegmenter model. STEM field is more biased towards the male, i.e. more bubble points are in the upper-right quadrant, than the NON-STEM field using the inaSpeechSegmenter model. Same holds for the measure, $\Delta E_{xp\text{prob}@n}$, see Figure 3 and Figure 5. These observations are also consistent with the previous conclusions that the inaSpeechSegmenter model produces higher probability values for the male gender – favouring the male gender over female, please refer to the bias scores of STEM field in Table III and Table VIII. For the NON-STEM field, the bubble points are only more dispersed with the inaSpeechSegmenter model. In Figure 4 and Figure 5, STEM is more biased than the NON-STEM towards the male – for the NON-STEM field there is no strong bias towards a specific gender group.

The results in Table XII indicate that the STEM field displays a higher bias when the probability distributions scores are taken from the inaSpeechSegmenter. This is also consistent with the results in Figure 1. On the other hand, for the NON-STEM field there is no statistically significant bias difference based on the results in Table XIII.

REFERENCES

[1] G. Gezici, A. Lipani, Y. Saygin, and E. Yilmaz, “Evaluation metrics for measuring bias in search engine results,” Information Retrieval Journal, vol. 24, no. 2, pp. 85–113, 2021.
[2] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ser. KDD ’16. New York, NY, USA: ACM, 2016, pp. 785–794. [Online]. Available: http://doi.acm.org/10.1145/2939672.2939785
[3] A. Liaw and M. Wiener, “Classification and regression by randomforest,” R News, vol. 2, no. 3, pp. 18–22, 2002. [Online]. Available: https://CRAN.R-project.org/doc/Rnews/
[4] G. Gezici, “Automated perceived gender bias pipeline in youtube,” arXiv preprint arXiv:2210.10517, 2022.
[5] G. Gezici and Y. Saygin, “Measuring gender bias in educational videos: A case study on youtube,” arXiv preprint arXiv:2206.09987, 2022.