

Balancing Fairness and Accuracy in Sentiment Detection using Multiple Black Box Models

Abdulaziz A. Almuzaini
Department of Computer Science
Rutgers University, New Brunswick, NJ
abdulaziz.almuzaini@rutgers.edu

Vivek K. Singh
School of Communication and Information
Rutgers University, New Brunswick, NJ
v.singh@rutgers.edu

ABSTRACT

Sentiment detection is an important building block for multiple information retrieval tasks such as product recommendation, cyber-bullying, fake news and misinformation detection. Unsurprisingly, multiple commercial APIs, each with different levels of accuracy and fairness, are now publicly available for sentiment detection. Users can easily incorporate these APIs in their applications. While combining inputs from multiple modalities or black-box models for increasing accuracy is commonly studied in multimedia computing literature, there has been little work on combining different modalities for increasing *fairness* of the resulting decision. In this work, we audit multiple commercial sentiment detection APIs for the gender bias in two-actor news headlines settings and report on the level of bias observed. Next, we propose a "Flexible Fair Regression" approach, which ensures satisfactory accuracy and fairness by jointly learning from multiple black-box models. The results pave way for fair yet accurate sentiment detectors for multiple applications.

CCS CONCEPTS

• **Computing methodologies** → **Ensemble methods; Regularization.**

KEYWORDS

sentiment analysis, black-box models, fairness, fusion

ACM Reference Format:

Abdulaziz A. Almuzaini and Vivek K. Singh. 2020. Balancing Fairness and Accuracy in Sentiment Detection using Multiple Black Box Models. In *Proceedings of the 2nd International Workshop on Fairness: Accountability, Transparency, and Ethics in MultiMedia (FATE/MM'20), October 12, 2020, Seattle, WA, USA*. ACM, New York, NY, USA, 7 pages. <https://doi.org/10.1145/3422841.3423536>

1 INTRODUCTION

There is an increasing concern amongst researchers that machine learning models developed with the best of intentions may exhibit biases, promote inequality, or perform unfairly for unprivileged groups. With the increasing usage of such models, a considerable amount of research has been conducted to recognize and address

these issues and their social impact [18, 31]. When models show signs of bias, then these models are referred as "unfair" which results in adverse effects on different groups.

Sentiment detection is an important building block for multiple applications such as content moderation, product recommendation, misinformation detection and recently natural language generation [14, 22, 24, 33]. To build these models, large training corpora from varied sources are used. Unfortunately, training data might already contain some bias and that bias can transmit to the learning phase; thus unprivileged groups get unfairly impacted. While the developers of such models often assess success by measuring the accuracy factor, a few have examined the fairness aspect.

Due to the aforementioned popularity, large companies such as Google¹, Amazon² and IBM³ provide users with black-box models that can be easily incorporated into any applications to provide a sentiment score for a given text. Although sentiment detection plays a significant role in such tasks, potential discriminatory treatments might exist for different populations and since these providers are normally perceived as trustworthy, a severe impact could hurt large populations [21]. For example, when a model often predicts a text snippet as toxic or abusive when a female pronoun is present but fails to do so for the male pronoun, this could amplify stereotypes, disenfranchise certain groups, and yield systemic misogyny (or conversely misandry) [30].

An important reason for success in multimedia computing is having multiple models (often involving different datasets) combined together to achieve better results. This approach has been widely used for combining multiple weak forecast models, such as in weather forecast. Similarly, for machine learning, "ensemble learning" and "multi-modal fusion" methods show astonishing results in terms of co-learning and accuracy enhancement [2, 19]. Recently, ensemble deep learning models have been utilized for various applications such as text, image, video, and speech recognition [8, 17].

Recognizing the importance of fairness in such applications, multiple researchers have proposed various metrics and bias mitigation methods [1, 18, 20, 31]. Many such methods either massage the incoming data (pre-processing approaches) or change the optimization parameters within the white-box machine learning model (in-processing) [6]. As both of these are not easily possible in multiple applications (e.g., news sentiment detection using Google API), we focus on a post-processing approach for combining the results from multiple and independent black box APIs (which we also refer to as "modalities" in this work).

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 ACM 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.

FATE/MM'20, October 12, 2020, Seattle, WA, USA

© 2020 Association for Computing Machinery.

ACM ISBN 978-1-4503-8148-2/20/10...\$15.00

<https://doi.org/10.1145/3422841.3423536>

¹Google Cloud <https://cloud.google.com/natural-language>

²Amazon Comprehend <https://aws.amazon.com/comprehend/>

³IBM Watson <https://www.ibm.com/watson>

In this paper, we examine the fairness aspect of three popular sentiment detection black-box models on crime news headlines in which misogynistic and/or misandristic bias might exist and we propose a method to mitigate this bias by combining the outputs from different models (black-box models). Specifically, we examine the sentiment detection across "**gendered interaction**" in news, such as "*a woman hurts a man in a bus*" vs "*a man hurts a woman in a bus*". Here, "gendered interaction" indicates that there are two actors both with clearly identified gender, and there is an interaction (action) taking place between them. In such settings, a model that produces a positive score for the first sentence but a negative score for the second is considered biased towards women (perpetrators) and unfair for men (victims) and vice versa. We apply these tests on crime news headlines dataset that has been collected specifically for this study. Experimental results show each of the publicly available APIs has inherent gender bias and also inaccuracies. On the positive side, the proposed "Flexible Fair Regression" approach was found to be useful to ameliorate both fairness and accuracy concerns.

Our main contributions in this paper are:

- To examine the fairness aspect of publicly available sentiment detection APIs that have been used extensively in various applications. We report that each of these models have inherent gender bias.
- To propose an optimization method "Flexible Fair Regression" to easily allow balancing between bias and accuracy when combining the outputs from multiple (semi-accurate and semi-biased) black box models.
- To share the newly created approach and resulting dataset for quantifying bias in "gendered interaction" scenarios ⁴.

Note that we consider the use of binary gender as a limitation of this work. The use of gender neutral pronouns and those inclusive of non-binary identities is still not common enough in news headlines and hence the problem of bias with binary gendered pronouns remains an important challenge [3, 20].

The rest of the paper is organized as follows. Section 2 provides an overview of the related work in bias detection and mitigation strategies that have been proposed. Then, Section 3 describes the proposed methods including the bias measurement approach and the bias mitigation method. The experimental setup including the data collection process and results are provided in Sections 4 and 5. Finally, in Section 6, a summary and future directions are shared.

2 RELATED WORK

There is significant research work devoted to fairness in algorithms and they can be divided into two general categories: (1) bias measurement, and (2) bias mitigation.

For the first category, a commonly used strategy is to compare the treatment differences for different groups [16, 25]. Multiple efforts use the approach of "word-swapping" where different sensitive variables are swapped in a fixed context. For example, in [22] authors use a template "*I hate <identity> people*" by replacing the sensitive variable with different identities such as (gay, Jewish, African, etc) to examine the models outputs for each identities. This process has been also used to measure bias in sentiment detection

[16, 28], coreference resolution [26] and recently in natural language models [13]. This technique provides a practical approach to examine the treatment for different sentences when a sensitive variable plays an important role in forming the sentence. Another set of metrics that have been widely used for classification tasks are measuring the treatment differences of accuracy, True Positive Rate (TPR), False Positive Rate (FPR), False Negative Rate (FNR) and so on for different groups [9, 22]. Unfortunately, these metrics are only applicable for classification tasks when the output is categorical, whereas in sentiment detection the output value is typically a continuous score, therefore different statistical measures must be utilized to examine unfair treatment.

Multiple studies have examined unfair treatments in settings in which a single sensitive variable exists in a text such as ("I hate women", or "Jews are bad"). In these settings, the sensitive variables are genders and identities, respectively and there is a single actor (victim) present in a text. Additionally, Rudinger et al., [26] in coreference resolution examine the bias of an inferred pronoun in a text that has (two actors) but no interaction. However, none have studied the bias issue in sentiment detection in cases where there are two different actors involved in a direct interaction.

For the second category, bias mitigation strategies can be used in different levels: pre-processing, in-processing and post-processing. Calmon et al., [6] propose a de-biasing method which uses a probabilistic transformation that edits the features and labels in the data with group fairness. Another pre-processing method presented by Zemel et al., [34] focuses on learning a fair representation technique that finds a latent representation which encodes the data well but obfuscates information about sensitive attributes which results in a fair model. Kamishima et al., [15] propose an in-processing algorithm known as Prejudice Remover to decrease bias by adding a discrimination-aware regularization term to the learning objective function. Celis et al., [7] put forward the idea of a meta fair classifier that takes fairness metric as part of the input and returns a classifier optimized with respect to a fairness metric. Other efforts such as [11, 23] bring forward the idea of calibrated equalized odds which is a post-processing technique that optimizes over calibrated classifier score.

Past work on multimodal fusion has proposed methods to enhance the accuracy by combining multiple modalities in data-levels, feature-levels or the decision-levels [2, 19]. These methods are also feasible in tasks when dealing with different and independent black-box models that have different levels of accuracy and different level of bias as well. Although these approaches have shown great success in the past, they are yet to be studied with the goal of fairness enhancement. Fairness in multimedia computing is a relatively nascent but fast-growing [1, 29] field and this work helps motivate and ground the need for fairness considerations in fusion research.

The approach in this paper is inspired by multi-modal fusion proposals for co-learning with weak learners to enhance the accuracy [2, 8, 17, 19]. It also adapts the in-processing technique for bias mitigation (fairness enhancement) using a fairness regularizer incorporated in the objective function to create a post-processing approach which can work well with multiple black box models in sentiment detection case.

⁴https://github.com/abdulazizasz/fairness_sentiment

3 METHODOLOGY

In this section, we define the problem settings and describe the proposed methods.

3.1 Preliminaries

We formulate the problem of fair sentiment detection as the following: we have k independent black-box models and their sentiment scores $x_k \in [-1, 1]$ and a ground truth score $y \in [-1, 1]$. Besides that, let there be a sensitive variable S that can divide the dataset $D = \{x_i, y_i\}_{i=1}^N$ into different groups, e.g., S_{male} , S_{female} . To simplify the settings, we can combine multiple k column vectors modalities $\{x_1, x_2, \dots, x_k\}$ in a matrix X .

In such a setting the goal of the algorithm is to minimize the loss as measured via a combination of accuracy error, bias penalty, and over-fitting penalty.

We describe the operationalization of the different terms above in the following subsections.

3.2 Measuring Bias

To measure the bias/fairness, we examine the mean difference of the scores for each black-box models with respect to a sensitive variable S , e.g., for a binary variable S^+ , S^- as follows:

$$\text{Mean Difference}(x, k) = \frac{\sum_{x_i \in S^+} x_i^k}{|S^+|} - \frac{\sum_{x_i \in S^-} x_i^k}{|S^-|}$$

A fair black-box model will result in a zero score which means the same sentiment score has been produced regardless of the sensitive variable; in other words different groups have been treated equally. Since we are focusing on the scores deviation among different groups we can use the Mean Absolute Difference/Error (MAE) to measure bias in a given model. Other methods proposed in the literature for measuring the bias can also be useful candidates (e.g. Correlation and delta of prediction accuracy [5]) in other settings.

3.3 Balancing Accuracy and Fairness

In this project, we use a regression model to find the best parameter to combine these independent and different black-box models X with respect to the target scores y . The regression model can be formulated as an optimization problem to find the parameter w that minimizes the following function:

$$\underset{w}{\text{minimize}} \quad \text{MSE}(w) = \frac{1}{N} \sum_{i=1}^N (w * x_i - y_i)^2 \quad (1)$$

Many variants of linear regression add a regularizer function to the regression model, which keeps a check on the number of parameters being used in the modeling [12]. Further, recent efforts have proposed adding a "fairness regularizer" to the regression [4]. Here, we adapt that approach to define a fair regularizer which penalizes the function when the sentiment scores differ between different groups; in other words when the model exhibits some notion of bias. For a binary variable, a model regularizes the difference between S^+ and S^- .

Thus, the modified objective function is trying to find the optimal parameter w that minimizes the Mean Squared Error (MSE) along

with the minimum bias between different groups present in the dataset.

To simplify the objective function, a bias matrix Δ which contains the sentiment score difference for each modality is calculated. For instance, for a modality k and a binary sensitive variable S , the bias vector can be calculated as follows:

$$\delta_k = |x_{S^+}^k - x_{S^-}^k|$$

Similarly, calculating the bias for other modalities and combining them in one matrix Δ yields:

$$\Delta = \begin{bmatrix} \delta_1^1 & \dots & \delta_1^k \\ \vdots & & \vdots \\ \delta_i^1 & \dots & \delta_i^k \\ \vdots & & \vdots \\ \delta_n^1 & \dots & \delta_n^k \end{bmatrix}$$

Using the w and Δ the "fairness" penalty function P is:

$$P(w) = \frac{1}{N} \sum_{i=1}^N (w * \delta_i)^2$$

Now the optimization function for "Flexible Fair Regression" is:

$$\underset{w}{\text{minimize}} \quad L(w) = \text{MSE}(w) + \beta P(w) + \lambda \|w\|^2 \quad (2)$$

where $\|w\|^2$ is the L_2 norm. Hyper parameters λ and β are used to control over-fitting and the fairness trade-off.

We call this approach "flexible fair regression" as it supports fairness in regression, and allows a system designer to flexibly pick the relative importance and thresholds that they want to balance the fairness and accuracy trade-off.

4 EXPERIMENTAL SETUP

In this section, we describe the process of the data-collection, annotations, baselines, and the bias-reduction results.

4.1 Dataset

To construct a dataset, we collected crime news headlines from Google News API ⁵. To do so, we used the API search criteria to retrieve only the news headlines that contain abusive verbs such as (*kill*, *murder*, *slap*, etc) along with at least two different subjects (*man*, *woman*) involved in an interaction. Using a carefully designed list of abusive verbs provided by Wiegand et al., [32], we collect a large number of data-points (crime news headlines) and we filter out the results in which there is no interaction between subjects.

Since in this project we are tackling the predictive learning problem as a regression, we need to label and assign a sentiment score to the collected dataset. Utilizing Figure-Eight ⁶, we asked 10 annotators to label and score each sentence (template) and to avoid bias in the annotation process, we anonymize the subjects in the sentence. Then, every annotator is presented with an anonymized template (see Fig. 1) and is asked to provide two pieces of information following the Valence-Arousal model [27]: (1) A valence label to a sentence such as "Positive or Negative" (2) An arousal score on a scale from 1 to 10. Thus, a sentence with a positive label and a score of 10 will have a sentiment score close to +1 whereas a sentence with a negative label and an arousal score of 5 will have

⁵<https://news.google.com>

⁶www.figure-eight.com

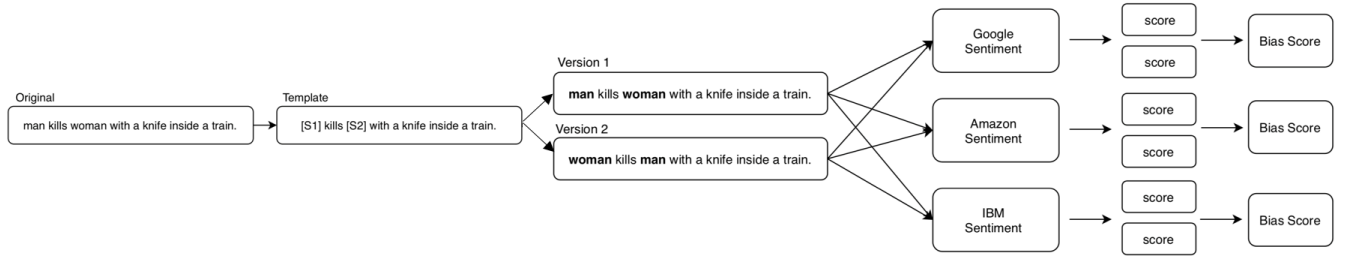


Figure 1: Constructing a template from the original news headlines, applying the gender-swap to create the two versions and getting the black-box models scores

a sentiment score -0.5. This process yielded scores in the range of $[-1, 1]$ and thus allows us to assign a ground-truth scores for each template.

Removing inconsistent and untrustworthy annotators, and using a seed of 200 templates, we did gender swapping in the sentences where the first sentence has men as perpetrators and women as victims and vice-versa. Also by applying different gender identities such as ("man-woman", "male-female", etc), we used 25 of such terms provided by [10] in our dataset. Thus, our corpus contains 10000 news headline sentences. The resulting dataset was then scored using Google, Amazon, and IBM APIs (see Fig. 1). Finally, we split the dataset into training and testing sets in a ratio of 70:30.

Table 1: A sample from the dataset for the two versions

Sentence	k_1	k_2	k_3	S	y
man hurts woman in ..	-0.9	-0.5	0.6	<i>m</i>	-0.7
woman hurts man in ..	-0.7	-0.8	-0.9	<i>f</i>	-0.7

A sample of a dataset is shown in Table 1. For each black box model, we have a sentiment score for the two versions that have been constructed from the template along with the gender of the perpetrator and the ground-truth score. From Table 1, we get the scores for each template by taking the average between the two versions since we are aiming at finding the optimal score for each template regardless of the presence of gender or the order of perpetrators-victims in a text; therefore the template is anonymized (see Table 2).

Table 2: Training dataset for each template regardless of gender. S1 and S2 stands for subject 1 and 2, respectively

Sentence	k_1	k_2	k_3	y
"[S1] hurts [S2] in ..	-0.8	-0.65	0.15	-0.7

4.2 Baselines

Since the scores that are generated from these black-box models are not always accurate and sometimes noisy, a fusion process is used to increase the accuracy by combining different modalities. Related

works (e.g., [2, 19]) provide various methods that are practical for fusing independent and weak modalities. In this project we are experimenting with three such methods:

Unweighted Average is the basic fusion process that assumes that independent modalities are equal in terms of accuracy. Thus, the predicted sentiment score for each template is calculated by a simple average:

$$\hat{y}_i = \frac{1}{N} (x_i^{\text{google}} + x_i^{\text{amazon}} + x_i^{\text{ibm}})$$

where N is the number of training data-points.

Weighted Average on the other hand weights each modality based on its historical accuracy (for the training set) an accurate models deserves a higher weight and vice-versa:

$$\hat{y}_i = w_{\text{Google}} * x_i^{\text{google}} + w_{\text{Amazon}} * x_i^{\text{amazon}} + w_{\text{IBM}} * x_i^{\text{ibm}}$$

where $\sum_{k=1}^3 w_k = 1$.

Multiple Regression is similar to our proposed method but it does consider the "fairness" penalty term in the learning phase. In this approach, black-box models outputs can be treated as independent features for a learning model.

All of the above methods are only used to optimize for the accuracy part and not considering the fairness aspect when using a fusion process.

Fairness Optimization is an additional baseline that optimizes (only) for fairness which weight each modality by its historical fairness scores in the training data. The "fairness" weight for a modality k is calculated as follows:

$$w^k = \sum_{i=1}^N \mathbb{1}_{\{|x_{\text{male}_i}^k - x_{\text{female}_i}^k| \leq \tau\}} \quad (3)$$

Here, if the sentiment scores differ by less than $\tau = 10\%$, we consider it as a fair treatment for that template. This baselines will help in determining the best parameter for the learning model.

The other baselines we use are fitting linear regression for each modality separately and by doing so, we can examine the fairness and accuracy separately and the trade-off among other methods.

5 RESULTS AND DISCUSSION

5.1 Auditing Sentiment Detection APIs

As a first step, we analyzed the results from the different black box models (sentiment detection APIs) to see if they are accurate and if there is a difference in the results obtained for sentences which are same, except for the genders represented for the perpetrators and the victims.

Table 3: Accuracy and Fairness in the original dataset

API	Acc. Error	Bias
Google	0.5611	0.0590
Amazon	0.6939	0.0581
IBM	0.7441	0.0545

To evaluate accuracy (compared to ground truth labels and scores obtained from multiple human labelers), we use Root Mean Squared Error (RMSE). To measure bias/fairness, we use Mean Absolute Error (MAE) metric for different groups. Table 3 shows the errors in accuracy and bias levels for each individual modality. The range for sentiment scores is $[-1, 1]$ and we see that there are noteworthy accuracy and bias issues with each modality.

Further, a pairwise t-test comparing the mean sentiment scores across genders yielded statistically significant differences in all three modalities. These issues motivate the need to combine the outputs from different modalities to improve both accuracy and fairness.

5.2 Improving Accuracy and Fairness

We implemented the Flexible Fair Regression approach (Eq. 2) on the created dataset using Python and Scipy library for optimization.

Table 4 provides the summary of the results. We can easily see the trade-off between the accuracy error and bias among these models (lower is better in both cases). Multiple Regression is performing well in terms of accuracy (Low RMSE) but has to contend with higher value for bias. Both Weighted Average and Unweighted Average methods yield higher errors in terms of accuracy than Multiple Regression but yield lower levels of bias.

Table 4: Mitigation Methods Analysis

Model Name	Acc. Error	Bias
Multiple Regression	0.5362	0.0738
Unweighted Average	0.6302	0.0435
Weighted Average	0.6153	0.0447
Fairness Optimization	0.7051	0.0173
Our Method	0.6026	0.0400

Note that it is also possible to optimize only for fairness (Fairness Optimization), and this can reduce the bias to very low value (close to zero). However, this comes with the price of a high accuracy error (see Fig. 2).

Lastly, our method allows for flexible trade-off between accuracy and fairness (see Eq. 2). The trade-off depends on the choice of

weight parameter β and different values of β yield different points on the purple curve (search space) in Fig. 3. The axes for the figure are accuracy error and bias levels. Lower is better for both of them and hence, the points on the lower left corner are ideal. As shown, all the points of the purple curve (our approach with different β values) either coincide with other baselines or strictly dominate them (i.e., yield better performance in terms of both accuracy and fairness).

The Multiple Regression and Fairness Optimization approaches can be considered extreme cases of our flexible fair regression approach, such that they optimize only for accuracy or only for fairness. Any point on the purple curve (varying values of β) will provide a trade-off between these two extremes. To find a specific suitable candidate for β value, we use Figure 3 to jointly consider the achievements of Fairness Optimization (FO) and Multiple Regression (MR). Hence, an ideal solution ("Utopia Point" marked as "X" in Fig. 3; unlikely to be achievable in practice), will yield accuracy error as low as MR and bias as low as FO (see Fig. 3). Hence, we consider the point closest (minimum distance) to the "Utopia Point" to be used as a suitable candidate to pick the β value. In the current work, β value of 0.002 is the closest point, which yields the results shown in Table 4 and Figure 2. Note that this result is pareto-optimal in the sense that there is no other feasible point that is lower in *both* accuracy error and bias. This can be seen from the points in Fig. 3 (which also includes points for models which use just one modality) and Table 4.

Another possible approach to joint optimization of two factors is to budget a fixed "cost of fairness" [4]. For example, losing a portion of accuracy could lead to a gain in fairness. In Figure 4, we provide an illustration of trade-off; a $\sim 10\%$ accuracy loss could yield a $\sim 38\%$ reduction in bias. Hence, a practitioner with an assigned 10% accuracy budget could gain up to 38% in terms of fairness. Other plausible budgets and impact can also be easily computed using this approach to allow for such decision making.

6 CONCLUSION

In this work, we deploy a regularized objective function that combines independent and different black box models to ensure an accurate and a fair learning model for sentiment detection.

Since we are dealing with disparate and independent black-box models, a fusion process helps combine multiple sources' results and build a more robust score for each template regardless of the subjects' genders. The proposed approach yields a family of pareto-optimal solutions compared to other baseline approaches. Further, our "fairness" penalty function performs well in terms of bias reduction and is more flexible compared to other baselines.

An important limitation of this work is the focus on binary genders which rarely exist in news headlines. Another critical challenge is that of constructing large and practical dataset, i.e., templates that cover enough amount of context in which various abusive verbs and identities can occur in news headlines. Additionally, the annotation process might result in biased scores; annotators have subjective opinion and that might result in a biased ground-truth labels. Thus, to mitigate this bias, we take the average of the scores from different annotators and remove inconsistent and untrustworthy annotators. Lastly, since we are solving an optimization

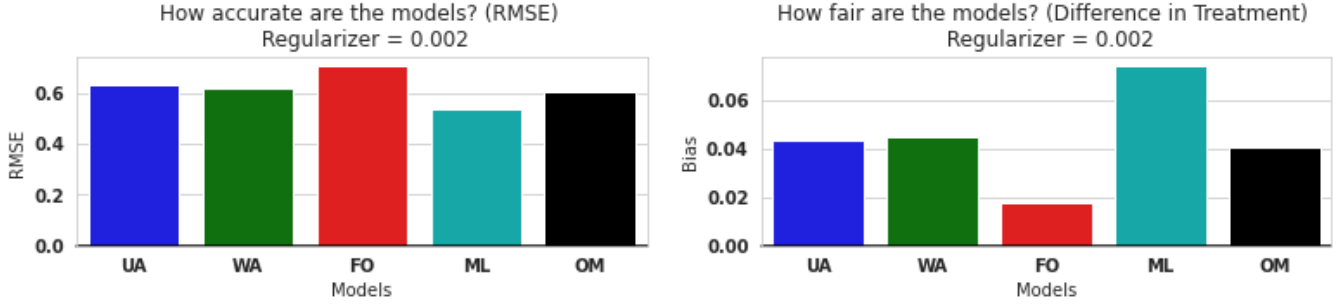


Figure 2: The accuracy error (RMSE) and the bias score (MAE) is shown for different baselines methods along with our method. The lower the bar the better. "UA: Unweighted Average, WA: Weighted Average, FO: Fairness Optimization, ML: Multiple Regression, OM: Our Method"

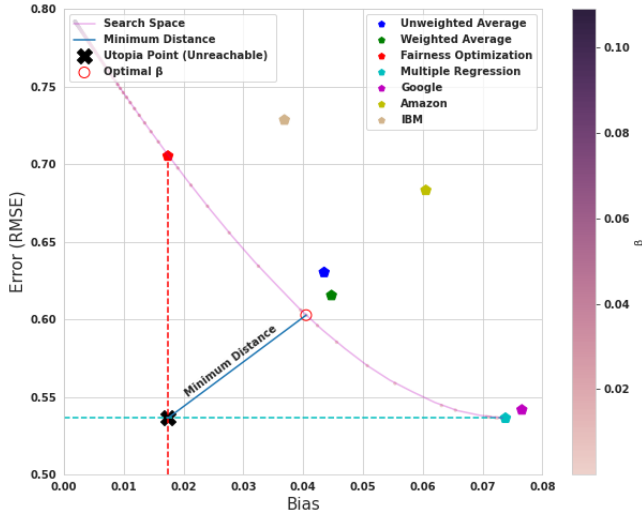


Figure 3: Using the intersected point, we choose the closest β which is $\beta = .002$ in Our Method

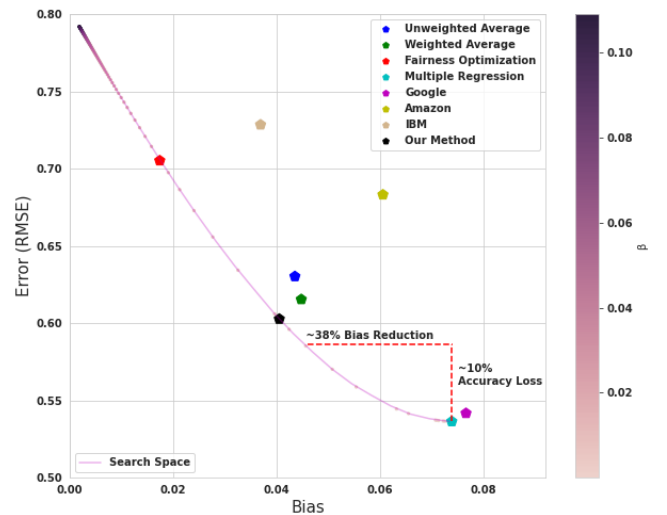


Figure 4: Trade-off between accuracy and fairness by manipulating the parameter τ (e.g., $\tau = 10\%$)

function, a convexity of equation is assumed and a solver has been used to find the optimal minima.

Despite the limitations, this paper marks a significant step forward toward fairness and accuracy in sentiment detection literature. The paper advances the fairness literature to consider multiple actor "gendered interactions", which has use cases in news analysis, abuse detection, and misinformation detection.

The public dataset and the proposed flexible approach can allow for fairness in a wide variety of scenarios where semi-accurate and semi-fair black box models need to be combined to obtain fair yet accurate predictions.

Future studies could incorporate additional APIs in the model and investigate the concept of associations that might exist between different black-box models. Initial experiment shows a sign of correlation between these different APIs that might have been caused by different factors such as using shared datasets or adapting similar learning algorithms; thus a more sophisticated fusion method can be developed to enhance both accuracy and fairness in this regard.

In addition, extending the dataset to contain more diverse structure can help us examine the underlying bias in these modalities and the factors that led to the disparate treatment for different groups. Finally, applying the same work on different machine learning applications such face-matching and misinformation might result in valuable results.

ACKNOWLEDGMENTS

We thank Ruoyuan Gao for useful and interesting discussion on this work and providing insightful advice on performing the annotation process for such task.

This material is in part based upon work supported by the National Science Foundation under Grant SES-1915790.

REFERENCES

- [1] Jamal Alasadi, Ahmed Al Hilli, and Vivek K Singh. 2019. Toward Fairness in Face Matching Algorithms. In *Proceedings of the 1st International Workshop on Fairness, Accountability, and Transparency in MultiMedia*. 19–25.
- [2] Pradeep K Atrey, M Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S Kankanhalli. 2010. Multimodal fusion for multimedia analysis: a survey. *Multi-media systems* 16, 6 (2010), 345–379.
- [3] Pinkesh Badjatiya, Manish Gupta, and Vasudeva Varma. 2019. Stereotypical bias removal for hate speech detection task using knowledge-based generalizations. In *The World Wide Web Conference*. 49–59.
- [4] Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. 2017. A convex framework for fair regression. *arXiv preprint arXiv:1706.02409* (2017).
- [5] Toon Calders, Asim Karim, Faisal Kamiran, Wasif Ali, and Xiangliang Zhang. 2013. Controlling attribute effect in linear regression. In *2013 IEEE 13th international conference on data mining*. IEEE, 71–80.
- [6] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. 2017. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems*. 3992–4001.
- [7] L Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K Vishnoi. 2019. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 319–328.
- [8] Li Deng and John C Platt. 2014. Ensemble deep learning for speech recognition. In *Fifteenth Annual Conference of the International Speech Communication Association*.
- [9] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 67–73.
- [10] Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *arXiv preprint arXiv:1903.03862* (2019).
- [11] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*. 3315–3323.
- [12] Arthur E Hoerl and Robert W Kennard. 1970. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* 12, 1 (1970), 55–67.
- [13] Po-Sen Huang, Huan Zhang, Ray Jiang, Robert Stanforth, Johannes Welbl, Jack Rae, Vishal Maini, Dani Yogatama, and Pushmeet Kohli. 2019. Reducing sentiment bias in language models via counterfactual evaluation. *arXiv preprint arXiv:1911.03064* (2019).
- [14] Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Unintended machine learning biases as social barriers for persons with disabilities. *ACM SIGACCESS Accessibility and Computing* 125 (2020), 1–1.
- [15] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 35–50.
- [16] Svetlana Kiritchenko and Saif M Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. *arXiv preprint arXiv:1805.04508* (2018).
- [17] Inwoong Lee, Doyoung Kim, Seoungyoon Kang, and Sanghoon Lee. 2017. Ensemble deep learning for skeleton-based action recognition using temporal sliding lstm networks. In *Proceedings of the IEEE international conference on computer vision*. 1012–1020.
- [18] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635* (2019).
- [19] Joao Mendes-Moreira, Carlos Soares, Alípio Mário Jorge, and Jorge Freire De Sousa. 2012. Ensemble approaches for regression: A survey. *Acm computing surveys (csur)* 45, 1 (2012), 1–40.
- [20] Debora Nozza, Claudia Volpetti, and Elisabetta Fersini. 2019. Unintended bias in misogyny detection. In *IEEE/WIC/ACM International Conference on Web Intelligence*. 149–155.
- [21] Cathy O’neil. 2016. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books.
- [22] Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. *arXiv preprint arXiv:1808.07231* (2018).
- [23] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. 2017. On fairness and calibration. In *Advances in Neural Information Processing Systems*. 5680–5689.
- [24] Elahieh Raisi and Bert Huang. 2019. Reduced-Bias Co-trained Ensembles for Weakly Supervised Cyberbullying Detection. In *International Conference on Computational Data and Social Networks*. Springer, 293–306.
- [25] Rachel Rudinger, Chandler May, and Benjamin Van Durme. 2017. Social bias in elicited natural language inferences. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*. 74–79.
- [26] Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. *arXiv preprint arXiv:1804.09301* (2018).
- [27] James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology* 39, 6 (1980), 1161.
- [28] Judy Hanwen Shen, Lauren Fratamico, Iyad Rahwan, and Alexander M Rush. 2018. Darling or babygirl? investigating stylistic bias in sentiment analysis. *Proc. of FATML* (2018).
- [29] Vivek K Singh, Elisabeth André, Susanne Boll, Mireille Hildebrandt, and David A Shamma. 2020. Legal and ethical challenges in multimedia research. *IEEE MultiMedia* 27, 2 (2020), 46–54.
- [30] Vivek K Singh, Mary Chayko, Raj Inamdar, and Diana Floegel. 2020. Female Librarians and Male Computer Programmers? Gender Bias in Occupational Images on Digital Media Platforms. *Journal of the Association for Information Science and Technology* (2020).
- [31] Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. *arXiv preprint arXiv:1906.08976* (2019).
- [32] Michael Wiegand, Josef Ruppenhofer, Anna Schmidt, and Clayton Greenberg. 2018. Inducing a lexicon of abusive words—a feature-based approach. (2018).
- [33] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771* (2019).
- [34] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning fair representations. In *International Conference on Machine Learning*. 325–333.