# Beyond the Binary: Detecting Gender Bias Using Explainability

**Gauri Gupta*, Supriti Vijay*, Krithika Ramesh***
Manipal Institute of Technology
MAHE, Manipal, 576104
[gaurigupta.315, supriti.vijay, kramesh.tlw]@gmail.com

## Abstract

Explanations for AI systems have been used to improve the trustworthiness of these systems. These explanations can be used to find the undesirable implicit biases that machine learning models can rely on for their outputs. We apply this concept to detect gender bias in sentiment analysis models for textual data. We employ two existing frameworks: LIME and SHAP, which produce explanations for text classifiers. With the help of an Equity Evaluation Corpus (EEC), we add a feminine, neutral, and masculine gender signals for otherwise identical input to the system and use explanations from LIME and SHAP to find a trend of bias, and identify terms that contribute the most to it.

## 1 Introduction

Language models have increasingly been found to reflect societal biases in their outputs, gender bias being one of them. Gender Bias here refers to when the gender signals in the input affect a model's predictions. Consider a set of three sentences fed into the model, only with different gender signals, such as "He was furious," "She was furious," and "They were furious." All three of these sentences should ideally predict the same final sentiment polarity. However, since the model learns from the training set, it learns undesirable associations between words which we observe as gender bias. Gender bias in language models can be harmful in more than one way, considering their potential impact on downstream tasks which have vast industrial applications.

In recent years, there has been a growing trend of work that focuses on measuring and mitigating gender bias (Bolukbasi et al., 2016; Gonen and Goldberg, 2019; Blodgett et al., 2020; Savoldi et al., 2021). There are limitations to some of these existing approaches, primarily considering that they are only applicable to the 'binary' gender. We aim to fill this gap by proposing the use of explainability to observe and uncover gender bias trends in model outputs and consequently measure the bias. Explainable AI is a field of research that transforms the classifier from a black box of information into a glass box that humans can easily interpret and understand. It helps us visualize the data in the training set by making the underlying contents of the model transparent.

This paper addresses whether explainable AI methods can be used to detect gender bias in textual data. We will be discussing the application of model-agnostic methods on the dataset and how these explanations help uncover the hidden bias.

## 2 Related Work

### 2.1 Gender Bias

Recent studies have shown that in sentiment analysis when models are trained on human-handwritten text, they predict biased outcomes, which may cause harm in the real-world applications of that model. Thus, it is imperative to detect and mitigate these biases to ensure fairness in sentiment analysis systems, as shown by existing studies (Ribeiro et al., 2020; Kiritchenko and Mohammad, 2018).

To measure gender bias, Zhao et al. (2018); Lu et al. (2019); Kiritchenko and Mohammad (2018) proposed gender-swapping, which referred to interchanging each male-defined word with its respective female equivalent and vice versa. An ideal fair model would predict equally for both sentences. However, since the model is biased, the difference in evaluation scores indicates the degree of gender bias detected in the model. Zhao et al. (2017) also proposed a method called Word Embedding Association Test (WEAT) to measure bias inside word embeddings.

Another method proposed was to use standard evaluation data sets to detect the degree of gender bias. However, since they contain more male references than females, we opt for specific data sets called Gender Bias Evaluation Test sets (GBETs)

| Template | #sent |
|---|---|
| <Person>feels <emotional state word>. | 280 |
| The situation makes <person>feel <emotional state word>. | 280 |
| I made <person>feel <emotional state word>. | 280 |
| <Person>made me feel <emotional state word>. | 280 |
| <Person>found himself/herself in a/an <emotional situation word>situation. | 280 |
| <Person>told us all about the recent <emotional situation word>events. | 280 |
| The conversation with <person>was <emotional situation word>. | 280 |
| **Total** | **1960** |

Table 1: Lists of template sentences and the sentence count

(Sun et al., 2019). One such example for GBET used in this paper is the Equity Evaluation Corpus (EEC) proposed by Kiritchenko and Mohammad (2018) for sentiment analysis. It generates test cases produced from 11 handcrafted templates. Each EEC sentence contains an emotional word (e.g., anger, fear, joy, sadness), intensities for each emotion, and a gender-specific term.

### 2.2 Explainable Sentiment Analysis

Sentiment Analysis refers to detecting sentiments and opinions present in textual datasets (Liu, 2012). The need for transparency in such models is imperative to understand why an instance of the textual dataset is attributed to a particular polarity (Zucco et al., 2018; So, 2020; Hase and Bansal, 2020; Silveira et al., 2019).

Explainable AI (XAI) is a field of artificial intelligence that develops explainable methods, which enable transparency and help understand ML models (Miller, 2019). Luo et al. (2016); Amrani et al. (2018); Jabreel and Moreno (2018) suggested using Scikit learn models like Support Vector Machines (SVM) (Steinwart and Christmann, 2008), Random Forest (RF) (Breiman, 2001), and XGB Boost Classifier (Chen and Guestrin, 2016) since they are regarded as good performing models in sentiment analysis. The recent work of Bodria et al. (2020) proposed attention-based techniques to provide explanations and extract meaningful sentiment scores to explore the internal behavior of deep neural network models.

### 2.3 Explainable approaches to measure fairness

Since XAI is a relatively new field, we could only find a limited number of research applying XAI methods to measure gender bias. Domnich and Anbarjafari (2021) implemented neural networks to show how gender bias affected the recognition of emotion. Jain et al. (2020) proposed a framework for evaluating explainable AI tools based on their capacity for detecting bias and fairness. They described statistical notions of fairness in terms of explanations given by the model. Alikhademi et al. (2021) addressed the needs of both XAI and Fair AI tools and proposed features for explaining fairness in ML models and data. They defined a rubric to create and evaluate XAI tools for fairness.

This paper has applied both local explainability for single instances using LIME and global explanations for the entire model using SHAP. We do this to compare the performance of both these explanations to understand gender bias in our dataset.

## 3 Experimental Setup

### 3.1 Equity Evaluation Corpus

To estimate the bias in the model's predictions, we used the Equity Evaluation Corpus (Kiritchenko and Mohammad, 2018), a corpus containing a series of templates of sentences for masculine and feminine genders. We extended this corpus in our work so that it could accommodate gender-neutral terms and omitted the creation of template sentences including terms for which we could not find the gender-neutral equivalent [1]. The total length of the corpus for each of the masculine, feminine and gender-neutral terms was 280 sentences for each of the emotions (anger, joy, fear, sadness), against which the bias was measured.

### 3.2 Models and Architectures Used

BERT, which stands for Bidirectional Encoder Representations from Transformers, uses an attention mechanism to learn the contextual relationship between words. (Devlin et al., 2019) It contains a nondirectional encoder which inputs the entire sequence of words at once. BERT utilizes a masking language model that randomly masks some of the

[1] For example, aunt/uncle

| Masculine | Feminine | Neutral |
|---|---|---|
| He | She | They |
| This man | This woman | This person |
| This boy | This girl | This child |
| My brother | My sister | My child |
| My son | My daughter | My sibling |
| My husband | My wife | My spouse |
| My boyfriend | My girlfriend | My partner |
| My father | My mother | My parent |

Table 2: List of gendered terms

| Anger | Fear | Joy | Sad |
|---|---|---|---|
| anger | anxious | ecstatic | depressed |
| annoyed | discouraged | excited | devastated |
| enraged | fearful | glad | disappointed |
| furious | scared | happy | miserable |
| irritated | terrified | relieved | sad |
| annoying | dreadful | amazing | depressing |
| displeasing | horrifying | funny | gloomy |
| irritating | shocking | great | grim |
| outrageous | terrifying | hilarious | heartbreaking |
| vexing | threatening | wonderful | serious |

Table 3: Emotion sets containing relevant words

tokens from the input to predict the original vocabulary of the masked word based on its contextual relationship with other words in the sentence.

For our experiments, we used XLM-RoBERTa, a popular BERT model, a transformer-based multilingual model trained on 100 different languages. (Conneau et al., 2020) It is a cross-lingual approach that uses both XLM and RoBERTa and can determine the correct language from the input data-id.

### 3.3 LIME

Locally Interpretable Model-Agnostic Explanations or better known as LIME is a framework that explains the predictions of any machine learning classifier (Ribeiro et al., 2016). With explanation defined as a model $g \in G$, where $G$ is a class of potentially interpretable models, the explanation from LIME is obtained with the help of the following:

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \mathcal{L}\left(f, g, \pi_x\right) + \Omega(g) \qquad (1)$$

where $\mathcal{L}(f, g, \pi_x)$ is a measure of how unfaithful $g$ is in approximating $f$ in the locality defined by $\pi_x$, given $f(x)$ is the probability (or a binary indicator) that $x$ belongs to a certain class, $\pi_x(z)$ is a proximity measure between an instance $z$ to $x$ used to define locality around $x$, and $\Omega(g)$ is measure of complexity.

This formulation can be used with different explanation families $G$, fidelity functions $\mathcal{L}$, and complexity measures $\Omega$.

Ribeiro et al. (2016) also extensively discusses how the framework could be used to derive insights on the undesirable correlations that the classifiers pick up during training. An instance of a 'husky' being classified as a 'wolf' was demonstrated as the classification model depending on the background behind the husky rather than the features of the husky itself. This ability of the framework motivates us to understand how a text classifier could depend on seemingly unrelated features of the input to produce an output that is biased.

### 3.4 SHAP

SHAP (SHapley Additive exPlanations) is a unified framework that provides interpretability for model predictions (Lundberg and Lee, 2017). This work introduced the concept of an explanation model, and it builds on the classical Shapely value estimation methods (Ŝtrumbelj and Kononenko, 2013; Lipovetsky and Conklin, 2001; Sen and Zick, 2016) towards additive feature attribution methods for the approximation of SHAP values.

SHAP values are proposed as "a unified measure of feature importance" and identify the contribution of each feature in the input, to the prediction. The Shapely values are obtained from a conditional expectation function of the original model providing the solution to the following theorem:

$$\phi_i(f, x) = \sum_{z' \subseteq x'} \frac{|z'|!(M-|z'|-1)!}{M!} \left[f_x\left(z'\right) - f_x\left(z' \backslash i\right)\right] \qquad (2)$$

where $|z'|$ is the number of non-zero entries in $z'$, and $z' \subseteq x'$ represents all $z'$ vectors where the non-zero entries are a subset of the non-zero entries in $x'$.

This would involve calculating the marginal contribution of every feature by considering all possible permutations (which totals to $N!$ permutations, where $N$ denotes the number of features). We consider the mean of the contributions of all features in order to estimate its aggregate contribution to the prediction of a particular class. In other words, we could say that the Shapely value is an indicator of how much weight a specific feature carries, which is also the difference between the actual prediction and the base prediction measured by the classifier.

## 4 Methodology

### 4.1 LIME

To estimate the bias present in our models, we've averaged the values of the contributions of the gendered features from Table 2 which are depicted in Tables 4, 5, 6 and 7. These contributions were calculated using the ratio of the gendered word's weightage to the probability of the prediction of the actual class. In these tables, we've also attempted to individually specify the number of samples that have a positive correlation and those that have a negative correlation with the ground truth label. Additionally, we have quantified the frequency of samples that have a positive contribution greater than 5% and 10% to the prediction of the correct class.

Although our initial study only takes into account the contributions of each feature, we have also calculated the differences in terms of the contribution of each gendered feature for a specific template sentence, and mapped out comparative distributions of these features, the results of which are included in the Appendix A. These distributions take into consideration only the cases where the gendered term has a positive correlation with the actual class. For instance, sentences using the template 'anger' should ideally be predicted as negative, and thus we only consider the gendered term's positive contribution to the prediction.

A shortcoming of LIME is its local interpretability, where it creates a linear local model around a singular data instance. There can be potential instability caused due to the variation in sampling of the data, which can lead to differences in the explanations for the same sample. In addition to this, the choice of hyperparameters can also affect the explanations produced for the same data sample. Thus the explanations produced, while somewhat accurate, may not necessarily be robust.

### 4.2 SHAP

Due to the disadvantages and uncertainty in LIME predictions, we turned to an alternative method, SHAP, to understand our model's predictions. The Shapely values allow us to gain an understanding of how exactly a particular feature impacts the probability of a particular prediction with respect to the original base prediction. SHAP is significantly more stable and does not rely on the same assumptions that the linear local model of LIME suffers from. In addition to this, it combines predictions at both a global and a local level to give an estimate of the feature's importance.

| Gender Signal | Positive Correlation | Negative Correlation | Freq > 0.05 | Freq > 0.1 |
|---|---|---|---|---|
| Masculine | 124 | 156 | 43 | 5 |
| Feminine | 135 | 145 | 66 | 16 |
| Neutral | 184 | 96 | 94 | 26 |

Table 4: Contributions of gendered terms to the LIME predictions for the template sentences associated with anger

| Gender Signal | Positive Correlation | Negative Correlation | Freq > 0.05 | Freq > 0.1 |
|---|---|---|---|---|
| Masculine | 140 | 140 | 40 | 5 |
| Feminine | 125 | 155 | 37 | 4 |
| Neutral | 87 | 193 | 13 | 1 |

Table 5: Contributions of gendered terms to the LIME predictions for the template sentences associated with joy

| Gender Signal | Positive Correlation | Negative Correlation | Freq > 0.05 | Freq > 0.1 |
|---|---|---|---|---|
| Masculine | 142 | 138 | 58 | 11 |
| Feminine | 164 | 116 | 70 | 19 |
| Neutral | 198 | 82 | 104 | 42 |

Table 6: Contributions of gendered terms to the LIME predictions for the template sentences associated with fear

| Gender Signal | Positive Correlation | Negative Correlation | Freq > 0.05 | Freq > 0.1 |
|---|---|---|---|---|
| Masculine | 128 | 152 | 53 | 17 |
| Feminine | 153 | 127 | 66 | 19 |
| Neutral | 196 | 84 | 104 | 49 |

Table 7: Contributions of gendered terms to the LIME predictions for the template sentences associated with sadness

Using SHAP, we've attempted to calculate the mean prediction value to estimate every gendered feature's influence on the model's prediction. This was done for each of the four emotion sets, and the results are presented in tables 9, 8, 10. We've noticed that while for some emotion sets, there appears to be a general trend in the correlations of the features for a particular gender, there is also a noticeable variation in whether these correlations are positive or negative. This could be an indicator that the bias learned by our models is less generalized for a specified gender than we assume, and thus further analysis using other explanatory algorithms would be required to study if this is the case.

## 5 Results and Discussion

Despite LIME's instability and its variation in predictions, we notice some generalizable trends in the

| Term | Anger | Joy | Fear | Sad |
|---|---|---|---|---|
| she | -0.1734 | 0.0771 | -0.1739 | -0.1516 |
| woman | -0.0845 | -0.0327 | -0.0567 | -0.0753 |
| girl | -0.0099 | 0.0140 | -0.0212 | -0.0302 |
| sister | -0.3003 | 0.0001 | -0.3883 | -0.3445 |
| daughter | 0.0057 | 0.0082 | -0.0078 | -0.0216 |
| wife | -0.0775 | 0.0433 | -0.0880 | -0.0796 |
| girlfriend | -0.0203 | -0.0810 | -0.0316 | -0.0305 |
| mother | -0.5185 | 0.5365 | -0.5656 | -0.5568 |
| her | 0.0067 | -0.0267 | 0.0167 | 0.0159 |

Table 8: SHAP feature contributions for feminine terms

| Term | Anger | Joy | Fear | Sad |
|---|---|---|---|---|
| he | -0.0557 | 0.1033 | -0.0452 | -0.0583 |
| man | 0.0370 | 0.0950 | 0.0406 | 0.0206 |
| boy | -0.0731 | 0.0442 | -0.0787 | -0.0694 |
| brother | -0.1413 | -0.1026 | -0.1327 | -0.1445 |
| son | 0.2099 | -0.1571 | 0.2002 | 0.2108 |
| husband | -0.0048 | -0.0098 | -0.0367 | -0.0407 |
| boyfriend | 0.0590 | -0.0881 | 0.0316 | 0.0666 |
| father | 0.1688 | -0.2052 | 0.1078 | 0.1030 |
| him | 0.0747 | 0.009 | 0.0612 | 0.048 |

Table 9: SHAP feature contributions for masculine terms

explanations produced. For instance, terms associated with gender neutrality seem to consistently display a positive correlation with predictions associated with 'negative' emotions, i.e. anger, fear and sadness, whereas we observe that the predictions experience a negative correlation with predictions associated with joy. These results are far less pronounced for the associations between the masculine terms and the emotion sets, as well as that of the feminine terms. However, there does appear to be a slightly stronger positive association of the feminine terms with fear and sadness, and a somewhat negative association with feminine terms associated with joy. That said, due to the variance in LIME explanations, the results are not conclusive.

An additional factor to consider is the pre-trained model itself, which was trained on data from social media (Twitter), and thus might contain stronger unwanted associations due to the opinionated and polarizing nature of social media.

As SHAP has shown us, these inclinations may vary based on the features within a particular gender set as well (some may have a positive correlation with the emotion, whilst some may have a negative correlation). In addition to this, several of the terms that contain higher values of association

| Term | Anger | Joy | Fear | Sad |
|---|---|---|---|---|
| they | 0.0826 | -0.2282 | 0.0416 | 0.0596 |
| person | -0.2362 | 0.0609 | -0.2265 | -0.2188 |
| child | 0.0098 | -0.1030 | 0.0166 | 0.0071 |
| sibling | 0.0744 | -0.0061 | 0.0622 | 0.0502 |
| spouse | 0.0049 | -0.1113 | -0.0126 | -0.0141 |
| partner | -0.0346 | -0.0612 | -0.0675 | -0.0756 |
| parent | -0.0459 | -0.0838 | -0.0618 | -0.0712 |
| them | 0.0604 | 0.0409 | 0.0183 | 0.0505 |

Table 10: SHAP feature contributions for gender-neutral terms

| | Masculine | Feminine | Neutral |
|---|---|---|---|
| **Anger** | 88.89% | 44.44% | 44.44% |
| **Joy** | 55.56% | 66.67% | 66.67% |
| **Fear** | 66.67% | 44.44% | 44.44% |
| **Sad** | 77.78% | 44.44% | 66.67% |

Table 11: Percentage of Agreement between the LIME and the SHAP values

indicate a positive correlation with the negative emotions and vice versa. As mentioned before, the bias from the system does not appear to be generalized across genders, and thus these results warranted an in-depth term-by-term analysis to assess the extent to which they contribute to bias in these systems.

Table 11 illustrates whether or not there is an established correlation of the LIME and SHAP values of the gendered features with the ground truth label, for each emotion set. The table indicates the percentage of gendered features whose LIME and SHAP values are in agreement with each other. The positive agreement of these values indicates that there is a strong likelihood that that particular term may contribute either positively or negative toward the model's prediction.

In order to isolate the terms that are most likely to contain such bias, we attempt to gauge if the LIME and SHAP values of these terms across all emotion sets are in agreement with one another. Table 12[2] indicates the terms that are most likely to be biased, and the terms that show too much variance in their contributions to the predictions to be biased in a specific direction. We determine that a term is likely to be biased if the contributions of its LIME and SHAP values across multiple classes are

[2]'Almost all' indicates the term's values are only in agreement for 3 values, 'almost none' indicates that they are only in agreement for one

| All | Almost All | Almost None | None |
|---|---|---|---|
| man, boy, father, son<br>girl, wife, child<br>child | he, boyfriend<br>she, sister<br>they, partner | NA<br>woman, daughter, girlfriend, her<br>parent, them | him<br>mother<br>person |

Table 12: Agreement between LIME and SHAP values

consistently in alignment with one another, and we conclude that it is not biased in a specific direction if there is a great degree of variance in the mutual alignment of these values.

## 6 Ethics Discussion

In their work that critically surveys work done in the field of bias in language technology, Blodgett et al. (2020) recommends the inclusion of a statement to describe what we consider harmful system behaviours. In this section, we define the ways in which bias could exist in our system and how it could be detrimental.

### 6.1 What is bias?

We consider bias to be: the questionable association of gender signals (including pronouns) with a particular behaviour or emotion such that it influences the output of sentiment analysis models. We inspect how this association could contribute to the probability output of a sentiment analysis model. These biases in the output cause representational harms to the users by furthering stereotypes[3]. Considering the social context, this would disproportionately harm those who are already marginalized and underrepresented.

### 6.2 Beyond the Binary

In computational studies discussing gender, there is often not a clear, critically informed understanding of gender (Brooke, 2019). Additionally, studies on gender bias heavily rely on pronouns to measure 'gender' bias which as discussed in Ramesh et al. (2021) our contributions could be more inclusive of the definitions of self-identity of pronouns (Zimman, 2019). The mitigation of this bias is a critical task, and so is making sure that pronouns are not likely to influence the outputs of a system. However, the social impact of this work is limited if it is carried only for the 'binary gender'.

Our work shows a trend of bias for individuals whose identity falls outside of the gender binary who would be most likely to use gender-neutral terms and hence face these biases due to the system. It is important to note that we included *they* pronouns as a part of our gender-neutral sentence sets, and the work may be expanded to the inclusion of neopronouns [4] (example: xe/xem). In further work, this may help the NLP community to find a way to discern and mitigate the systemic biases that non-binary individuals might experience that the language models may be amplifying.

## 7 Conclusion

Through frameworks of explainable AI, we attempt to see if there is an established bias trend for masculine, feminine, and neutral gender signals in sentiment analysis and conclude that there is much variance. However, through LIME's correlations and SHAP's feature contributions, we observe that the model is drawing on the information from these gendered features. Identifying this bias for neutral gender signals opens scope for work for bias identification beyond the binary. Aside from this, in our expansion of this work, we aim to look at a broader range of models and tasks and see if the conclusions drawn by our methods used to form explanations agree with more traditional bias estimation metrics similar to those within Word Embedding Fairness Evaluation framework. We also look to explore more robust methods for developing these explanations and identify more features related to race, gender, etc that exhibit such societal bias.

## References

Kiana Alikhademi, Brianna Richardson, Emma Drobina, and Juan E. Gilbert. 2021. Can explainable ai explain unfairness? a framework for evaluating explainable ai.

Yassine Al Amrani, M. Lazaar, and Kamal Eddine El Kadiri. 2018. Random forest and support vector machine based hybrid approach to sentiment analysis. *Procedia Computer Science*, 127:511–520.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is

[3] We are omitting explicit examples of the stereotyping to avoid further stereotyping

[4] https://lgbta.wikia.org/wiki/Neopronouns

power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Francesco Bodria, A. Panisson, A. Perotti, and Simone Piaggesi. 2020. Explainability methods for natural language processing: Applications to sentiment analysis. In *SEBD*.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *CoRR*, abs/1607.06520.

Leo Breiman. 2001. Random forests. *Machine learning*, 45(1):5–32.

Sian Brooke. 2019. "condescending, rude, assholes": Framing gender and hostility on Stack Overflow. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 172–180, Florence, Italy. Association for Computational Linguistics.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Artem Domnich and Gholamreza Anbarjafari. 2021. Responsible ai: Gender bias assessment in emotion recognition.

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. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 609–614, Minneapolis, Minnesota. Association for Computational Linguistics.

Peter Hase and Mohit Bansal. 2020. Evaluating explainable ai: Which algorithmic explanations help users predict model behavior?

Mohammed Jabreel and Antonio Moreno. 2018. Eitaka at semeval-2018 task 1: An ensemble of n-channels convnet and xgboost regressors for emotion analysis of tweets.

Aditya Jain, Manish Ravula, and Joydeep Ghosh. 2020. Biased models have biased explanations.

Svetlana Kiritchenko and Saif M. Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems.

Stan Lipovetsky and Michael Conklin. 2001. Analysis of regression in game theory approach. *Applied Stochastic Models in Business and Industry*, 17(4):319–330.

B. Liu. 2012. Sentiment analysis and opinion mining. In *Synthesis Lectures on Human Language Technologies*.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2019. Gender bias in neural natural language processing.

Scott Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *CoRR*, abs/1705.07874.

Fang Luo, Cheng Li, and Zehui Cao. 2016. Affective-feature-based sentiment analysis using svm classifier. *2016 IEEE 20th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, pages 276–281.

Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.*, 267:1–38.

Krithika Ramesh, Gauri Gupta, and Sanjay Singh. 2021. Evaluating gender bias in Hindi-English machine translation. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pages 16–23, Online. Association for Computational Linguistics.

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. *CoRR*, abs/1602.04938.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of nlp models with checklist.

Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Gender bias in machine translation. *CoRR*, abs/2104.06001.

Shayak Sen and Yair Zick. 2016. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. pages 598–617.

Thiago Silveira, H. Uszkoreit, and Renlong Ai. 2019. Using aspect-based analysis for explainable sentiment predictions. In *NLPCC*.

Chaehan So. 2020. What emotions make one or five stars? understanding ratings of online product reviews by sentiment analysis and xai.

Ingo Steinwart and Andreas Christmann. 2008. *Support vector machines*. Springer Science & Business Media.

E. Ŝtrumbelj and I. Kononenko. 2013. Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, 41:647–665.

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.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods.

Lal Zimman. 2019. Trans self-identification and the language of neoliberal selfhood: Agency, power, and the limits of monologic discourse. *International Journal of the Sociology of Language*, 2019:147–175.

Chiara Zucco, Huizhi Liang, Giuseppe Di Fatta, and Mario Cannataro. 2018. Explainable sentiment analysis with applications in medicine. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1740–1747.

# A Appendix

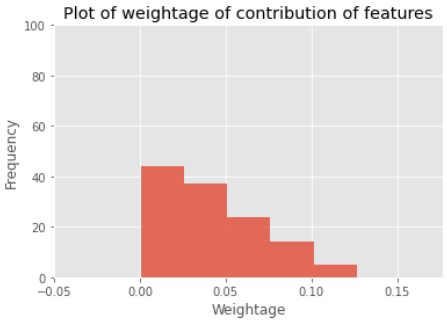


Figure 1: Distribution of the contribution of masculine features to the 'negative' class in LIME (Anger)

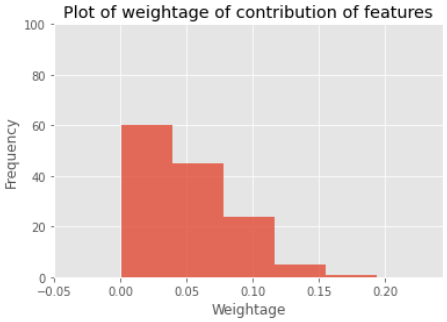


Figure 2: Distribution of the contribution of feminine features to the 'negative' class in LIME (Anger)

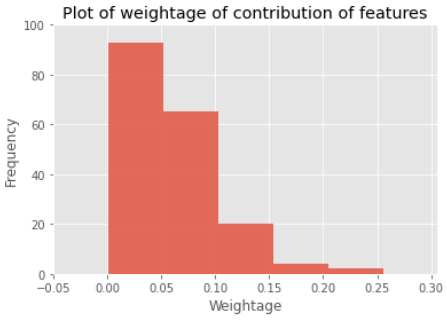


Figure 3: Distribution of the contribution of gender neutral features to the 'negative' class in LIME (Anger)

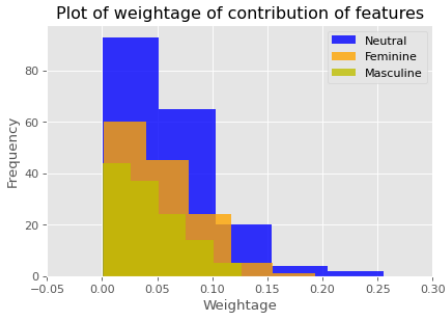


Figure 4: Comparative graph of all the distributions of features to the 'negative' class in LIME (Anger)

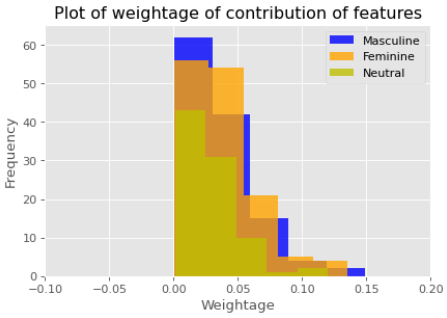


Figure 8: Comparative graph of all the distributions of features to the 'positive' class in LIME (Joy)

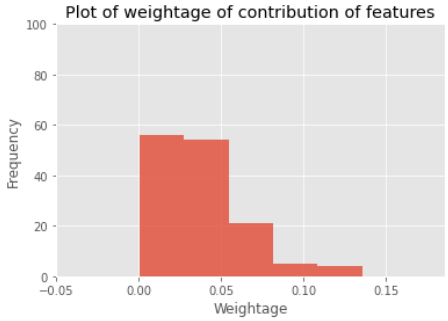


Figure 5: Distribution of the contribution of masculine features to the 'positive' class in LIME (Joy)

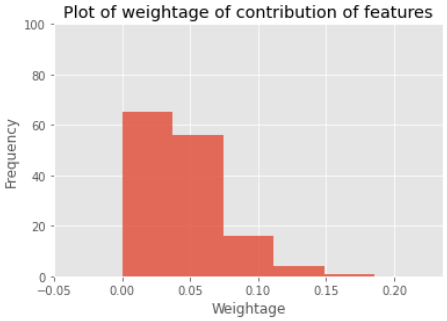


Figure 9: Distribution of the contribution of masculine features to the 'negative' class in LIME (Fear)

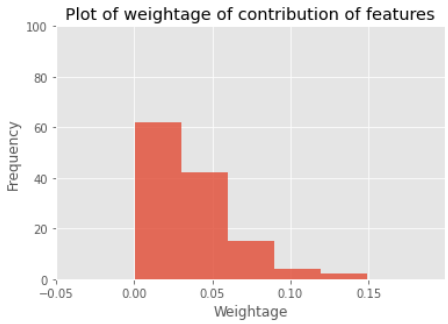


Figure 6: Distribution of the contribution of feminine features to the 'positive' class in LIME (Joy)

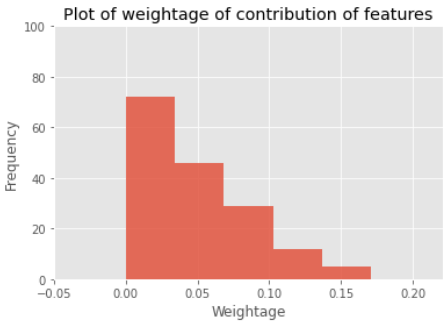


Figure 10: Distribution of the contribution of feminine features to the 'negative' class in LIME (Fear)

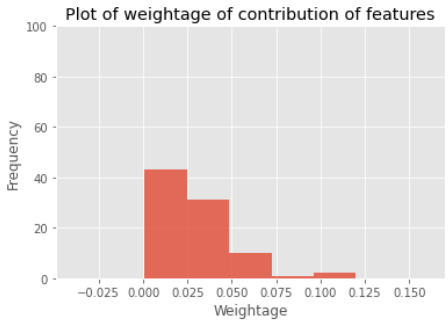


Figure 7: Distribution of the contribution of gender neutral features to the 'positive' class in LIME (Joy)

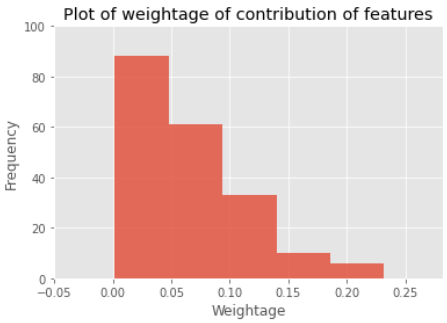


Figure 11: Distribution of the contribution of gender neutral features to the 'negative' class in LIME (Fear)

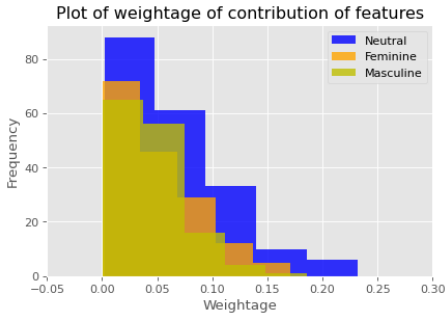


Figure 12: Comparative graph of all the distributions of features to the 'negative' class in LIME (Fear)

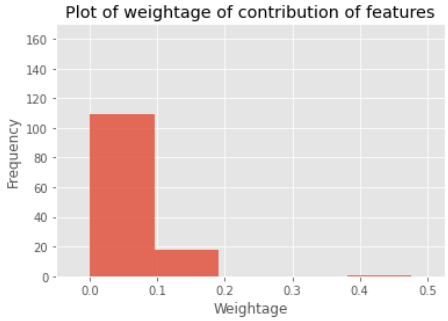


Figure 13: Distribution of the contribution of masculine features to the 'negative' class in LIME (Sad)

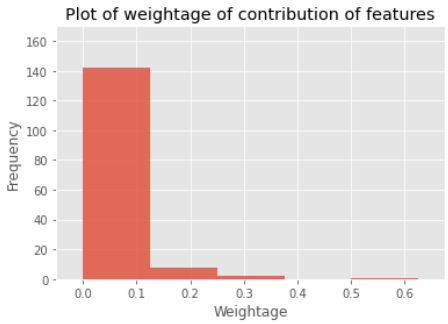


Figure 14: Distribution of the contribution of feminine features to the 'negative' class in LIME (Sad)

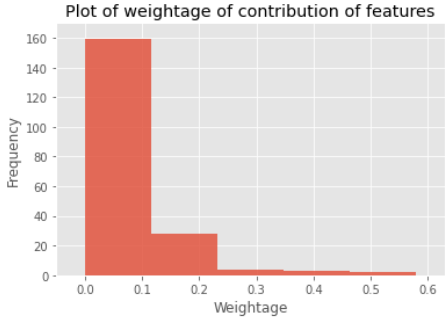


Figure 15: Distribution of the contribution of gender neutral features to the 'negative' class in LIME (Sad)

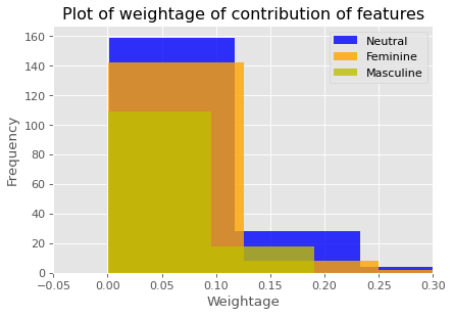


Figure 16: Comparative graph of all the distributions of features to the 'negative' class in LIME (Sad)