An Explainable Approach to Understanding Gender Stereotype Text

Manuela Nayantara Jeyaraj *
Technological University Dublin
Ireland
manuela.n.jeyaraj@mytudublin.ie

Sarah Jane Delany
Technological University Dublin
Ireland
sarahjane.delany@tudublin.ie

Abstract

Gender Stereotypes refer to the widely held beliefs and assumptions about the typical traits, behaviours, and roles associated with a collective group of individuals of a particular gender in society. These typical beliefs about how people of a particular gender are described in text can cause harmful effects to individuals leading to unfair treatment. In this research, the aim is to identify the words and language constructs that can influence a text to be considered a gender stereotype. To do so, a transformer model with attention is fine-tuned for gender stereotype detection. Thereafter, words/language constructs used for the model's decision are identified using a combined use of attention- and SHAP (SHapley Additive exPlanations)-based explainable approaches. Results show that adjectives and verbs were highly influential in predicting gender stereotypes. Furthermore, applying sentiment analysis showed that words describing male gender stereotypes were more positive than those used for female gender stereotypes.

1 Introduction

Gender stereotypes (GS) are the perceptions about the typical physical, emotional, and social characteristics displayed by men and women (Wiegand et al., 2021; Blumer et al., 2013; Ellemers, 2018; Morgan and Davis-Delano, 2016). Thus, gender stereotypes function as text that can be used to directly or indirectly infer that individual's gender. These perceptions/beliefs assumed by society about an individual based on their gender can lead to gender bias negatively impacting that individual's life.

For example, Andrich and Domahidi (2022) studied descriptions about U.S. Political candidates. Their study showed that Facebook comments posted by users were gender stereotypical in the way that the male candidates were described with stronger masculine traits associated to a political career than the female candidates. This discrepancy and power inequality in traditionally assumed feminine/masculine gender stereotypes has the potential to negatively influence the voters' decisions thus penalizing the candidates based on their gender (Eagly, 2013). Another similar instance occurred during the 2017 Labor leadership election in Britain. An analysis of the language used in news articles about the candidates showed discrepancies in how they were described that were related to their gender[1]. These examples illustrate how language used to describe the subject based on their gender may perpetuate gender stereotypes and lead to gender bias and/or unfair treatment of individuals based on their gender. Hence, it is important to understand gender stereotypes that could potentially lead to gender bias and discrimination against individuals based on their gender.

The aim of this paper is to use explainable AI (XAI) approaches when predicting gender stereotypes to understand the words or language that suggest a gender stereotype. A challenge with using AI prediction models is that they are black-boxes. It makes it hard for humans to understand why models arrived at the particular decisions that they predicted (Xu et al., 2019). Therefore, XAI approaches aim to improve the transparency and interpretability of AI models by offering explanations as to how or why the predicted result was inferred.

XAI approaches are generally categorized as transparency design explanations and post-hoc explanations (Lipton, 2018). Transparency design approaches explain how the model functions in the view of the developer such as the model's structure, understanding the individual components of

[1]Gender bias in Political description of candidates: https://www.theguardian.com/technology/2017/ apr/13/ai-programs-exhibit-racist-and-sexist-biases-research-reveals

Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP), pages 45–59

the model, its underlying training algorithm, etc. Post-hoc explanations provide an understanding of why a prediction was inferred; the components of the input that influenced the output (Xu et al., 2019). In this work, we use post-hoc explanation approaches such as attention and SHAP to identify words that influenced the model's prediction of a gender stereotype and anti-gender stereotype text.

Since the idea of attention was introduced in Vaswani et al. (2017), it has been used in understanding text for various NLP tasks as the attention mechanism helps a model to capture the context of words and to focus on the relevant parts of a text when making decisions about the prediction (Chen et al., 2019; Bai, 2018; Liu et al., 2020). Attention captures the importance of the word to the model's prediction corresponding to that particular input text. Therefore, it has been considered to be a local-level of explanation surrounding that particular input instance (Danilevsky et al., 2020).

On the other hand, XAI explanations like SHAP enable a more sophisticated understanding of how the words are important on a global-level to the whole model. Therefore, SHAP is said to generate global explanations of a model's prediction offering a global understanding of which words are important.

Our approach is to fine-tune a transformer model with attention to classify textual input as a gender stereotype or anti-gender stereotype. Thereafter, using the attention and SHAP-based explanations, we identify the words that influence the model's decision to categorise the input text as a gender stereotype. In addition, we perform a sentiment analysis on the identified top-influential words to study the emotion associated with the choice of words used for gender stereotypes about men and women.

Our analysis of top-influential words and language constructs show that adjectives and verbs highly impact gender stereotype predictions. In addition, sentiment analysis shows that gender stereotypes associated with the male gender are more positive than those associated with the female gender.

The rest of this paper is structured as follows. Section 2 presents the related works on gender stereotypes and gender stereotype detection. Section 3 outlines the datasets and model architecture implemented, the explainable approaches used and how we obtain the top-influential words that suggest gender stereotypes. We present the results of our evaluation in section 4 and discuss the observations. We conclude by presenting our key findings and some limitations in our current work.

2 Related Work

Often gender stereotype and gender bias are considered synonymous though their focus and scope differ (Blodgett et al., 2020). Gender bias is a more specific and technical term that refers to the intentional or unintentional discrimination against individuals based on their gender (Costa-jussà, 2019). More generally, gender stereotypes refer to the widely held beliefs and assumptions about the typical traits, behaviors, and roles that are associated with men and women in society (Wiegand et al., 2021; Ellemers, 2018; Morgan and Davis-Delano, 2016; Blumer et al., 2013).

Although the definition of gender stereotypes roots from the attribution of characteristics or traits to the group, the bias itself rises from the discrimination an individual faces by being assumed and assigned the same characteristics or traits of the group. Hence, this paper discusses stereotyping from the perspective of an individual as driven by the motivating examples in the introduction.

Most of the work in existing literature focuses on identifying and understanding gender bias using ML rather than on gender stereotypes (Hoyle et al., 2019). For example, researchers investigated the existence and/or the mitigation of gender bias in word embeddings (Bolukbasi et al., 2016; Zhao et al., 2019; Caliskan et al., 2022), Language models (Bordia and Bowman, 2019; Kurita et al., 2019; Vig et al., 2020; Nadeem et al., 2021), coreference resolution (Rudinger et al., 2018; Zhao et al., 2018; Cao and Daumé III, 2019), machine translation (Stanovsky et al., 2019; Prates et al., 2020; Savoldi et al., 2021), Parts-of-Speech (POS) tagging (Garimella et al., 2019), natural language generation (Sheng et al., 2020), etc.

Existing work on analysing gender stereotypes is mainly focused on the use of pre-defined lexicons of gender-specific words and actions curated through manual and psychological studies (Bem, 1974; Rosenkrantz et al., 1968; Spence Janet and Joy, 1974). Herdağdelen and Baroni studied the association between gender and actions related to gender stereotypes. They extracted verb-noun pairs from the OMCS Common sense database and analyzed the occurrence of the verb-noun pairs in the tweets. Their results showed that there

are clear gender associations with certain actions, such as women being more associated with cooking and cleaning, while men were more associated with driving and building.

Rubegni et al. explored how children perceive gender stereotypes by analyzing the characters in text written by children in the form of storytelling. They found that male antagonists were described using a limited set of negative adjectives which are demeaning descriptors, while female antagonists were defined using a richer and more varied set of negative qualities.

A more recent study by Cryan et al. used a self-compiled dataset of web-posts and news articles which were annotated through crowd-sourcing to identify instances of gender stereotypes. This supervised learning-based method involved training a machine learning model on a set of annotated data to classify texts as to whether the description of an individual in text conformed or contradicted to the intended gender of the subject. The most frequently used words which were used for the gender-conforming and gender-non-conforming predictions were presented in their work.

In the past, while machine learning models remained black boxes, using the attention mechanism was a popular approach to understand the predictions of models by looking at the parts of text that were highly attended to as the model was making its decision (Xu et al., 2015; Bahdanau et al., 2014). When a transformer model processes each word in the input text, it calculates attention scores for each word. This attention score indicates how much weight or attention the model should give to that word when it decides the predicted class. Various studies have found attention to be unreliable explanations (Abnar and Zuidema, 2020). Although the attention score captures the absolute importance of the token, researchers have contradicted the idea of how this instance-level understanding can be approximated to get a global understanding of the feature's importance to the whole model's prediction understanding (Sun and Lu, 2020). And, the scaling factor used to calculate the attention score can affect the interpretability of the feature's importance in terms of the attention.

According to Jain and Wallace (2019), attention is not a robust indicator. Attention was found to loudly predict the overall relevance of the input components (the words) to a model (Serrano and Smith, 2019). Moreover, Danilevsky et al. (2020) question the extent to which attention can provide explainability of feature importance. Attention weight measures the relative importance of the token within a specific input sequence. So though the attention score captures the absolute importance of the token, researchers have contradicted the idea of how this instance-level understanding can be approximated to get a global understanding of the feature's importance to the whole model's prediction understanding (Sun and Lu, 2020). Nevertheless, there are works that strongly challenge this claim of the attention not being an explanation of feature importance (Wiegreffe and Pinter, 2019). And, researchers have been using the attention score to understand and interpret top words influencing the predictions of machine learning models (Vashishth et al., 2019; Tal et al., 2019).

Recently, the concept of XAI has paved way for these black-box ML model predictions to be interpreted as glass-box explanations (Holzinger, 2018; Rudin and Radin, 2019). There are a wide variety of approaches through which these explanations can be derived (Mathews, 2019; Gunning et al., 2019; Vilone and Longo, 2020). But most of these are based on post-hoc explanations of a surrogate model that render model-agnostic explanations. Some such XAI approaches are SHAP and LIME (Local Interpretable Model-Agnostic Explanations).

SHAP provides a global explanation of the output of any ML model by assigning each feature an importance value (SHAP value) in the prediction process (Lundberg and Lee, 2017). SHAP values take into account the token interactions based on whether a word is present or absent across the predicted instances and builds a model based on these changes to explain the predictions in the context of other words. Work done by Bosco et al. (2023) used SHAP values to study explanation of racial stereotypes. This study identified the words that were most influential in categorizing text into different categories of hate speech based on their SHAP values.

3 Methodology

This section outlines the datasets, the model architecture, and the approach used to identify the most influential words for a prediction.

In this research, rather than looking at

male/female as a biological sex assigned at birth, we consider male/female as a gender. As defined in (Albert and Delano, 2022), "Gender refers to a person's gender identity (how they see themselves or experience their own gender) but also involves other factors such as how a person is perceived by others or experiences differential treatment related to their perceived gender".

Three gender stereotype datasets, see Table 1, were used.

Dataset	#Samples	Min chars	Max chars	Distribution of samples as a % of the whole dataset			
				GS		Anti-GS	
				Male	Female	Male	Female
SSet	1,986	14	165	24	22	30	24
CC	4,550	14	45,242	25	25	25	25
CR	3,221	7	889	34	30	16	20

Table 1: Dataset description and statistics where GS means gender stereotype.

The **StereoSet (SSet)** dataset (Nadeem et al., 2021) contains 4 stereotypical categories (gender, race, religion, occupation) of which we use the gender category instances for our research. To create this dataset the authors compiled target terms that represented the different target categories (e.g., for gender "woman", for race "Asian", etc.) based on Wikidata associations found in triples related to the above categories. Then, crowd-workers were asked to write two sentences describing people using these target terms where one sentence suggests a gender stereotype while the other does not. We require the gender of the subject discussed in the text but gender is not explicitly identified in this dataset. We manually labelled the gender identity of the subject as describing a male or a female person. There were 55 instances where the gender of the subject described in the text was not identifiable, these instances were excluded from our analysis.

Cryan's content (CC) dataset was specifically compiled to study gender stereotyping (Cryan et al., 2020). Using crowd-sourcing crowd workers were asked to find articles that describe a person (male/female) and label them as whether the description is consistent or contradictory to common gender stereotypes as perceived by that crowd-worker. This dataset has 4 labels, consistent with or contradictory to male/female. Translating these labels to a binary classification for our experiments, the male/female consistent labels become gender stereotypes (GS) and the contradictory ones, anti-gender stereotypes (anti-GS).

The crowd-workers who were compiling and labelling articles for Cryan et al.'s research were also requested to provide their reason for labelling an article as consistent with a gender stereotype or contradictory to a gender stereotype which was not used in their study. Reviewing these texts provided as reasons by the annotators, we found them to be valid and direct perceptions of why a person (crowd-worker) would consider a certain text as a GS or an anti-GS. We used these reason texts to generate a dataset which we called **Cryan's Reasons (CR)** and labelled it manually as a GS or anti-GS text. To label the data, it was divided into 4 subsets of approximately 1000 text samples each, and 3 annotators labelled each subset of text samples. Annotators were asked to label if they considered the text was a gender stereotype or not. They were also asked to select if they thought the text described a "male", "female", "non-binary" gendered person or was "not related to a person".

The inter-annotator agreement (IAA) for the GS/anti-GS label for each subset was calculated using the Fleiss kappa (Fleiss et al., 1981). One subset of labelled text samples with an IAA less than 0.8 was dropped and the other 3 subsets with IAAs of 0.89, 0.89 and 0.9 were retained giving an average IAA across all retained labelled samples of 0.89.

To arrive at a consensus label for the gender and gender stereotype/anti-GS labels, the label assigned to each instance was based on a majority vote, i.e. the value chosen by 2 out of 3 annotators. Instances where the 3 raters' gender labels were all different were dropped. Then, we removed the instances where the consensus gender label was "not related to a person". Only 37 samples were about non-binary people (11 GS and 26 anti-GS). This was not sufficient to train and test a classifier model for our study. Therefore, we retained the male and female samples, a total of 3221 samples: 1081 male GS, 958 female GS, 528 male anti-GS and 654 female anti-GS samples.

Following a similar approach to Cryan et al. (2020), we use a transformer model based on the BERT architecture, which is a pre-trained deep neural network architecture used to process sequential input data, such as text. We chose BERT due to its bidirectional nature. In addition, its context aware embeddings capture relationships between words. And researchers have been successfully fine-tuning BERT for downstream tasks in the past within the domain (Huo and Iwaihara,

2020; Mohammadi and Chapon, 2020; Xinxi, 2021; Qasim et al., 2022).

We fine-tuned BERT for the gender stereotype detection task on each dataset and added a classification head to predict if a new unseen text was a GS or anti-GS. The pre-trained BERT model is fine-tuned on the labeled training datasets and optimized for the best hyper-parameters using Optuna (Akiba et al., 2019) which is an open-source hyper-parameter optimization framework based on Bayesian optimization. Performance is measured as the average class recall (due to imbalance in the class distribution of the data) over three iterations of 5-fold cross validation on each dataset.

The sole use of one XAI approach is not a reliable measure of the influential words contributing to the prediction (Fryer et al., 2021). Attention scores can sometimes be sensitive to noise or outliers in the data leading to misleading interpretations (Serrano and Smith, 2019). And although the fundamental workings of ML models remain unclear, XAI methods approximate the model's behaviour based on the predictions. Therefore, the post-hoc explanations produced by XAI methods like SHAP alone may not be as fully accurate at capturing how the ML model arrived at a decision (Zhong and Negre, 2022) either. Hence, we looked into capturing the words' importance in making a prediction using more than one approach.

Abnar and Zuidema (2020) state that though SHAP values are not attention scores, the attention flows which are an extension of attention weights obtained after post-processing align with SHAP values. So, we use the attention score along with the SHAP value to identify the words that influence the model's prediction. We combine the attention score and SHAP values to get an influence score $IScore(w_i)$ for the word w_i as shown in Equation 1.

$$IScore(w_i) = \frac{AS(w_i)}{SV(w_i)} \quad (1)$$

where $AS(w_i)$ is the attention score and $SV(w_i)$ is the SHAP value of the corresponding word.

We ranked the words in each instances by their influence scores. We selected the top three words with the highest word influence score for analysis. The words with word influence scores lower than these top three were typically article words (a, an, the), prepositions (in, under), conjunctions (and, but) and determiners (some, many).

4 Results and Discussion

Figure 1 reports the mean and std.deviation of the average class recall on the three datasets across three iterations of 5-fold cross validation for the gender stereotype detection task.

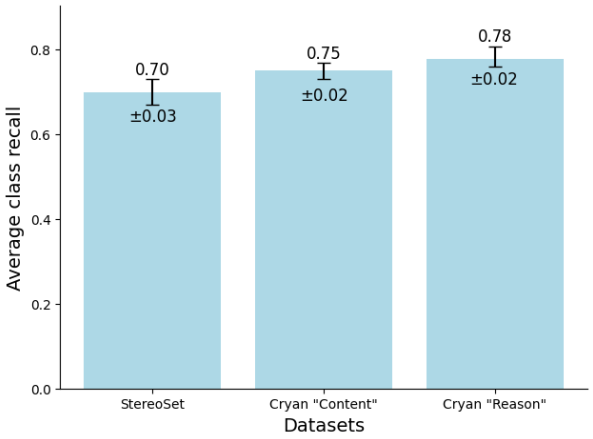


Figure 1: Average class recall across the three datasets.

SS, CC and CR datasets obtained average class recalls of 0.7, 0.75 and 0.78, respectively with the CR dataset achieving the best performance. Further analysis was carried out on the words and type of language constructs that influenced the predictions.

4.1 Influence of gendered and non-gendered words

First, we analysed the influence of gendered and non-gendered words on the predictions by identifying the proportion of gendered words from the top three words considered as the most influential words by the model based on our influence score. The gendered words were manually identified as a list of words consisting of gendered pronouns ("he/she", "him/her", etc.), words explicitly ending on '-man/men', '-woman/-women' ("policeman", "businesswoman", etc.), and gendered terms ("mother", "sister", "actress", etc.) compiled from the ESCWA Gender-Sensitive Language Guidelines released by the United Nations [2].

Figure 2 illustrates the percentage of gendered words found in the words that most influenced the

[2]ESCWA Gender-Sensitive Language Guidelines: https://archive.unescwa.org/sites/www.unescwa.org/files/page_attachments/1400199_0.pdf

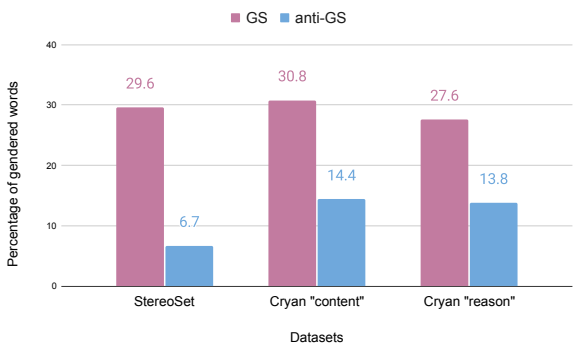


Figure 2: Percentage of gendered words with a high word importance score associated with the prediction of each class across all datasets.

prediction.

This shows that across all datasets the model uses a higher proportion of gendered words to predict GS than it does to predict anti-GS. This can be attributed to the presence of gendered pronouns or words with lexical gender from which the gender can be directly inferred. For example, the text "She liked to bake cookies and pies all day" was correctly predicted as a GS by focusing on the gendered word "she" along with the other two top words "liked" and "bake" in that text. And the word "bake" being associated with a female-gendered word "she" shows how women are associated with typically feminine, gender-stereotypical gender roles. However, the text "She is outside doing yard work" was incorrectly predicted as a stereotype as the perception of a gender stereotype is tied to the gender performing the task mentioned in the text which was not clearly captured for the above sample prediction.

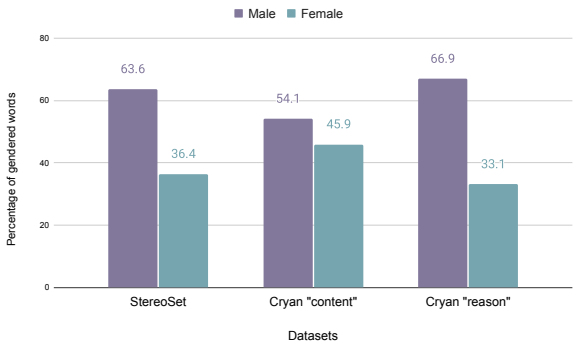
We evaluated if gendered words are more prominently associated with one gender over the other when it comes to predicting gender stereotypes. Figures 3a and 3b visualize the percentage of gendered words associated with male/female instances for the GS and anti-GS predictions respectively.

Figure 3a shows that more of the gendered words for GS predictions are associated with a male instance than a female instance. This pattern can be tied to tradition where gender stereotypes have depicted men as powerful, authoritative, and capable, whereas women are frequently represented in caring or submissive positions. Because preconceptions about men are more often represented in a manner that is considered neither harmful or derogatory to the male gender, those gender stereotypes continue to be used in society. Hence, this bias may result in a stronger connection of gendered phrases with male gender stereotype examples.

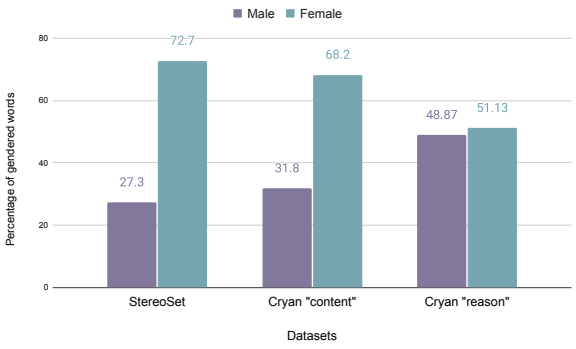
However, figure 3b shows a significantly higher percentage of gendered words used for anti-GS are associated with females than males. The growing awareness around gender-inclusivity and bias against women may have caused a larger inclination for people to use gendered terms with female examples in anti-GS situations. This may also indicate a deliberate effort to fight and confront preconceptions that paint women in a gender-stereotypical manner.

4.2 Influence of Parts of Speech

Contrary to *lexical gender*, which refers to the inherent gender classification of a word based on its meaning (e.g. businessman, actress, etc.) (Siemund and Dolberg, 2011), *social gender*



(a) for GS predictions.



(b) for anti-GS predictions.

Figure 3: Percentage of gendered words associated with predictions of both GS and anti-GS.

refers to the implicit inference of an individual's gender from words (such as adjectives, verbs, etc.) where the gender is not obvious (McDowell, 2015). This inference roots from cultural and social roles, behaviors, and expectations associated with masculinity and femininity in a society or community (Fausto-Sterling, 2019). A definition in (Ackerman, 2019) terms the social gender as Biosocial gender which is "the gender of a person based on phenotype, socialisation, cultural norms, gender expression, and gender identity". Out of these, in this research the concepts of *gender expression* and *gender roles* (Benwell, 2006; Soundararajan et al., 2023) in gender stereotypes are studied further.

Gender expression refers to the way an individual presents their gender to the world through their appearance and characteristic traits (Rubin and Greene, 1991). In terms of language and parts-of-speech (POS) in text, an individual's appearance, i.e., gender expression, is typically described using adjectives (Hamon, 2004; Hattori et al., 2007; Otterbacher, 2015; Ismayanti and Kholiq, 2020).

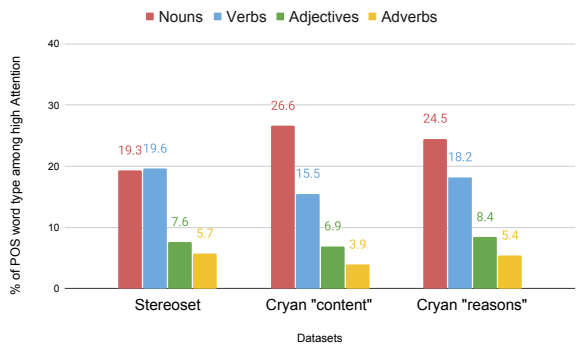
Gender roles are societal expectations or norms associated with gender, including behaviors, actions, and activities that are considered appropriate for men and women (Gabriel et al., 2008). Language-wise, the actions/roles one performs are typically described using verbs (Semin and Fiedler, 1988; Bower et al., 1979; Sanford and Garrod, 1998; Van Atteveldt et al., 2017; Clark et al., 2018).

In order to build a generic view of what type of language constructs, including these implicit gendered words, suggest a text to be as a gender stereotype, we analysed the influence of different POS on predictions. Figure 4a shows the distribution of different parts of speech across all instances in the three datasets. This is compared to Figure 4b which shows the distribution of different POS-tagged adjectives (gender expression descriptors), verbs (gender role descriptors), adverbs (action/gender role modifiers) and nouns that influenced the predictions.

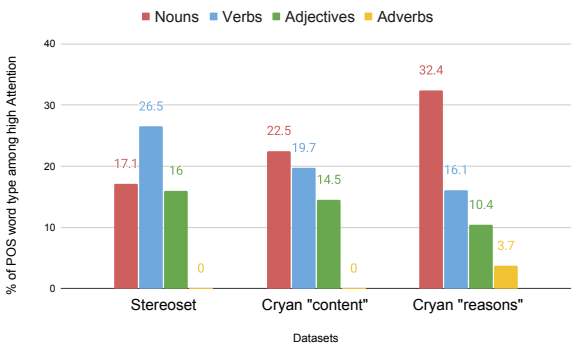
Although there are comparatively fewer adjectives across all the instances in the datasets, the model has focused mostly on adjectives and verbs to make predictions. Also, though there are more nouns across all three datasets, they are significantly lower in proportion among the most influential words in the SSet and CC datasets with a slight exception in the CR dataset. This shows that nouns are not as influential as adjectives or verbs in detecting gender stereotypes. This aligns with the social gender concepts of gender expression, captured by adjectives, and gender roles including behaviour and actions, captured by verbs, showing that both adjectives and verbs are significant indicators in identifying gender stereotypes.

Research by Ye et al. revealed that the overall usage frequencies of personality adjectives used to describe men and women across two centuries were higher for men than women. Hence, we further analysed the different POS among the most influential words based on the gender that they were associated with. Figure 5a confirms that there is a higher percentage of adjectives associated with males than females across all datasets.

Figure 5b shows that slightly more top nouns were associated with males than females. This pattern agrees with the existing social bias where the world is used to viewing generic experiences

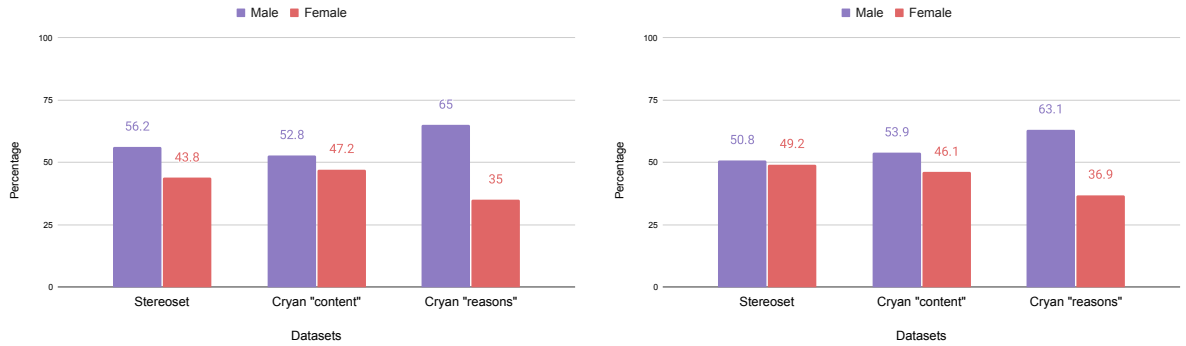


(a) across the entire dataset.



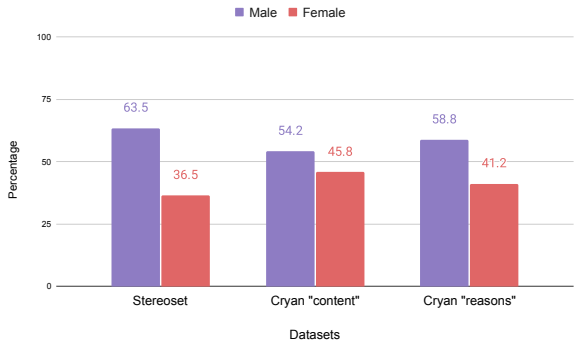
(b) across most influential words used for the model's prediction.

Figure 4: Distribution of different POS types across the datasets and predictions.



(a) Adjectives

(b) Nouns



(c) Verbs

Figure 5: Distribution of different POS types across the most influential words used for predictions, associated with gender.

and descriptions as mostly relevant to men [3]. Models trained on datasets inadvertently learn and capture biases present in the training data. Since our analysis found that there was a higher likelihood of top nouns appearing in sentences that were labeled by human annotators as text suggesting a male-GS, it shows that our model has merely learned to reflect this behaviour and is assigning more importance to certain nouns when the context is associated with a male stereotype text i.e., the discussion or description of males. This may reflect human perception by capturing the biases on how people have been traditionally described in terms of their personality traits.

Figure 5c reflects the distribution of most influential verbs across genders in the prediction of stereotypes. Once again, there are slightly more verbs associated with males than with female instances. In the statistical analysis done in the study conducted in (Haines et al., 2016) regarding the perceptions of gender stereotypes for the past 3 decades from 1983-2014, there were fewer women participating in actions related to politics, sports, etc. And the stereotypical beliefs associated with women were either more tied to characteristic traits or traditional gender roles assumed to be feminine (e.g., caring for family). This observation regarding verbs (gender role descriptors), is also supported by our motivating example about the 2017 British Labor leadership Elections where the 2 female elections candidates were discussed more in terms of their fathers and their family where the actual modern shift in gender roles in the present-society is not being reflected. Women have begun taking up new gender roles in fields such as politics or sports which were not traditionally considered to be feminine. Thus, in reality, the gap between the gender roles taken up by men and women is being bridged. However, this shift in equivalence of gender roles taken up by men and women is not reflected by traditional gender

[3] Article on Gender Sensitive Communication by European Institute of Gender Equality: https://eige.europa.eu/publications-resources/toolkits-guides/gender-sensitive-communication/challenges/invisibility-and-omission/do-not-use-gender-biased-nouns-refer-groups-people

stereotypes which are more associated with men as seen in our data. This possibly implies how traditional gender stereotypes perceived by society (as captured in the datasets) do not reflect the reality of modern gender roles (described using verbs) being equally taken up by both genders.

There were no adverbs among the influential words for the SSet and CC datasets. Only the CR had more male-associated adverbs than female-associated adverbs in predicting GS.

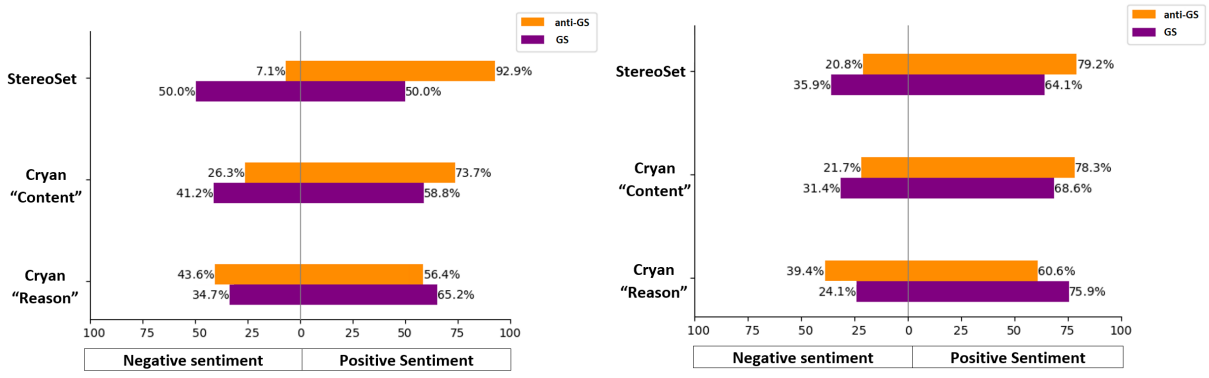
4.3 Sentiment Analysis of predictive words

In order to examine whether the emotions associated with the most influential words were related to specific genders, we analysed the sentiment of the most influential adjectives and verbs used in predictions. We used SentiWordNet 3.0 (Baccianella et al., 2010) to get the sentiment associated with a word. Figure 6 shows the percentage of most influential adjectives and verbs associated with a positive/negative sentiment for predictions across the three datasets. The orange bar represents the most influential adjectives (see figure 6a)/verbs (see figure 6b) used to predict anti-GS text samples while the purple bar represents the adjectives/verbs used to predict GS text. The portion of the bar lying on the right side of the origin along the x-axis represents the proportion of those adjectives/verbs associated with a positive sentiment. And the portion of the bar lying on the left side of the origin along the x-axis represents the proportion of those adjectives/verbs associated with a negative sentiment.

For the three datasets, the adjectives used in the prediction of anti-GS text (see figure 6a) convey a more positive sentiment. Though the adjectives used to predict GS text have a slightly more positive sentiment as observed in the CC and CR datasets, this difference is not significant. Hence, this suggests that anti-GS text tends to bear a slightly more positive social perspective of characteristic traits pertaining to the genders. The same evaluation was carried out for verbs in figure 6b which shows that verbs associated with a more positive sentiment prompt anti-GS predictions in general. This is similar to the pattern displayed by the sentiment associated with top adjectives (Figure 6a).

We also examined whether the sentiment associated with the adjectives/verbs were tied to a specific gender. In the following graphs, the green bar represents the most influential adjectives/verbs used to predict GS/anti-GS text about a female and the blue bar, a male. The portion of the bar lying on the right side of the origin along the x-axis represents the proportion of those adjectives/verbs associated with a positive sentiment. And the portion of the bar lying on the left side of the origin along the x-axis represents the proportion of those adjectives/verbs associated with a negative sentiment.

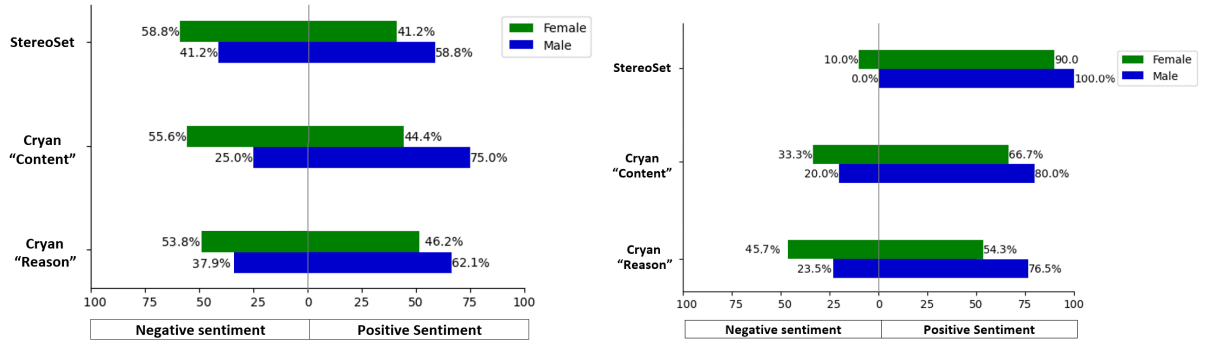
Figure 7a shows that GS characteristic traits of females described using adjectives (i.e., gender expressions) are associated with a slightly more negative sentiment whereas adjectives used to de-



(a) Adjectives

(b) Verbs

Figure 6: Sentiment associated with different influential words corresponding to the parts of speech.
(Orange bar: proportion of most influential adjectives (6a) / verbs (6b) used to predict anti-GS text samples. Purple bar: proportion of most influential adjectives/verbs used to predict GS text. Portion of the bar lying on the right side of the origin along the x-axis represents the proportion of those adjectives/verbs associated with a positive sentiment. Portion of the bar lying on the left side of the origin along the x-axis represents the proportion of those adjectives/verbs associated with a negative sentiment.)



(a) for GS predictions

(b) for anti-GS predictions

Figure 7: Sentiment associated with most influential adjectives.
(Green bar represents the proportion of the most influential adjectives used to predict GS (7a) / anti-GS (7b) text about a female and the blue bar, a male. The portion of the bar lying on the right side of the origin along the x-axis represents the proportion of those adjectives associated with a positive sentiment. And the portion of the bar lying on the left side of the origin along the x-axis represents the proportion of those adjectives associated with a negative sentiment.)

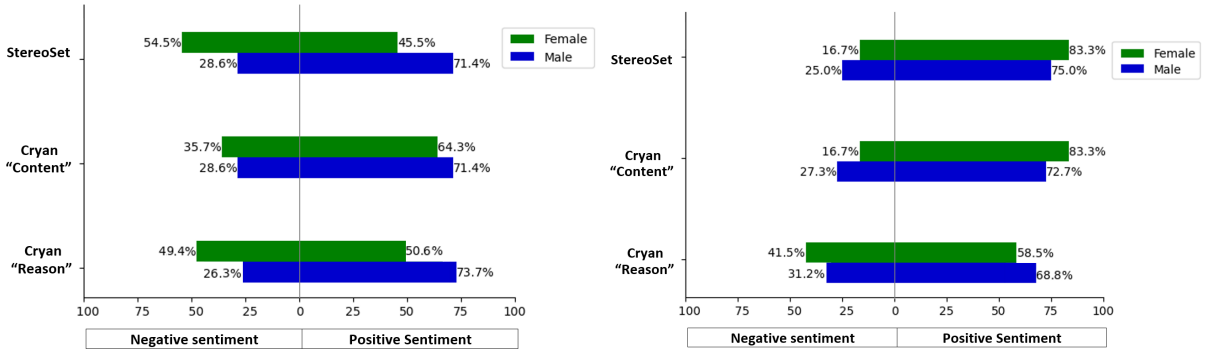
scribe males are significantly more positive. This can suggest the existing gender bias in society where gender expression or characteristic traits expected of women are associated with traditional standards of beauty and appearance (Cash and Brown, 1989; Lavin and Cash, 2001; Heflick et al., 2011). When a modern female deviates from these established norms, it can be negatively perceived by society (Biefeld et al., 2021; Plaza-del Arco et al., 2024). However, the same shift in gender expressions and characteristic traits illustrated by men are not accentuated perceived in a similar negative sense (Shyian et al., 2021).

Figures 7b shows that adjectives used to predict anti-GS are associated with a more positive sentiment for both genders than they are with predicting GS across all datasets.

The same evaluations were performed on verbs and are shown in figure 8a and 8b for GS and anti-GS respectively.

Figure 8a shows that verbs used to predict GS were significantly more positive for males than females. However, words used to predict anti-GS were associated with a positive sentiment for both genders (see Figure 8b), which is consistent with the pattern displayed by adjectives used to describe males/females.

This behaviour of describing males and females using gender expression and gender role descriptors that are associated with different sentiments shows that the model has learned some biases from the training data which may reflect the societal gender biases against males and females. The words (adjectives, verbs) that are more influential



(a) for GS predictions

(b) for anti-GS predictions

Figure 8: Sentiment associated with most influential verbs.
(Green bar represents the proportion of the most influential verbs used to predict GS (8a) / anti-GS (8b) text about a female and the blue bar, a male. The portion of the bar lying on the right side of the origin along the x-axis represents the proportion of those verbs associated with a positive sentiment. And the portion of the bar lying on the left side of the origin along the x-axis represents the proportion of those verbs associated with a negative sentiment.)

are mirroring society's negative perception when it comes to describing the characteristic traits and expected gender roles of women. However, society has been accustomed to describing men in a more positive manner, be it their characteristic traits or expected gender roles (Fast et al., 2016). The presence of this biased societal perception is supported by our experiments and results.

As such, we found that adjectives that are gender expression/characteristic trait descriptors and verbs that are gender role/action descriptors are highly influential in prompting gender stereotypes. Moreover, we found that words describing a male gender stereotype are more positive than those used to describe the female gender stereotype.

5 Conclusion

Gender stereotypes manifest in the way people express themselves through gender expression/characteristic traits described using adjectives or their gender roles/actions described using verbs. These gender stereotypes can prompt harmful effects leading to gender bias if not captured. In this research, we fine-tune a transformer model with attention to classify gender stereotypes. A proposed combination of attention and SHAP explainable approach is used to identify the words/language constructs that influence a text to be considered as a gender stereotype or not. Our findings showed that adjectives (gender expression descriptors) and verbs (gender role descriptors) highly impact a text to suggest a gender stereotype. Furthermore, a sentiment analysis of identified top-influential words also revealed that top-influential words used to describe males were more positive than those chosen to describe females. This partiality towards the way in which genders are described represents gender bias where humans evaluate expressions related to men more positively than those related to women.

Limitations and Future work

In this work, we have only used attention and SHAP to identify the words and thereby, the language that influences gender stereotypes. In our ongoing extension of this research, we will explore the use of other post-hoc explainable AI approaches such as LIME, Captum, etc. to understand the features that influence a text to be predicted a gender stereotype about a male or a female. Also, in this work, due to the current lack of data to study non-binary gender stereotypes (Nozza et al., 2022), we focus on identifying the type of words prompting binary (male/female) gender stereotypes and the sentiment associated with those words.

Ethics Statement

We have handled all datasets and pre-processing in an ethical manner complying with the ACL code of ethics. Due to practical reasons and existing lack of datasets, we limited our research to only the binary genders. However, we understand the importance of inclusion and will consider extending our study, where possible, to non-binary genders.

References

Samira Abnar and Willem Zuidema. 2020. Quantifying attention flow in transformers. *arXiv preprint arXiv:2005.00928*.

Lauren Ackerman. 2019. Syntactic and cognitive issues in investigating gendered coreference. *Glossa: a journal of general linguistics*, 4(1).

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2623–2631.

Kendra Albert and Maggie Delano. 2022. Sex trouble: Sex/gender slippage, sex confusion, and sex obsession in machine learning using electronic health records. *Patterns*, 3(8).

Aliya Andrich and Emese Domahidi. 2022. A leader and a lady? a computational approach to detection of political gender stereotypes in facebook user comments. *International Journal of Communication*, 17:20.

Stefano Baccianella, Andrea Esuli, Fabrizio Sebastiani, et al. 2010. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *Lrec*, volume 10, pages 2200–2204.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Xuemei Bai. 2018. Text classification based on lstm and attention. In *2018 Thirteenth International Conference on Digital Information Management (ICDIM)*, pages 29–32. IEEE.

Sandra L Bem. 1974. The measurement of psychological androgyny. *Journal of consulting and clinical psychology*, 42(2):155.

Bethan Benwell. 2006. *Discourse and identity*. Edinburgh University Press.

Sharla D Biefeld, Ellen A Stone, and Christia Spears Brown. 2021. Sexy, thin, and white: The intersection of sexualization, body type, and race on stereotypes about women. *Sex Roles*, 85(5):287–300.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050*.

Markie LC Blumer, Mary S Green, Nicole L Thomte, and Parris M Green. 2013. Are we queer yet?: Addressing heterosexual and gender-conforming privilege. In *Deconstructing Privilege*, pages 151–168. Routledge.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Shikha Bordia and Samuel R Bowman. 2019. Identifying and reducing gender bias in word-level language models. *arXiv preprint arXiv:1904.03035*.

Cristina Bosco, Viviana Patti, Simona Frenda, Alessandra Teresa Cignarella, Marinella Paciello, and Francesca D'Errico. 2023. Detecting racial stereotypes: An italian social media corpus where psychology meets nlp. *Information Processing & Management*, 60(1):103118.

Gordon H Bower, John B Black, and Terrence J Turner. 1979. Scripts in memory for text. *Cognitive psychology*, 11(2):177–220.

Aylin Caliskan, Pimparkar Parth Ajay, Tessa Charlesworth, Robert Wolfe, and Mahzarin R Banaji. 2022. Gender bias in word embeddings: a comprehensive analysis of frequency, syntax, and semantics. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 156–170.

Yang Trista Cao and Hal Daumé III. 2019. Toward gender-inclusive coreference resolution. *arXiv preprint arXiv:1910.13913*.

Thomas F Cash and Timothy A Brown. 1989. Gender and body images: Stereotypes and realities. *Sex roles*, 21:361–373.

Jindong Chen, Yizhou Hu, Jingping Liu, Yanghua Xiao, and Haiyun Jiang. 2019. Deep short text classification with knowledge powered attention. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6252–6259.

Peter Clark, Bhavana Dalvi, and Niket Tandon. 2018. What happened? leveraging verbnet to predict the effects of actions in procedural text. *arXiv preprint arXiv:1804.05435*.

Marta R Costa-jussà. 2019. An analysis of gender bias studies in natural language processing. *Nature Machine Intelligence*, 1(11):495–496.

Jenna Cryan, Shiliang Tang, Xinyi Zhang, Miriam Metzger, Haitao Zheng, and Ben Y Zhao. 2020. Detecting gender stereotypes: lexicon vs. supervised learning methods. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–11.

Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. 2020. A survey of the state of explainable ai for natural language processing. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 447–459.

Alice H Eagly. 2013. *Sex differences in social behavior: A social-role interpretation*. Psychology Press.

Naomi Ellemers. 2018. Gender stereotypes. *Annual review of psychology*, 69:275–298.

Ethan Fast, Tina Vachovsky, and Michael Bernstein. 2016. Shirtless and dangerous: Quantifying linguistic signals of gender bias in an online fiction writing community. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 10, pages 112–120.

Anne Fausto-Sterling. 2019. Gender/sex, sexual orientation, and identity are in the body: How did they get there? *The Journal of Sex Research*, 56(4-5):529–555.

Joseph L Fleiss, Bruce Levin, Myunghee Cho Paik, et al. 1981. The measurement of interrater agreement. *Statistical methods for rates and proportions*, 2(212-236):22–23.

Daniel Fryer, Inga Strümke, and Hien Nguyen. 2021. Shapley values for feature selection: The good, the bad, and the axioms. *Ieee Access*, 9:144352–144360.

Ute Gabriel, Pascal Gygax, Oriane Sarrasin, Alan Garnham, and Jane Oakhill. 2008. Au pairs are rarely male: Norms on the gender perception of role names across english, french, and german. *Behavior research methods*, 40(1):206–212.

Aparna Garimella, Carmen Banea, Dirk Hovy, and Rada Mihalcea. 2019. Women's syntactic resilience and men's grammatical luck: Gender-bias in part-of-speech tagging and dependency parsing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3493–3498.

David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang. 2019. Xai—explainable artificial intelligence. *Science robotics*, 4(37):eaay7120.

Elizabeth L Haines, Kay Deaux, and Nicole Lofaro. 2016. The times they are a-changing… or are they not? a comparison of gender stereotypes, 1983–2014. *Psychology of Women Quarterly*, 40(3):353–363.

Philippe Hamon. 2004. What is a description? *Bal, M. Narrative Theory: Critical Concepts in Literary and Cultural Studies*, 1:309–340.

Shun Hattori, Taro Tezuka, and Katsumi Tanaka. 2007. Mining the web for appearance description. In *Database and Expert Systems Applications: 18th International Conference, DEXA 2007, Regensburg, Germany, September 3-7, 2007. Proceedings 18*, pages 790–800. Springer.

Nathan A Heflick, Jamie L Goldenberg, Douglas P Cooper, and Elisa Puvia. 2011. From women to objects: Appearance focus, target gender, and perceptions of warmth, morality and competence. *Journal of Experimental Social Psychology*, 47(3):572–581.

Amaç Herdağdelen and Marco Baroni. 2011. Stereotypical gender actions can be extracted from web text. *Journal of the American Society for Information Science and Technology*, 62(9):1741–1749.

Andreas Holzinger. 2018. From machine learning to explainable ai. In *2018 world symposium on digital intelligence for systems and machines (DISA)*, pages 55–66. IEEE.

Alexander Hoyle, Hanna Wallach, Isabelle Augenstein, Ryan Cotterell, et al. 2019. Unsupervised discovery of gendered language through latent-variable modeling. *arXiv preprint arXiv:1906.04760*.

Hairong Huo and Mizuho Iwaihara. 2020. Utilizing bert pretrained models with various fine-tune methods for subjectivity detection. In *Web and Big Data: 4th International Joint Conference, APWeb-WAIM 2020, Tianjin, China, September 18-20, 2020, Proceedings, Part II 4*, pages 270–284. Springer.

Eni Ismayanti and Abdul Kholiq. 2020. An analysis of students'difficulties in writing descriptive text. *E-link Journal*, 7(1):10–20.

Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. In *Proceedings of NAACL-HLT*, pages 3543–3556.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. *arXiv preprint arXiv:1906.07337*.

Melissa Ann Lavin and Thomas F Cash. 2001. Effects of exposure to information about appearance stereotyping and discrimination on women's body images. *International Journal of Eating Disorders*, 29(1):51–58.

Zachary C Lipton. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57.

Zhenyu Liu, Haiwei Huang, Chaohong Lu, and Shengfei Lyu. 2020. Multichannel cnn with attention for text classification. *arXiv preprint arXiv:2006.16174*.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc.

Sherin Mary Mathews. 2019. Explainable artificial intelligence applications in nlp, biomedical, and malware classification: a literature review. In *Intelligent Computing: Proceedings of the 2019 Computing Conference, Volume 2*, pages 1269–1292. Springer.

Joanne McDowell. 2015. Masculinity and non-traditional occupations: Men's talk in women's work. *Gender, Work & Organization*, 22(3):273–291.

Samin Mohammadi and Mathieu Chapon. 2020. Investigating the performance of fine-tuned text classification models based-on bert. In *2020 IEEE 22nd International Conference on High Performance Computing and Communications; IEEE 18th International Conference on Smart City; IEEE 6th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*, pages 1252–1257. IEEE.

Elizabeth M Morgan and Laurel R Davis-Delano. 2016. How public displays of heterosexual identity reflect and reinforce gender stereotypes, gender differences, and gender inequality. *Sex Roles*, 75(5-6):257–271.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. Stereoset: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371.

Debora Nozza, Federico Bianchi, Anne Lauscher, Dirk Hovy, et al. 2022. Measuring harmful sentence completion in language models for lgbtqia+ individuals. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*. Association for Computational Linguistics.

Jahna Otterbacher. 2015. Crowdsourcing stereotypes: Linguistic bias in metadata generated via gwap. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 1955–1964.

Flor Miriam Plaza-del Arco, Amanda Cercas Curry, Alba Curry, Gavin Abercrombie, and Dirk Hovy.

2024. Angry men, sad women: Large language models reflect gendered stereotypes in emotion attribution. *arXiv preprint arXiv:2403.03121*.

Marcelo OR Prates, Pedro H Avelar, and Luís C Lamb. 2020. Assessing gender bias in machine translation: a case study with google translate. *Neural Computing and Applications*, 32:6363–6381.

Rukhma Qasim, Waqas Haider Bangyal, Mohammed A Alqarni, and Abdulwahab Ali Almazroi. 2022. A fine-tuned bert-based transfer learning approach for text classification. *Journal of healthcare engineering*, 2022.

P Rosenkrantz, H Bee, S Vogel, and I Broverman. 1968. Sex-role stereotypes and self-concepts in college students. *Journal of Consulting and Clinical Psychology*, 32(3):287–295.

Elisa Rubegni, Monica Landoni, Antonella De Angeli, and Letizia Jaccheri. 2019. Detecting gender stereotypes in children digital storytelling. In *Proceedings of the 18th ACM International Conference on Interaction Design and Children*, pages 386–393.

Donald L Rubin and Kathryn L Greene. 1991. Effects of biological and psychological gender, age cohort, and interviewer gender on attitudes toward gender-inclusive/exclusive language. *Sex Roles*, 24:391–412.

Cynthia Rudin and Joanna Radin. 2019. Why are we using black box models in ai when we don't need to? a lesson from an explainable ai competition. *Harvard Data Science Review*, 1(2):10–1162.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of NAACL-HLT*, pages 8–14.

Anthony J Sanford and Simon C Garrod. 1998. The role of scenario mapping in text comprehension. *Discourse processes*, 26(2-3):159–190.

Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Gender bias in machine translation. *Transactions of the Association for Computational Linguistics*, 9:845–874.

Gün R Semin and Klaus Fiedler. 1988. The cognitive functions of linguistic categories in describing persons: Social cognition and language. *Journal of personality and Social Psychology*, 54(4):558.

Sofia Serrano and Noah A Smith. 2019. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2020. Towards controllable biases in language generation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3239–3254.

Oksana M Shyian, Larysa F Foster, Tatiana M Kuzmenko, Larysa V Yeremenko, and Nina P Liesnichenko. 2021. Socio-psychological criteria of the formation of gender stereotypes of appearance. *Journal of Intellectual Disability-Diagnosis and Treatment*, 9:651–666.

Peter Siemund and Florian Dolberg. 2011. From lexical to referential gender: An analysis of gender change in medieval english based on two historical documents.

Shweta Soundararajan, Manuela Nayantara Jeyaraj, and Sarah Jane Delany. 2023. Using chatgpt to generate gendered language. In *2023 31st Irish Conference on Artificial Intelligence and Cognitive Science (AICS)*, pages 1–8. IEEE.

T Spence Janet and Stapp Joy. 1974. The personal attributes questionnaire: A measure of sex-role stereotypes and masculinity-femininity. In *Journal Supplement Abstract Service: Catalog of Selected Documents in Psychology*, volume 4, page 43.

Gabriel Stanovsky, Noah A Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684.

Xiaobing Sun and Wei Lu. 2020. Understanding attention for text classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3418–3428.

Omer Tal, Yang Liu, Jimmy Huang, Xiaohui Yu, and Bushra Aljbawi. 2019. Neural attention frameworks for explainable recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 33(5):2137–2150.

Wouter Van Atteveldt, Tamir Sheafer, Shaul R Shenhav, and Yair Fogel-Dror. 2017. Clause analysis: Using syntactic information to automatically extract source, subject, and predicate from texts with an application to the 2008–2009 gaza war. *Political Analysis*, 25(2):207–222.

Shikhar Vashishth, Shyam Upadhyay, Gaurav Singh Tomar, and Manaal Faruqui. 2019. Attention interpretability across nlp tasks. *arXiv preprint arXiv:1909.11218*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. *Advances in neural information processing systems*, 33:12388–12401.

Giulia Vilone and Luca Longo. 2020. Explainable artificial intelligence: a systematic review. *arXiv preprint arXiv:2006.00093*.

Michael Wiegand, Josef Ruppenhofer, and Elisabeth Eder. 2021. Implicitly abusive language–what does it actually look like and why are we not getting there? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 576–587.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20.

Zhang Xinxi. 2021. Single task fine-tune bert for text classification. In *2nd International Conference on Computer Vision, Image, and Deep Learning*, volume 11911, pages 434–439. SPIE.

Feiyu Xu, Hans Uszkoreit, Yangzhou Du, Wei Fan, Dongyan Zhao, and Jun Zhu. 2019. Explainable ai: A brief survey on history, research areas, approaches and challenges. In *Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part II 8*, pages 563–574. Springer.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR.

Shenglu Ye, Simin Cai, Chuansheng Chen, Qun Wan, and Xiuying Qian. 2018. How have males and females been described over the past two centuries? an analysis of big-five personality-related adjectives in the google english books. *Journal of Research in Personality*, 76:6–16.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. In *Proceedings of NAACL-HLT*, pages 629–634.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 2.

Jinfeng Zhong and Elsa Negre. 2022. Shap-enhanced counterfactual explanations for recommendations. In *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing*, pages 1365–1372.