Detecting Gender Stereotypical Language using Model-agnostic and Model-specific Explanations

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

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

Abstract

AI models learn gender-stereotypical language from human data. So, understanding how well different explanation techniques capture diverse language features that suggest gender stereotypes in text can be useful in identifying stereotypes that could potentially lead to gender bias. The influential words identified by four explanation techniques (LIME, SHAP, Integrated Gradients (IG) and Attention) in a gender stereotype detection task were compared with words annotated by human evaluators. All techniques emphasized adjectives and verbs related to characteristic traits and gender roles as the most influential words. LIME was best at detecting explicitly gendered words, while SHAP, IG and Attention showed stronger overall alignment and considerable overlap. A combination of these techniques, combining the strengths of model-agnostic and model-specific explanations, performs better at capturing gender-stereotypical language. Extending to hate speech and sentiment prediction tasks, annotator agreement suggests these tasks to be more subjective while explanation techniques can better capture explicit markers in hate speech than the more nuanced gender stereotypes. This research highlights the strengths of different explanation techniques in capturing subjective gender stereotype language in text.

1 Introduction

Gender stereotypes refer to commonly held beliefs and expectations about the characteristic traits, roles, and behaviours of groups of individuals based on their gender (Ellemers, 2018). These stereotypes often arise from subconscious beliefs that could influence how an individual perceives another based on established culture and social norms.

These stereotypes can potentially lead to gender bias when a particular individual is believed to conform to or contradict the established beliefs that individual should appear or behave based on their gender. For example, women are often perceived as more nurturing and empathetic, while men are perceived as strong and decisive (Cuddy et al., 2008). Although these traits may appear advantageous (positive gender stereotypes), they can limit individuals by confining them to specific roles due to preconceived beliefs. Negative gender stereotypes are the more harmful or restrictive perceptions. A common negative gender stereotype is the idea that women are too emotional to be effective leaders (Rudman and Glick, 2021). For instance, in the study conducted by (Andrich and Domahidi, 2022) on the descriptions of the US political candidates, they showed that the users' facebook comments described the male candidates with stronger masculine traits associated with a career in politics than the female candidates.

In the past, machine learning models have always been black-boxes until the concept of explainability, the ability to interpret these models as glass boxes, was introduced (Holzinger, 2018; Rudin and Radin, 2019). With post-hoc XAI approaches on natural language prediction systems, the most influential words in a prediction task can be identified by a score assigned to each word which is indicative of each word's contribution to the model's prediction.

The aim of this paper is to evaluate how different explanation approaches that produce these word-level scores, used on a gender stereotype detection model align with human perceptions of language that suggest gender stereotyping. This can assist with identifying gender stereotypes that could potentially lead to gender bias in text.

The words identified as the most influential by four explanation approaches that return word-level importance scores (LIME, SHAP, attention and Integrated gradients) were analyzed against the ground truth annotated by humans as indicative

Proceedings of Recent Advances in Natural Language Processing,pages 481–490
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-057

of gender stereotypes using an overlap measure, the Jaccard index.

While no technique performed perfectly, an analysis of the influential words from different explanation approaches showed that a local XAI explanation-based approach such as LIME identified different words than global explanation-based approaches such as SHAP, or model-specific approaches like attention and IG. A combination of explanation approaches performed better than individual approaches.

We found that human evaluators' consensus on the ground truth suggested that gender stereotype detection was a subjective task as the annotators had diverse views on what they considered as words suggesting gender stereotype in a particular text. We observed how these explanation approaches performed on other subjective tasks including hate speech detection and sentiment analysis. The consensus across the ground truth for these tasks showed that they are more subjective than gender stereotype detection. However, these explanation techniques performed better at capturing words that aligned with what human evaluators considered to be hate speech than they did for gender stereotypes. This may be due to the strong word markers used in hate speech.

The rest of the paper is structured as follows; section 2 covers related work, section 3 details the methodology used, while section 4 presents the results and analysis. The paper concludes in section 5.

2 Background

Most of the major workplace discrimination is based on gender stereotypes. For example, discrimination against female political candidates due to the use of gender stereotypical language may lead to influencing the voters decision to choose a male candidate (Andrich and Domahidi, 2022).

While gender stereotypes refer to general social beliefs about gender roles, gender bias represents the partiality that stems from these stereotypes (Avitzour et al., 2020). As such, gender stereotypes are the underlying beliefs, while bias is the action or consequence that results from these beliefs (Ellemers, 2018). For example, the stereotype that men are better at mathematics can lead to gender bias against women in hiring for technical roles (Vuletich et al., 2020).

Due to the strong negative impact that gender bias has, much of the previous research in this area has been on identifying and mitigating gender bias. Researchers have investigated gender bias within the domain of 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), co-reference 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), and more.

Previously, researchers have used lexicon-based methods to analyze gender stereotypes using predefined gender-specific word lists derived from psychological studies (Bem, 1974; Rosenkrantz et al., 1968; Spence Janet and Joy, 1974). Some approaches extracted verb-noun pairs from databases like OMCS (Open Mind Common Sense) (Singh et al., 2002) to examine stereotypical associations such as "women" with "cooking", "men" with "building", etc. (Herdağdelen and Baroni, 2011).

More recently, machine learning has been used to train a model that classified texts based on whether an individual's description aligned or contradicted the expected gender norms of the subject to explore how gender expectations are reflected in individual descriptions (Cryan et al., 2020). A dataset of web posts and news articles containing descriptions of people was complied and annotated by crowd-sourcing to determine whether the annotators considered the descriptions to be consistent with or contradictory to the gender stereotype of the subject. The study listed the most frequent words that contributed to gender-conforming and gender-non-conforming predictions.

Most of the previous research that used XAI approaches explored explicit bias (De Keijzer, 2025; Mehta and Passi, 2022; Hofeditz et al., 2022) and not the implicit bias that comes in the form of gender stereotypes. And despite the growing interest in the field of XAI techniques, not many works have conducted human-centered evaluations which meant that there is a gap between what the XAI techniques claimed and what humans considered as actual explanations (Suh et al., 2025).

However, the work by (Jeyaraj and Delany, 2024) has applied explainability techniques to identify words associated with gender stereotypes in text using SHAP and attention. Using a BERT-

based transformer model, the study classified text for gender stereotypes and introduced an influence score combining SHAP values and attention weights. This approach was used to analyze the linguistic patterns of influential words linked to gender stereotypes. The findings showed that certain word types are strongly associated with such stereotypes, highlighting the potential of explanation methods to identify language that may contribute to gender bias in text.

3 Approach

The aim of this work is to evaluate how different explanation approaches used on a gender stereotype detection model align with the words that humans identify as those that suggest gender stereotyping. The approach used firstly involved training a transformer model to detect gender stereotype text. Explanation techniques including SHAP, LIME, Integrated gradients and Attention were used to get the most influential words for the correctly predicted gender stereotype texts. These words were evaluated against human-annotated ground truth.

3.1 Dataset

The dataset used in this research, referred to as the CR dataset, consists of short text statements that are either consistent or contradict gender stereotypes. These texts were originally compiled and annotated through crowd-sourcing by Cryan et al. (2020), who asked annotators to label each statement as gender-conforming or gender-non-conforming and to provide reasons for their decisions. In the work conducted by Jeyaraj and Delany (2024), the authors validated these reasons as texts conveying gender stereotypes, labelled 'GS' or anti-gender stereotypes, labelled 'anti-GS', using crowd-sourcing. The anti-GS texts consist of both neutral texts that do not reinforce or challenge gender stereotypes as well as texts that actively counter and challenge gender stereotypes and intentionally highlight non-stereotypical behaviours (e.g. "That man was a compassionate nurse who loves taking care of children."). To maintain a balanced training set across male/female GS and anti-GS categories, half of the male and female anti-GS samples and approximately 20% of the GS samples were used for training, resulting in an unconventional train-test split of around 27.9% - 72.1% (≈ 30/70). This dataset is shown in Table 1.

A BERT-based transformer model was trained on this binary text classification task to predict whether a text was a gender stereotype / anti-gender stereotype[1] This achieved a precision of 0.66, recall of 0.81 and f1-score of 0.71 on the test set.

Dataset	Size	Gender Stereotype		Anti-Gender Stereotype	
		Male	Female	Male	Female
Whole CR dataset	2818	2818 (100%)			
		2198 (78%)		620 (22%)	
		1155 (41%)	1043 (37%)	282 (10%)	338 (12%)
Training set	789	254 (9%)	226 (8%)	141 (5%)	169 (6%)
Test set	2029	902 (32%)	818 (29%)	141 (5%)	169 (6%)

Table 1: CR dataset used for training and testing the transformer model. Sample counts are shown first, followed by percentages (based on a total of 2818 samples).

3.2 Building the ground truth datasets

Each correctly predicted gender stereotype text from the test set was annotated by crowd workers on Amazon MTurk[2]. Annotators were asked to highlight the words in the text that made them consider the text as a gender stereotype. Each text was annotated by three annotators. Samples that were incomplete and samples that had been carelessly highlighted (annotations including highlights of solely white spaces or partial words highlighted across two or more words) were removed. 1371 texts that had three valid annotations were retained.

The consensus agreement across annotators for each text instance was calculated using the Jaccard agreement (A). This measures the overlap between two sets of words, S_1 and S_2, as shown in equation 1.

$$A(S_1, S_2) = \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|} \quad (1)$$

Two datasets (D) were extracted from the set of 1371 valid texts based on the level of agreement among the annotators for each text. The first dataset called the weak consensus dataset (D_{weak}) consists of 1328 texts. Here, each text has a set of words where at least two out of three annotators agreed on the words selected. The second dataset, called the strong consensus dataset (D_{strong}), includes 540 texts, where all three annotators agreed on the words selected. It is significantly smaller than D_{weak}.

[1]The model was a bert-base-uncased transformer model where we tuned the learning-rate, batch size, epochs, max sequence length and weight decay using Bayesian optimization with optuna.

[2]Amazon MTurk: https://www.mturk.com/

The average agreement ($A(D)$) across a dataset (D) is the overall annotator agreement across that dataset as shown in equation 2.

$$A(D) = \frac{1}{|D|} \sum_{i=1}^{|D|} A(S_1, S_2)_i \quad (2)$$

$A(D_{\text{weak}})$ was calculated as 0.57 where S_1 is the set of words selected by at least two out of three annotators for each text and S_2 is the set of words selected by all three annotators for each text.

$A(D_{\text{strong}})$ was calculated as 0.34, where S_1 is the set of words selected by all 3 annotators for a text i and S_2, similar to $A(D_{\text{weak}})$, is the set of words selected by all three annotators for text i. $A(D_{\text{weak}})$ is higher than the $A(D_{\text{strong}})$ as there is more agreement when only two annotators have to agree than when all three have to agree. This suggests that gender stereotype detection is a subjective task with varying views on what humans consider as language suggestive of gender stereotypes in text.

3.3 Analysing the ground truth datasets

To understand what humans perceive as gender-stereotypical language, the words selected by the annotators from D_{weak} and D_{strong} datasets were analysed. The frequency distribution of the top words selected by the annotators in the D_{weak} and D_{strong} ground truth datasets is shown in Figure 1.

In D_{strong}, the frequency distribution of words unanimously agreed upon by the annotators is significantly skewed as the common words are nearly all explicitly (or lexically) gendered words such as "women" and "men". The weak consensus ground truth dataset, which considered words selected by at least two annotators, is less heavily skewed and includes more words that are are not explicitly or lexically gendered.

In spite of the the skew in the frequency distribution, most of the words identified by the annotators as words suggesting gender stereotypes were non-gendered words: 95% and 87% of D_{weak} and D_{strong} datasets respectively. This shows that humans rely heavily on non-gendered words in recognizing gender stereotypes. This suggests that gender bias is often conveyed indirectly through context rather than explicit gender markers. Lexically gendered words like "manly", "female", "mother", "husband" and "feminine" have higher occurrence in D_{strong} dataset (12.7%) than the D_{weak} (5.5%) suggesting that there is more agreement across explicitly gendered words.

The distribution of the male to female words

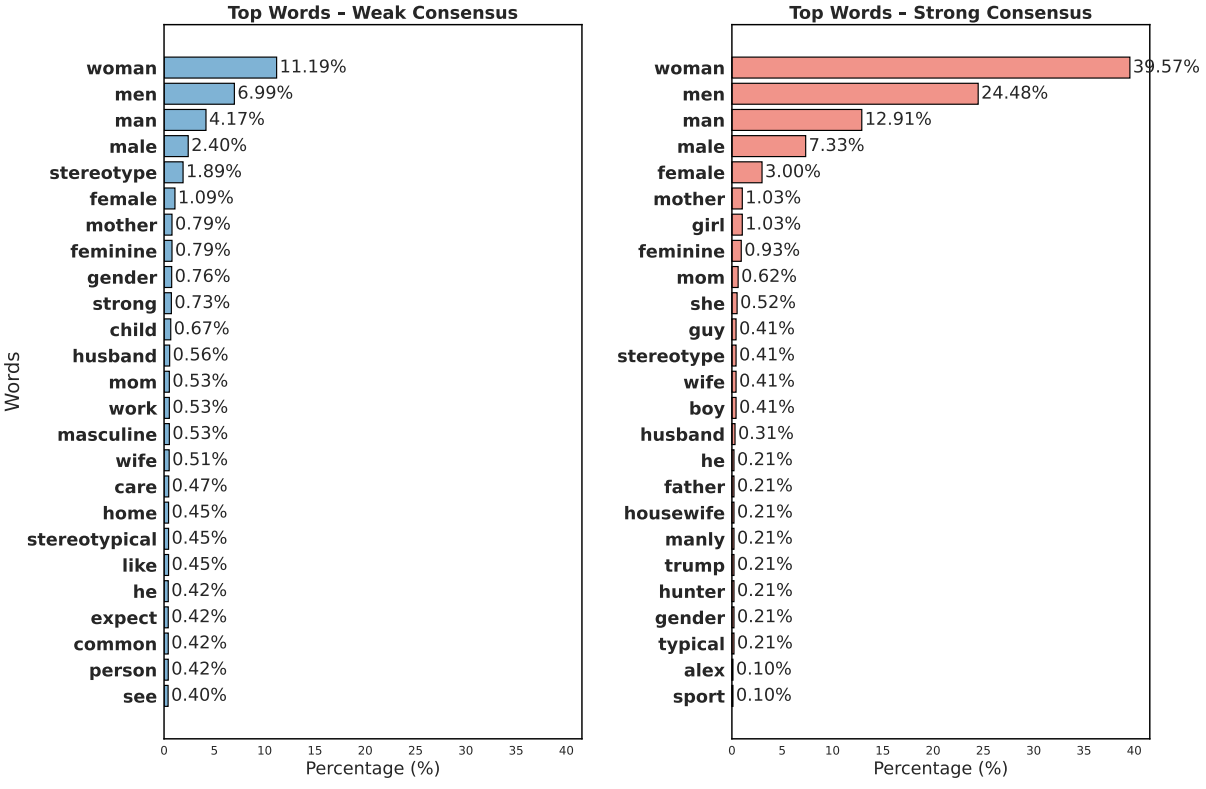

Figure 1: Top word frequency distribution for D_{weak} and D_{strong} ground truths (CR – GS Dataset)

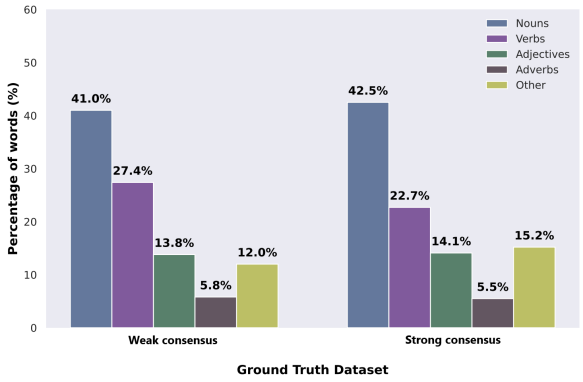

Figure 2: Percentage of different POS types in the weak and strong consensus ground truth datasets.

within the lexically gendered words is 44%/56% and 46%/54% for D_{weak} and D_{strong} datasets respectively suggesting that humans are more sensitive to gender stereotypical language about women.

Figure 2 shows the distribution of the different parts of speech (POS) types of words across D_{weak} and D_{strong}. In both, annotators agree that nouns are more suggestive of gender-stereotypical language in text. This is followed by verbs which describe actions and behaviour, reflecting gender role descriptors, and adjectives that describe characteristic traits, reflecting gender expression descriptors. This indicates that verbs and adjectives play a significant role in perceiving gender stereotypes confirming previous findings by Jeyaraj and Delany (2024).

3.4 Extracting the most influential words

Post-hoc XAI approaches for text classification tasks analyze and emphasize which specific components of a given text (words, phrases, or sentence structures) contribute to the model's decision-making process. The post-hoc techniques used in this work include SHAP (Lundberg and Lee, 2017) and LIME (Ribeiro et al., 2016), along with model-specific approaches like Integrated Gradients (IG) (Sundararajan et al., 2017) and the Attention mechanism (Vaswani et al., 2017; Serrano and Smith, 2019), which provide insights into a model's inner workings.

While LIME and SHAP are both model-agnostic approaches which means they do not rely on the internal architecture of the model, IG requires access to the model's internal gradients and Attention is inherently part of the model's architecture rather than a post-hoc explanation. Therefore, IG and Attention can deliver model-specific explanations.

LIME explains predictions by training a local surrogate model such as a linear model around the prediction of a single text instance. It treats the original model as a black box, relying only on input-output behaviour without using any internal architecture of the original model. Therefore, its explanations are local and capture how the prediction changes within the context of that input.

SHAP treats the model as a black box but does so by estimating a value called Shapley value to quantify the contribution of each word. It begins with a baseline prediction, like the model's output on an empty or neutral text, and then gradually changes the input by removing or changing words. In doing so, SHAP determines the relative importance of each word based on how these modifications influence the prediction. Therefore, its aggregations across instances gives more globally consistent value of word importance.

In contrast, IG and attention are model-specific methods. IG computes attributions by integrating the gradients of the model's output with respect to the input along a path from a baseline to the actual input. This requires full access to the model's internal gradients. IG gives a global feature importance of words.

Attention is a mechanism which is part of the transformer model. It determines how much focus each input word in the text receives during prediction. In previous work, research has contested the idea that attention is a reliable explanation (Abnar and Zuidema, 2020). However, attention weights are instance-level and model-specific, directly reflecting the learned behaviour of the model therefore capturing a different aspect of the word's contribution to the model's predictions.

Each of these techniques were applied to each instance in the ground truth datasets and returned a score for each word in the text. This score represents the contribution of that word to the model's prediction. Scores for stopwords and special characters were not considered.

To identify the most influential words, 90% of each dataset was used to tune a word score threshold for each method. The threshold that achieved the highest agreement with the ground truth on the remaining 10% of the data was selected. The influential words for a technique are those above the identified threshold.

4 Results and Discussion

Figure 3 shows the agreement between the most influential words from each explanation technique and D_{weak}/D_{strong} ground truth datasets measured using Equation 2 with S_1 representing the set of D_{weak}/D_{strong} words for a text and S_2 being the corresponding influential words according to a particular explanation technique for that text.

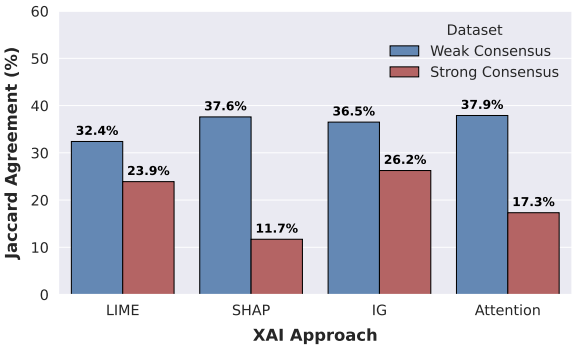

Figure 3: Agreement between ground truth and words captured by each explanation method.

This agreement is higher and more stable for the weak consensus ground truth than the strong consensus dataset. In general, all the explanation methods perform better in capturing weak consensus ground truth words than the strong consensus words. However, the small size of the strong consensus dataset may contribute to this.

The distribution of the frequency of influential words captured by the different methods is shown in Figure 4.

Figure 4 shows that LIME captures fewer words per text than the other three methods. None of the explanation techniques match the distribution of the human annotators who, unsurprisingly, generally choose the fewest words.

More than 95% of influential words identified by all methods are non-gendered words, ie. non lexically gendered. Interestingly, gender pronouns such as "he", "she", "her", etc. are rarely captured by the XAI methods. LIME selects the most lexically gendered words, which is still significantly less than human annotators.

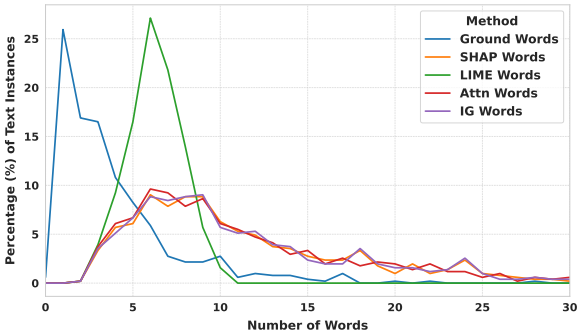

Figure 4: Word count distribution of most influential words.

The D_{weak} ground truth includes a slightly higher proportion of female-gendered words (∼57%) compared to male-gendered words (∼43%). However, the explanation techniques identified gendered words in near balanced proportions, ranging between 48-50% male and 50-52% female. This indicates that while human evaluators demonstrate a slight bias in associating gender stereotypes more with females than males, explanation methods do not reflect as much difference in this regard.

Figure 5 shows the agreement between words captured by different explanation methods for the gender stereotype texts measured using Equation 2 where S_1 and S_2 are the set of words identified by the two corresponding techniques compared.

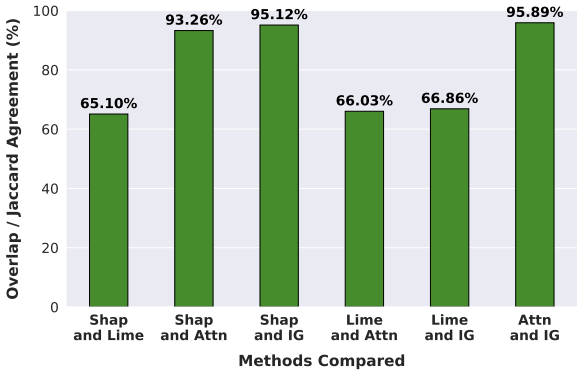

Figure 5: Agreement between words captured by different methods.

Overall, while SHAP, attention and IG show high agreement with each other, LIME aligns less strongly with the others. This may be due to the difference in LIME providing local explanations, interpreting individual predictions. Although attention is local to the instance, it learns during training to focus on semantically or syntactically important tokens. This learning often aligns attention with importance measures from gradient-based methods. And this may account for the high overlap between the influential words according to attention and global explanations like IG or SHAP.

We explored how well a combination of influential words from multiple explanation techniques would perform at aligning with what humans per-

ceive as words suggesting gender stereotype in text. We explored different combinations of 2 to all 4 techniques. The set of most influential SHAP words (S_{SHAP}) and attention words (S_{ATTN}), combined with the set of most influential words found by LIME (S_{LIME}) but not present in the set of words found by IG (S_{IG}) had the best agreement of 38.24% with the ground truth, as shown in Equation 3.

$$S_{\text{GS}} = S_{\text{SHAP}} \cup S_{\text{ATTN}} \cup (S_{\text{LIME}} \setminus S_{\text{IG}}) \quad (3)$$

This was followed very closely by an agreement of 38% by combining S_{SHAP} and S_{ATTN} words alone. This suggests that although LIME has the lowest agreement with the ground truth on its own, it may capture meaningful words that were not captured by the other methods.

4.1 Comparing explanation techniques on other subjective tasks

We have shown above that gender stereotype detection is a subjective task based on the annotator agreement. We compared how the words highlighted by XAI techniques match with human-annotated words for other tasks such as hate speech detection and sentiment analysis. To observe if the explanation techniques aligned with human judgement in identifying words that influenced prediction across tasks other than gender stereotype detection, we considered other tasks including hate speech detection and sentiment analysis.

For hate speech detection, a subset of the Hate Speech 18 dataset (HS18) (de Gibert et al., 2018) and for sentiment analysis, a subset of the IMDB movie reviews datasets (Maas et al., 2011) were used. Details about these datasets are presented in Table 2.

Dataset	Label	Train Set (%)	Test Set (%)
HS18	No Hate	50% (500 / 1000)	50% (500 / 1000)
	Hate	50% (500 / 1000)	50% (500 / 1000)
IMDB	Negative	50% (1000 / 2000)	50% (1000 / 2000)
	Positive	50% (1000 / 2000)	50% (1000 / 2000)

Table 2: Hate Speech 18 (HS18) and IMDB Movie review (IMDB) dataset descriptions used for training and testing the transformer model.

Each dataset was split 50%/50% for training and testing. The same transformer model architecture used for the gender stereotype detection task was used to build models to predict hate/no hate text and positive and negative sentiment respectively. Hyperparameters were tuned individually for each task. Table 3 shows the performance of the classifier on these tasks.

Task	Class	Precision	Recall	f1-score
Hate speech detection	Hate	0.80	0.84	0.82
	No hate	0.84	0.79	0.81
Sentiment analysis	Positive	0.90	0.85	0.87
	Negative	0.86	0.91	0.88

Table 3: Classifier performance on other tasks.

The correctly predicted hate-labeled, positive and negative movie review texts from these tasks were annotated on Amazon MTurk by 3 annotators following a similar process to that of the gender stereotype annotation task. D_{weak} and D_{strong} ground truth datasets for these tasks were created, similar to that of the gender stereotype detection task, details of each are shown in Table 4.

Task	Total	D_{weak}	D_{strong}
Hate Speech	420	417 (99.3%)	150 (35.7%)
Sentiment (Pos)	850	92 (10.8%)	20 (2.4%)
Sentiment (Neg)	910	379 (41.7%)	89 (9.8%)

Table 4: Distribution of D_{weak} and D_{strong} ground truth subsets for each task. (Sample counts are shown first, followed by percentages based on the respective total samples).

Figure 6 shows the agreement between the annotators on the D_{weak} and D_{strong} datasets for the hate speech and sentiment analysis computed using Equation 2 and compared with the gender stereotype detection ground truth.

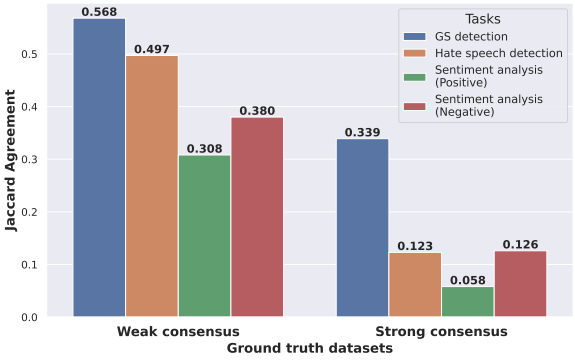

Figure 6: Agreement between annotators for different subjective tasks.

Figure 6 shows that identifying words that indicate positive sentiment is the most subjective of the

tasks evident due to the lowest agreement among annotators. Human evaluators seem to have diverse views on what suggests a text to be considered a positive or negative sentiment although they agree more on what words influence negative than positive sentiment. This is followed by hate speech detection being moderately subjective and gender stereotype detection being less subjective than the other two tasks.

The thresholds for the explanations' word scores were tuned in a similar way to the gender stereotype detection task to identify the set of influential words from each explanation approach for each task.

Figure 7 shows the agreements between the influential words of the different explanation approaches against their respective ground truths for all tasks.

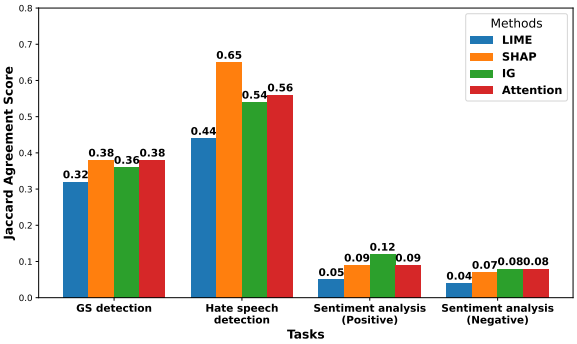

Figure 7: Jaccard agreement of most influential XAI words with their ground truth words.

All the explanation methods perform well on hate speech detection. We found that hate speech detection is a subjective task according to human annotators. Humans selected a wide variety of words that suggest hate speech not necessarily agreeing on what they consider as hate speech as much as we might expect. Explanation techniques performed well, tending to highlight obvious strong marker words such as bitch, homo, kill, etc.. Similar to the gender stereotype detection task, LIME performs poorly in capturing words suggesting hate speech or negative/positive sentiment compared to the other three approaches. All techniques perform poorly in capturing words that align with human ground truth for the sentiment analysis task which is not surprising as this task was found to be very subjective according to the low agreement between the annotators as shown in Figure 6.

5 Conclusion

To identify language indicative of gender stereotypes, we trained a transformer model to classify gender stereotype and anti-gender stereotype texts and compiled a ground truth dataset of words for the correctly predicted gender stereotype texts using crowd-workers on Amazon MTurk. According to human evaluators, gender stereotype detection is a highly subjective task with varying views on gender-stereotypical language in text.

We applied model-agnostic eXplainable AI (XAI) approaches such as SHAP and LIME, along with model-specific explanations such as Integrated Gradients, and Attention on these correctly predicted gender stereotype texts to identify the most influential words in the classification task. The words identified by each approach were compared against the ground truth using the average overlap agreement across text samples. Our findings showed that while all methods mainly identified non-gendered words, LIME captures the most gendered words. This could be due to LIME's local explanation approach allowing it to identify words missed by the other methods which have significant overlap among the words they capture.

We combine the strengths of model-agnostic and model-specific explanation in identifying words suggesting gender stereotype in text. The proposed approach incorporates influential words from SHAP and Attention, along with words from LIME not captured by Integrated Gradients. This combination showed the best agreement with the ground truth dataset achieving an agreement of 38.24%.

Further analysis across other tasks such as hate speech detection and sentiment analysis showed that all four explanation approaches performed well in identifying words related to hate speech. This could be attributed to the presence of strong, objective word markers suggesting hate speech.

In conclusion, this study provides a detailed comparison of explanation methods in the context of gender stereotype detection. It presents the importance of combining different approaches to improve interpretability and highlights the challenges posed by the subjectivity of the gender stereotype detection task.

References

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

arXiv:2005.00928.

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.

Eliana Avitzour, Adi Choen, Daphna Joel, and Victor Lavy. 2020. On the origins of gender-biased behavior: The role of explicit and implicit stereotypes. Technical report, National Bureau of Economic Research.

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

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*.

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*.

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.

Amy JC Cuddy, Susan T Fiske, and Peter Glick. 2008. Warmth and competence as universal dimensions of social perception: The stereotype content model and the bias map. *Advances in experimental social psychology*, 40:61–149.

Adrianus De Keijzer. 2025. Identifying and analyzing provocative text: An xai approach to classification and feature selection.

Naomi Ellemers. 2018. Gender stereotypes. *Annual review of psychology*, 69(1):275–298.

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.

Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate Speech Dataset from a White Supremacy Forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 11–20, Brussels, Belgium. Association for Computational Linguistics.

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.

Lennart Hofeditz, Sünje Clausen, Alexander Rieß, Milad Mirbabaie, and Stefan Stieglitz. 2022. Applying xai to an ai-based system for candidate management to mitigate bias and discrimination in hiring. *Electronic Markets*, 32(4):2207–2233.

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.

Manuela Jeyaraj and Sarah Delany. 2024. An explainable approach to understanding gender stereotype text. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 45–59.

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

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.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Harshkumar Mehta and Kalpdrum Passi. 2022. Social media hate speech detection using explainable artificial intelligence (xai). *Algorithms*, 15(8):291.

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.

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.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*.

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.

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.

Laurie A Rudman and Peter Glick. 2021. *The social psychology of gender: How power and intimacy shape gender relations*. Guilford Publications.

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.

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.

Push Singh, Thomas Lin, Erik T Mueller, Grace Lim, Travell Perkins, and Wan Li Zhu. 2002. Open mind common sense: Knowledge acquisition from the general public. In *On the Move to Meaningful Internet Systems 2002: CoopIS, DOA, and ODBASE: Confederated International Conferences CoopIS, DOA, and ODBASE 2002 Proceedings*, pages 1223–1237. Springer.

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.

Ashley Suh, Isabelle Hurley, Nora Smith, and Ho Chit Siu. 2025. Fewer than 1% of explainable ai papers validate explainability with humans. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–7.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.

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.

Heidi A Vuletich, Beth Kurtz-Costes, Erin Cooley, and B Keith Payne. 2020. Math and language gender stereotypes: Age and gender differences in implicit biases and explicit beliefs. *Plos one*, 15(9):e0238230.

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.