Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI-23)
Special Track on AI for Good

# Promoting Gender Equality through Gender-biased Language Analysis in Social Media

**Gopendra Singh** , **Soumitra Ghosh** and **Asif Ekbal**
Department of Computer Science and Engineering, Indian Institute of Technology Patna, Patna, India
{gopendra.99, ghosh.soumitra2}@gmail.com, asif@iitp.ac.in

## Abstract

Gender bias is a pervasive issue that impacts women's and marginalized groups' ability to fully participate in social, economic, and political spheres. This study introduces a novel problem of Gender-biased Language Identification and Extraction (GLIdE) from social media interactions and develops a multi-task deep framework that detects gender-biased content and identifies connected causal phrases from the text using emotional information that is present in the input. The method uses a zero-shot strategy with emotional information and a mechanism to represent gender-stereotyped information as a knowledge graph. In this work, we also introduce the first-of-its-kind Gender-biased Analysis Corpus (GAC) of 12,432 social media posts and improve the best-performing baseline for gender-biased language identification and extraction tasks by margins of 4.88% and 5 ROS points, demonstrating this through empirical evaluation and extensive qualitative analysis. By improving the accuracy of identifying and analyzing gender-biased language, this work can contribute to achieving gender equality and promoting inclusive societies, in line with the United Nations Sustainable Development Goals (UN SDGs) and the Leave No One Behind principle (LNOB). We adhere to the principles of transparency and collaboration in line with the UN SDGs by openly sharing our code and dataset.

## 1 Introduction

Gender equality is a key component of sustainable development. Gender-biased language (GL) can reinforce gender stereotypes and discrimination against women and girls, and hinder efforts to promote gender equality and sustainable development. In some cultures, GL is deeply ingrained in the language and is considered normal. In developing and underdeveloped countries, women and girls, as well as other marginalised groups, may face greater challenges in accessing education, healthcare, and other services. Furthermore, GL can hinder effective communication, particularly in contexts with diverse gender identities and expressions.

One drawback of existing studies on gender bias detection is their limited ability to identify gender bias in contextual and emotionally charged language, such as social media interactions. Additionally, existing methods typically focus on identifying biased words or phrases, without considering the underlying causal connections between them. Table 1 shows some instances from our introduced *GAC* dataset that highlight manually annotated spans of gender-biased content in social media posts. The task of simultaneous detection of gender-biased content and identification of connected causal phrases from text using emotional information and gender-stereotyped external knowledge can help bridge this gap. By incorporating emotional information and leveraging external knowledge, this task can capture the contextual nuances of language and identify the underlying causal relationships between gender-biased content, allowing for more accurate detection and analysis of gender bias in language.

This study addresses this gap by introducing a novel problem of *Gender-biased Language Identification and Extraction (GLIdE)* from social media interactions and developing a multi-task, deep framework that detects gender-biased content and identifies connected causal phrases from the text using emotional information that is present in the input. The method proposed in this study includes a zero-shot strategy to integrate emotional information into the training process

| Sentence | Class |
|---|---|
| I'm not sexist as sexism is wrong and **I'm a man so I'm never wrong** | *GB* |
| Call me sexist but I think some **women are seriously lacking knowledge** | *GB* |
| This is because - no matter how much you #notallmen or argue a false equivalence with race or other characteristics - **males pose a risk to females**. It's not a coincidence that **most sex offenders are male and most of their victims are female**. Those figures are not unconnected. | *GB* |
| A **female can't tell me nothing about sports**. Sorry I grew up in a sports crazed house. | *GB* |
| Gender diversity is a key to good performance | *Non-GB* |
| I understood her point quite clearly. I still think her idea needs a little more development. | *Non-GB* |

Table 1: Sample instances from our *GAC* dataset. Span annotation(s) is highlighted in bold. GB: Gender-biased

and a unique mechanism to incorporate gender-stereotyped information as external knowledge. Additionally, the study introduces the *Gender-biased Analysis Corpus (GAC)*, a collection of 12,432 social media posts manually annotated with gender-biased stereotyped spans (as shown in Table 1). The *GAC* can provide valuable insights to researchers and policymakers, allowing them to gain a better understanding of the extent to which gender bias is present in social media posts.

The main contributions are summarised below:

- We propose the task of *Gender-biased Language Identification and Extraction (GLIdE)* from social media.
- Using emotional information and gender-stereotyped external knowledge, we propose a multi-task, deep framework to simultaneously detect gender-biased content and identify connected causal phrases from the text.
- Our method features 'structural embedding' to learn entity representations from triplets; 'textual embedding' using BERT and a vocabulary graph to capture lexical relationships within a language; emotion features extracted from BERT for automatic generation of emotion class semantic features.
- We propose a zero-shot loss objective to minimise the difference between the feature of the input text and the semantic feature of the emotion label, enabling model optimization.
- A corpus of 12,432 social media posts, annotated with gender-biased stereotyped spans, is created as the first *Gender-biased Analysis Corpus (GAC)*.
- Access the code and data at 1. https://www.iitp.ac.in/~ai-nlp-ml/resources.html#GLIdE-GAC, 2. https://github.com/Soumitra816/GLIdE-GAC.

**Task Relevance with respect to UN SDGs and LNOB.** Identifying and analysing gender-biased language accurately can promote inclusive societies and gender equality, in line with the UN Sustainable Development Goals (SDGs) and the Leave No One Behind principle (LNOB). Gender equality is a human right that fosters sustainable economic growth, reduces poverty, and builds peaceful and just societies. This work aligns with several UN SDGs, including Goal 5: Gender Equality, Goal 10: Reduced Inequalities, and Goal 16: Peace, Justice, and Strong Institutions.

1. **Goal 5 - Achieving Gender Equality:** *GLIdE* promotes gender equality by identifying language that reinforces gender stereotypes and discrimination against women and girls. It also ensures inclusivity for transgender and non-binary individuals.
2. **Goal 10 - Reducing Inequalities:** *GLIdE* reduces discrimination against marginalised groups, such as women, girls, LGBTQ+ individuals, and others, by identifying biased language.
3. **Goal 16 - Promoting Peace, Justice, and Strong Institutions:** *GLIdE* identifies language that reinforces prejudice and stereotypes, promoting peace and respect for diversity. It ensures marginalized groups are not left behind in building strong institutions.
4. **Leave No One Behind Principle:** *GLIdE* fosters inclusivity and diversity by detecting and analyzing language that reinforces prejudice and stereotypes, promoting the value and amplification of all individuals' voices.

The rest of the paper is organised as follows. Section 2 summarises some previous works in this area. We discuss the dataset preparation in Section 3. Section 4 addresses our proposed methodology in depth, followed by the results and analysis in Section 5. Finally, we conclude our discussion in Section 6 and define the scope of future work.

## 2 Related Work

It is well-documented that gender bias impacts women and marginalised groups' ability to fully participate in social, economic, and political spheres [Eagly and Karau, 2002; Brescoll and Uhlmann, 2008]. The emergence of social media platforms has provided a new space for communication and interaction, but also a new space for gender bias to manifest [Kramer *et al.*, 2014]. As such, there is an urgent need to develop effective methods for identifying and analysing gender-biased language in online interactions.

Various approaches, including rule-based, statistical, and machine-learning-based methods, have been proposed to combat sexism and online harassment. [Megarry, 2014] explored different types of online harassment, while [Waseem and Hovy, 2016] proposed a classification system for sexism using a dataset of 16,000 tweets (only 200 were available at the time of the study). [Jha and Mamidi, 2017] identified a new category of benevolent sexism based on tweets exhibiting "protective paternalism", "complementary gender differentiation", or "heterosexual intimacy". [Lewis *et al.*, 2017] documented the top ten categories of abuse experienced by female activists on social media, while [Vitis and Gilmour, 2017] analyzed the threatening nature of online harassment against women using the definition of [Citron, 2009].

Previous work has focused on detecting gender bias in language using automated methods. [Bolukbasi *et al.*, 2016] proposed a method to identify gender bias in word embeddings, while [Caliskan *et al.*, 2017] introduced a method to detect gender bias in text corpora. However, the interpretability of these models has received little attention, despite the social and legal consequences of erroneous predictions. Furthermore, these studies primarily focus on detecting gender bias at the sentence or document level and do not consider the causal relations between different parts of the text. Knowledge graphs have become popular in NLP due to their capacity to represent external knowledge, as demonstrated by [Han *et al.*, 2018]. Emotional information is also important in language and can improve various Natural Language Processing (NLP) tasks, including sentiment analysis [Kumar *et al.*, 2019], depression detection [Ghosh *et al.*, 2022a], etc.

Despite prior research on gender bias in language, improved methods are needed to identify and analyse gender-biased content, especially in social media interactions. The proposed *Gender-biased Language Identification and Extraction (GLIdE)* framework meets this need by using emotional information and gender stereotypes as external knowledge to improve the accuracy of identifying gender-biased language.

Hereon, we refer to the tasks of gender-biased language identification and extraction as *GLI* and *GLE*, respectively.

## 3 Dataset

We discuss the data collection and annotation details in the following subsections.

### 3.1 Data Collection

In this study, we create the *Gender-biased Analysis Corpus* (*GAC*) corpus by consolidating instances from the following three benchmark datasets.

- Workplace Sexism [Grosz and Conde-Cespedes, 2020]: Publicly available on Github[1], this dataset includes 1100+ examples of workplace sexism, covering certain and ambiguous cases, and features examples of sexism towards both genders. It differs from previous Twitter datasets by filtering out rare scenarios, removing duplicates, and using formal language instead of slang.
- Call Me Sexist: [Samory *et al.*, 2021] retrieved data from Twitter's Search API by using the phrase "call me sexist, but" and then annotated the retrieved sentences using crowd-sourcing. A pilot study showed that annotators tended to assume anything following the phrase was sexist if interpreted as a disclaimer.
- *EXIST@IberLEF*: [Rodríguez-Sánchez *et al.*, 2021] compiled prevalent sexist terms and phrases in English and Spanish by extracting them from Twitter messages that women encounter regularly. The terms and expressions were frequently employed to devalue and underestimate women's roles in society.

A number of sexist datasets are compared in Table 2. None of the existing datasets is marked with spans for gender bias, and *GAC* is the first-of its kind. Our *GAC* corpus includes all instances from "Workplace Sexism" and "EXIST 2021" datasets, and only the *Sexist* and *Non-sexist* samples from the "Call Me Sexist" dataset[2]. All instances under the 'sexist' category are categorized under the 'Gender-biased' (GB) class of the newly formed *GAC* dataset. Rest are placed under the 'Non-Gender-biased' (Non-GB) category.

| Datasets | Labels | Size | Spans |
|---|---|---|---|
| Waseem & Hovy [Waseem and Hovy, 2016] | Racist, Sexist, Normal | 16k | x |
| AMI@IberEval [Fersini *et al.*, 2018] | Misogynous, Not Misogynous, | 8k | x |
| Exist@IberLEF [Rodríguez-Sánchez *et al.*, 2021] | Sexist, Not Sexist, | 11k | x |
| Call me Sexist [Samory *et al.*, 2021] | Sexist, Not Sexist, Toxicity | 14k | x |
| ***GAC (Ours)*** | **Gender-biased, Non-Gender-biased** | **12k** | ✓ |

Table 2: Comparisons of different sexist datasets

[1]https://github.com/dylangrosz/Automatic_Detection_of_Sexist_Statements_Commonly_Used_at_the_Workplace

[2]We exclude instances of the 'Toxicity' class.

### 3.2 Data Annotation

Before starting the annotation task, annotators are informed that it may contain hate or offensive content. We provide works by [Poria *et al.*, 2021; Ghosh *et al.*, 2022b] to aid in classification and span annotation. For instances labelled as *Gender-biased*, we ask two PhD linguistics and one PhD computer science student to highlight text portions that contain terms that could justify the annotation. These span annotations help explore manifestations of gender bias. Following [Poria *et al.*, 2021; Ghosh *et al.*, 2022b], annotators mark multiple causal spans for a *gender-biased* post. We determine the final causal span using the span-level aggregation method in [Gui *et al.*, 2016] and assess inter-rater agreement using the macro-F1 measure, achieving an F1-score of 0.74, indicating high-quality annotation.

Table 3 contains the details of the distribution of instances over the *GB* and *Non-GB* classes for the constituent datasets that form the *GAC* corpus. We also show the number of causal spans annotated for each case. The average number of tokens highlighted per gender-biased post is 12.25 while the average number of tokens per gender-biased post is 24.59. We show some examples from our introduced *GAC* dataset in Table 1. The highlighted text(s) in each example is the manually annotated causal spans for gender-biased content. We generate a word cloud (as shown in Figure 1) of trigrams formed from the important combinations of words in the GB posts of the *GAC* dataset. Specifically, we find the noun phrases (important keywords combination) in the text that would help to find out what entities are being talked about in the given text. The bigger the phrase in the visual, the more often it appeared in the posts.

| Dataset | GB | Non-GB | Total | 1 cause | 2 causes |
|---|---|---|---|---|---|
| *Workplace* | 627 | 515 | 1142 | 626 | 12 |
| *Call me Sexist* | 1809 | 3837 | 5646 | 1806 | 10 |
| *EXIST* | 2794 | 2850 | 5644 | 2782 | 229 |
| **Total (GAC)** | **5230** | **7202** | **12432** | **5214** | **251** |

Table 3: Dataset details. GB: Gender-biased


Figure 1: Word Cloud from *GB* posts of the *GAC* dataset.

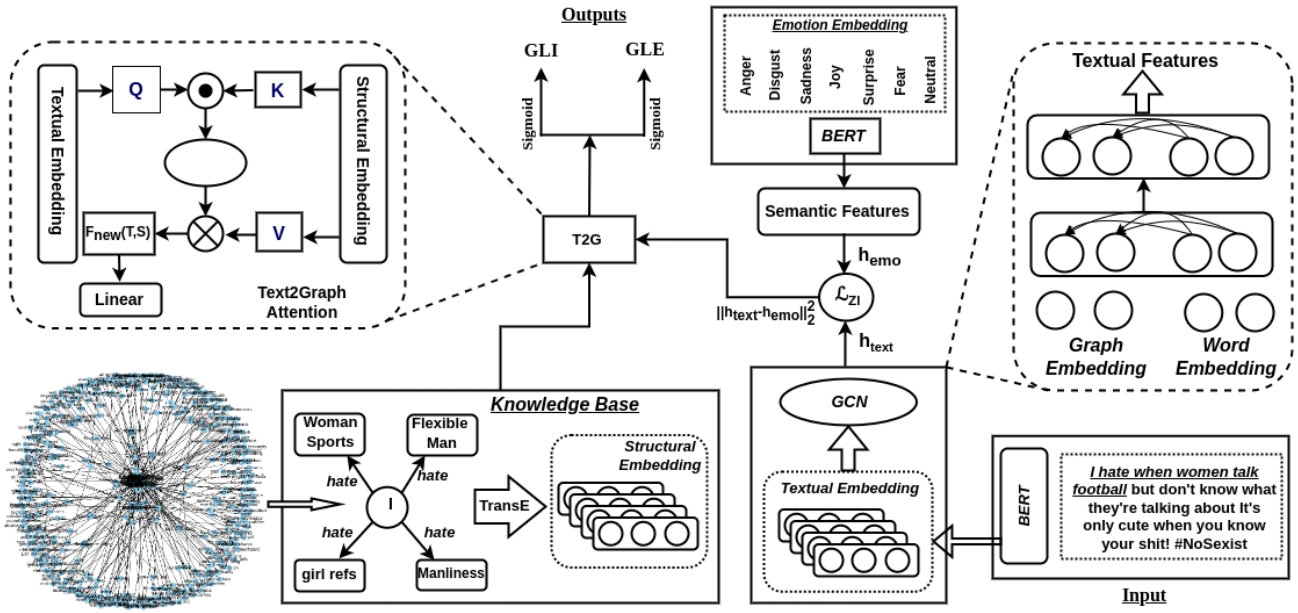


Figure 2: Illustration of the proposed *Gender-biased Language Identification and Extraction (GLIdE)* framework

## 4 Methodology

In this section, we present the *GLIdE* framework, a multi-tasking system for *Gender-biased Detection and Cause Extraction* from social media posts. Our approach utilizes a zero-shot strategy to incorporate emotional information during training and incorporates a novel fusion mechanism to combine features from multimodal inputs. Figure 2 illustrates the overall architecture of our method.

### 4.1 Problem Formulation

The task at hand involves analysing a social media post (SMP), represented by a sequence of sentences denoted as SMP = [$s_1, \cdots s_i \cdots, s_t$]. Each sentence $s_i$ can be further broken down into a sequence of words. The total number of sentences in the post is denoted by *t*. The main objective is to determine whether the social media post contains any gender bias or not. This determination is represented by a binary value, with 0 indicating the absence of gender bias and 1 indicating the presence of gender bias. Additionally, it is also required to identify and extract all possible causal spans within the post that supports the prediction of gender bias.

### 4.2 Structural Embedding

Various techniques can be employed to obtain structured entity embeddings for knowledge graphs. TransE [Bordes *et al.*, 2013] was chosen for learning entity representations from triplets due to its simplicity. The triplets are formed from the annotated gender-biased spans of the *GAC* dataset. TransE takes in triplets of the form *(h, r, t)*, where *h* is the head entity, *r* is the relationship between *h* and *t*, and *t* is the tail entity. TransE then learns an embedding $x_e \in \mathbb{R}^M$ for each entity *e* in the graph. The embedding is a vector representation of the entity in an M-dimensional space[3]. The way TransE works is by regarding the relationship *r* as a translation vector that connects the head entity *h* to the tail entity *t*. This means that the vector representation of *h* plus the vector representation of *r* should be close to the vector representation of *t*. Mathematically, this is expressed as $\boldsymbol{h + r = t}$.

[3]In this scenario, M=100

### 4.3 Textual Embedding

Our text encoder combines BERT [Devlin *et al.*, 2019] with a vocabulary graph that captures lexical relationships and word co-occurrences within documents. BERT extracts localized information, while the vocabulary graph represents the global vocabulary network. Together, they generate an embedding representation based on the input text. Attention mechanisms are applied to this embedding in combination with the input text. The vocabulary graph construction utilizes Normalized Point-wise Mutual Information (NPMI) [Bouma, 2009]. Figure 2 provides a visual illustration of the process.

$$NPMI(x, y) = \frac{1}{\log p(x, y)} \log \frac{p(x, y)}{p(x)p(y)} \qquad (1)$$

The variables $a$ and $b$ represent words, while $p(x, y) = \frac{W(x,y)}{W}$, $p(x) = \frac{W(x)}{W}$, and $p(y) = \frac{W(y)}{W}$ are the probabilities of the word pairs and individual words, respectively, based on their frequency in sliding windows. Here, $W(*)$ represents the frequency of a word or word pair, and W represents the total number of windows. The NPMI range is from -1 to +1, with +1 indicating a high semantic correlation between words and -1 indicating no correlation. We set an empirical threshold of 0.1 to 0.4 for NPMI.

To construct the lexical graph based on the vocabulary instead of documents, we use the Graph Convolution Network (GCN) [Kipf and Welling, 2017]. In Equation 2, we define the convolution layer in GCN for a single document repre-

sented as a row vector *r* comprising words from the lexicon.

$$hidd = (r^T E^{\sim})^T W = rE^{\sim}W \tag{2}$$

Where $E^{\sim}$ represents our vocab graph, $rE^{\sim}$ pulls the relevant component of our vocab graph from the input sentence r. The one-layer graph convolution is:

$$HIDD = RE^{\sim}W \tag{3}$$

Finally, we compute 4 as follows:

$$HIDD = ReLU(R_{bv}E_v^{\sim}W_{v*hidd}) \tag{4}$$

where *b* represents the batch size, *v* represents vocab size, *hidd* represents the hidden size, and $R_{bv}$ is a vector that contains features extracted from the word embeddings of BERT. This equation uses several iterations of graph convolution to combine input sentence words with their corresponding vocabulary graph terms, with the goal of identifying the relevant portions of the graph in relation to the input.

Instead of taking just the incoming phrase's word embeddings as input, the BERT transformer takes both those and the vocabulary graph embedding derived in equation 5. This method not only records the phrase's lexical order but also the context provided by the GCN. A self-attention encoder uses layer-by-layer interaction to completely merge the local and global embeddings. Thus, the final embedding corresponding to this may be written as:

$$F_E = ReLU(R_{bv}E_v^{\sim}W_{vhidd}) \tag{5}$$

where $e$ is the embedding size (dimension).

### 4.4 Emotion Features

We employ the pre-trained BERT (*base*) [Devlin *et al.*, 2019] model to encode the semantic feature information for the Ekman's [Ekman, 1992] basic emotion classes (*Anger*, *Disgust*, *Sad*, *Joy*, *Surprise*, *Fear*). Additionally, we consider the *Neutral* class to accommodate instances that do not fall in the scope of Ekman's categorization. Fetching the features from BERT obviates the need for further human annotation.

**Zero-Shot Loss**

The objective of our model is to minimize the difference between the feature of the text, represented by $\theta(h_{text})$, and the semantic feature of the emotion label, represented by $\phi(h_{emo})$, through optimization. We achieve this using the following function:

$$\mathcal{L}_{zl} = ||\theta(h_{text} - \phi(h_{emo})||_2^2$$

### 4.5 Text2Graph (T2G)

We introduce T2G Attention, a cross-modality attention module that captures local features specified in a knowledge graph (e.g., hate, dislike) to capture the nuanced interplay between the semantic features of the graph and the linguistic features of the text. By incorporating these local mappings, we expect our model to be more generalized for bias detection.

Our T2G Attention module takes token-wise embeddings of the text $f_t(T) \in \mathbb{R}^{N*a}$ and the knowledge base $f_t(S) \in \mathbb{R}^{M*a}$ as input and uses text tokens as queries to search the knowledge base for relevant terms. Specifically, we use $Q = f_p(T)W_q$ as textual queries, $K = f_t(S)W_k$ as knowledge text keys, and $V = f_t(S)W_v$ as knowledge text values, where $W_q$, $W_k$, and $W_v$ are learnable linear transformations of size $a * a$.

To calculate the cross-modal attention $Att(T, S) \in \mathbb{R}^{N*M}$, the T2G Attention module computes a dot product between each text token and knowledge pair, followed by a softmax.

$$Att(T, S) = softmax(\frac{QK^{tra}}{\sqrt{a}})V \tag{6}$$

We use the attention matrix to generate novel feature representations $f_{new}(T, S) \in \mathbb{R}^{N*a}$ for all sentence tokens. Specifically, $f_{new}(T, S) = Att(T, S)V$, which recomputes the sentence token embeddings by incorporating token-wise embeddings of relevant words from the knowledge base. To obtain the sentence token-level embedding, we apply global pooling on the token dimension N of the new sentence token embeddings $f_{new}(T, S)$, resulting in $f^{\sim}new(T, S) \in \mathbb{R}^{1*a}$. We do not use skip connections, unlike earlier cross-modal attention blocks, to achieve a linearly separable representation. Instead, we apply a simple linear layer to calculate the alignment vector (*AS*) for each sentence-knowledge base pair.

$$AS(T, S) = Linear(f^{\sim}new) \tag{7}$$

**Task-specific layers.** The output from the most recent T2G unit's output, which corresponds to the target utterance, is fed into two task-specific dense layers and the output layers for the *GLI* and *GLE* tasks. For the *GLE* assignment, a linear layer with sigmoid activation is used to calculate the span start and end logits, with a threshold of 0.4. This layer serves as the output layer, producing the probability of up to three causal spans, as indicated by the output of the probabilities of the three first and three last tokens.

**Calculation of Loss**

Throughout the training process of the model, we adopt a uniform loss function as shown in equation 8. For the *GLI* and *GLE*, we utilize categorical cross-entropy loss and binary cross-entropy loss, respectively.

$$L = \sum_{\omega} W_{\omega} L_{\omega} \tag{8}$$

Here, $\omega$ refers to the two tasks, *GLI* and *GLE*. We update the weights ($W_{\omega}$) using back-propagation for the specific loss of each task.

## 5 Experiments and Results

In this section, we discuss the experiments performed along with the results and analysis.

### 5.1 Baselines

In this work, we consider the following systems as baselines for the thorough assessment of our proposed *GLIdE* approach and the presented *GAC* dataset: BiRNN-Attn [Liu and Lane, 2016], CNN-GRU [Zhang *et al.*, 2018], BiRNN-HateXplain [Mathew *et al.*, 2021], BERT [Liu *et al.*, 2019], BERT-HateXplain [Mathew *et al.*, 2021], SpanBERT [Liu *et al.*, 2019] and Cascaded Multitask System with External Knowledge Infusion (CMSEKI) [Ghosh *et al.*, 2022a].

| *Models* | GLI Task | | GLE Task | | | | |
|---|---|---|---|---|---|---|---|
| | F1 (%) | ACC. (%) | FM | PM | HD | JF | ROS |
| BiRNN-Attn [Liu and Lane, 2016] | 65.27 | 66.39 | 24.51 | 28.33 | 0.48 | 0.65 | 0.70 |
| CNN-GRU [Zhang *et al.*, 2018] | 66.62 | 68.31 | 25.21 | 29.99 | 0.50 | 0.67 | 0.72 |
| BERT [Liu *et al.*, 2019] | 69.36 | 71.51 | 31.32 | 33.48 | 0.55 | 0.71 | 0.74 |
| SpanBERT [Liu *et al.*, 2019] | 70.69 | 72.23 | 33.98 | 35.31 | 0.58 | 0.73 | 0.76 |
| BiRNN-HateXplain [Mathew *et al.*, 2021] | 67.41 | 68.21 | 28.10 | 30.27 | 0.50 | 0.69 | 0.71 |
| BERT-HateXplain [Mathew *et al.*, 2021] | 71.29 | 73.41 | 31.51 | 35.41 | 0.59 | 0.74 | 0.76 |
| CMSEKI [Ghosh *et al.*, 2022a] | 73.81 | 75.66 | 34.29 | 36.55 | 0.63 | 0.77 | 0.78 |
| ***GLIdE (Proposed)*** | **77.86** | **78.29** | **37.21** | **38.89** | **0.65** | **0.80** | **0.81** |

Table 4: Results from the *GLIdE* model and the various baselines on the *GAC* dataset. Here, the bolded values indicate maximum scores. Here, GLI: Gender-biased Language Identification, GLE: Gender-biased Language Extraction.

## 5.2 Experimental Setup

Our proposed model is developed using PyTorch[4], a deep learning package based on Python. For our experiments, we import BERT from the huggingface transformers[5] package, using a 12-layer and 12-head self-attention between graph embedding and word embedding. All experiments are conducted on an NVIDIA GeForce RTX 2080 Ti GPU. Our empirical results show that a structural embedding size of 100 is optimal. For optimization, we use Adam [Kingma and Ba, 2015] with a learning rate of 0.05 and a dropout of 0.3. Stochastic gradient descent has a learning rate of 1e-4, weight decay of 1e-3, and momentum of 0.5. The activation function is set as ReLU with a slope of 0.2. We perform 5-fold cross-validation on the *GAC* dataset for training and testing, running experiments for 200 epochs. To account for non-determinism in Tensorflow GPU operations, we report averaged scores after 5 runs of the experiments.

## 5.3 Evaluation Metrics

To evaluate the *Gender-biased Language Identification* task, we use accuracy and macro F1-scores. For the *Gender-biased Language Extraction* task, following the approach in [Ghosh *et al.*, 2022b], we report the full match (FM), partial match (PM), Hamming Distance (HD), Jaccard Similarity (JS), and Ratcliffe-Obershelp Similarity (ROS) scores.

## 5.4 Results and Analysis

Table 4 shows the results of the proposed *GLIdE* framework on the introduced *GAC* dataset.

### Comparison With Existing Works

The results presented in Table 4 indicate that CMSEKI is the best-performing baseline, which is unsurprising given its ability to utilise external knowledge sources to comprehend input information. Nonetheless, our proposed *GLIdE* model surpasses CMSEKI on all metrics, achieving a notable improvement of 4.05% F1 for the *GLI* task and 3 ROS points for the *GLE* task. While BERT-HateXplain is the top-performing baseline that does not rely on external information, it falls short when compared to our *GLIdE* framework, demonstrating a 4.88% F1 and 5 ROS point deficit for the *GLI* and *GLE* tasks, respectively. The poor performance of BERT [Liu *et al.*, 2019], SpanBERT [Liu *et al.*, 2019], and BERT-HateXplain [Mathew *et al.*, 2021] highlights the challenge that even powerful language models face in comprehending span extraction for gender-biased content, a crucial task.

### Human Evaluation

In order to qualitatively evaluate the identified causes of the model, human review was conducted on 300 randomly selected posts from the test dataset. Three well-defined metrics were used for the assessment process [Singh *et al.*, 2022], and a score ranging from 0 to 5 was awarded based on these metrics. The most incorrect responses received a score of 0, while the best received a score of 5. The evaluator examined Fluency[6], Knowledge Consistency[7], and Informativeness[8]. In Table 5, the proposed framework performed well for all the manual evaluation measures compared to various baselines. The proposed approach resulted in a higher *Knowledge Consistency* score, ensuring that the extracted causal spans were

[4]https://pytorch.org/

[5]https://huggingface.co/docs/transformers/index

[6]*Fluency*: This determines whether or not the extracted span is fluent and natural. Natural and regular answers are assigned a score of 5, whereas inarticulate ones receive a 0.

[7]*Knowledge consistency*: This determines whether or not the produced answer has used the appropriate knowledge. If the model generates responses based on irrelevant information, it must get a score of 0, while the selection of pertinent knowledge must receive a score of 5.

[8]*Informativeness*: This metric is used to assess how informative the produced replies are. Here, a score of 0 means that the replies are uninformative, and a score of 5 means that they are.

| Models | KC | Inf | F |
|---|---|---|---|
| BERT-HateXplain | 2.39 | 2.40 | 2.97 |
| SpanBERT | 2.74 | 3.01 | 3.21 |
| CMSEKI | 2.91 | 3.18 | 3.41 |
| ***GLIdE (Proposed)*** | **3.20** | **3.81** | **3.61** |

Table 5: Results of human evaluation. Here, KC: Knowledge Consistency, Inf: Informativeness, F: Fluency

| Model | Text | Label |
|---|---|---|
| **1. Human Annotator** | **you're a woman and you don't equate to shit** bc you aren't a man and you can't do anything as good as a man ever will | GB |
| BERT-HateXplain | you're a **woman** and you don't equate to *shit* bc you aren't a man and you can't do anything as good as a man ever will | GB |
| SpanBERT | **you're a woman** and you don't equate to shit bc you aren't a man and you can't do anything as good as a man ever will | GB |
| CMSEKI | **you're a woman** and you don't equate to shit bc you aren't a man and you can't do anything as good as a man ever will | GB |
| **Proposed** | **you're a woman and you don't equate to shit** bc you aren't a man and you can't do anything as good as a man ever will | GB |
| **2. Human Annotator** | I'm not sexist, but **some bitches can't drive for SHIT.** | GB |
| BERT-HateXplain | I'm not **sexist**, but **some bitches** can't drive for SHIT. | GB |
| SpanBERT | **I'm not sexist**, but some bitches can't drive for SHIT. | Non-GB |
| CMSEKI | **I'm not sexist, but some bitches** can't drive for SHIT. | GB |
| **Proposed** | I'm not sexist, but some **bitches can't drive for SHIT.** | GB |

Table 6: Sample predictions from the various systems. Span annotation(s)/prediction(s) is highlighted in bold.

consistent with annotated causal spans. The *Informativeness* and *Fluency* of the proposed framework were also of high quality. Therefore, this demonstrates the model's strong ability to comprehend offensive information and produce results comparable to those of human annotators.

### Ablation Study

We perform ablation experiments on our *GLIdE* framework and report the results in Table 7. The experiments involved removing one module at a time from the *GLIdE* architecture. The three ablation experiments performed were: (1) removing any input from the knowledge graph (KG) to eliminate external knowledge from the model, (2) removing the textual embedding generation through Vocab graph and GCN and directly passing the BERT features as input to the T2G module after optimizing the $\mathcal{L}_{zl}$, and (3) replacing the T2G module with a simple linear concatenation operation to fuse the representations from KG and GCN. The results indicate a significant decrease in scores in all three experiments compared to the original *GLIdE* framework, with the KG module being the most crucial component for improving performance.

### Qualitative Analysis

We thoroughly examined the predictions made by the different systems. Consider the examples in Table 6. The top row displays the tokens (or 'causes') that human annotators noted and that they consider representing the causes for the post being *gender-biased*. The next four rows show the extracted tokens from the various models. We observe that the proposed *GLIdE* model correctly categorizes the examples as GB and also extracts good-quality causal spans. Existing span extraction system, such as SpanBERT highlights the explicit occurrences of gender-biased terms ('sexist'), whereas more advanced systems such as BERT-HateXplain and CMSEKI manages to highlight the stress markers ('some bitches'). However, all these systems fail to predict the gender-stereotyped fact ('bitches can't drive'), which our proposed *GLIdE* system predicted correctly.

| Setup | $F1^{GLI}$ (%) | $HD^{GLE}$ | $JF^{GLE}$ | $ROS^{GLE}$ |
|---|---|---|---|---|
| *[GLIdE]*$_{-KG}$ | 74.38 (-3.48) | 0.61 (-0.04) | 0.76 (-0.04) | 0.75 (-0.06) |
| *[GLIdE]*$_{-GCN}$ | 75.18 (-2.68) | 0.62 (-0.03) | 0.76 (-0.04) | 0.75 (-0.06) |
| *[GLIdE]*$_{-T2G}$ | 75.47 (-2.39) | 0.63 (-0.02) | 0.77 (-0.03) | 0.74 (-0.07) |
| **GLIdE** | **77.86** | **0.65** | **0.80** | **0.81** |

Table 7: Results of ablation experiments. KG: Knowledge Graph, GCN: Graph Convolutional Network, T2G: Text2Graph. The % fall in scores is shown in brackets.

## 6 Conclusion

In conclusion, this study proposes a novel task of *Gender-biased Language Identification and Extraction (GLIdE)* from social media interactions and presents a multi-task deep framework for detecting gender-biased content and identifying connected causal phrases from the text. The method incorporates emotional information and gender-stereotyped external knowledge and includes a unique mechanism to integrate this information into the training process. The study introduces the *Gender-biased Analysis Corpus (GAC)*, a collection of annotated social media posts that can aid researchers and policymakers in gaining a better understanding of gender bias in social media interactions. The proposed method and corpus could contribute to various domains, including natural language processing, social media analysis, and gender studies. The study provides open-sourced code and data to assist researchers in replicating the experiments and developing new methods to address gender bias in language.

The *GLIdE* method and *GAC* present a base for future research in identifying gender bias in social media. One possible extension of this work is to apply the *GLIdE* method to analyze gender bias in other languages and domains, such as online news, academic writing, and public speeches. This can involve adapting the proposed method to handle linguistic and cultural differences in different languages and domains. Additionally, expanding the *GAC* to include more diverse and representative data from different social media platforms and user demographics, as well as data from other domains, can provide a more comprehensive understanding of the prevalence and impact of gender bias in language.

## Ethical Statement

Our resource was developed using publicly available datasets, following the data use guidelines and ensuring no copyright infringement.

## Acknowledgements

Authors acknowledge the support from the project "HELIOS: Hate, Hyperpartisan, and Hyperpluralism Elicitation and Observer System", sponsored by Wipro Ltd., India.

## Contribution Statement

Gopendra Singh and Soumitra Ghosh contributed equally to this work and are joint first authors.

## References

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

[Bordes *et al.*, 2013] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26, 2013.

[Bouma, 2009] Gerlof Bouma. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, 30:31–40, 2009.

[Brescoll and Uhlmann, 2008] Victoria L Brescoll and Eric Luis Uhlmann. Can an angry woman get ahead? status conferral, gender, and expression of emotion in the workplace. *Psychological science*, 19(3):268–275, 2008.

[Caliskan *et al.*, 2017] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

[Citron, 2009] Danielle Keats Citron. Law's expressive value in combating cyber gender harassment. *Mich. L. Rev.*, 108:373, 2009.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, 2019.

[Eagly and Karau, 2002] Alice H Eagly and Steven J Karau. Role congruity theory of prejudice toward female leaders. *Psychological review*, 109(3):573, 2002.

[Ekman, 1992] Paul Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200, 1992.

[Fersini *et al.*, 2018] Elisabetta Fersini, Paolo Rosso, and Maria Anzovino. Overview of the task on automatic misogyny identification at ibereval 2018. *Ibereval@ sepln*, 2150:214–228, 2018.

[Ghosh *et al.*, 2022a] Soumitra Ghosh, Asif Ekbal, and Pushpak Bhattacharyya. A multitask framework to detect depression, sentiment and multi-label emotion from suicide notes. *Cogn. Comput.*, 14(1):110–129, 2022.

[Ghosh *et al.*, 2022b] Soumitra Ghosh, Swarup Roy, Asif Ekbal, and Pushpak Bhattacharyya. Cares: Cause recognition for emotion in suicide notes. In *European Conference on Information Retrieval*, pages 128–136. Springer, 2022.

[Grosz and Conde-Cespedes, 2020] Dylan Grosz and Patricia Conde-Cespedes. Automatic detection of sexist statements commonly used at the workplace. In *Trends and Applications in Knowledge Discovery and Data Mining: PAKDD 2020 Workshops, DSFN, GII, BDM, LDRC and LBD, Singapore, May 11–14, 2020, Revised Selected Papers 24*, pages 104–115. Springer, 2020.

[Gui *et al.*, 2016] Lin Gui, Dongyin Wu, Ruifeng Xu, Qin Lu, and Yu Zhou. Event-driven emotion cause extraction with corpus construction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1639–1649, Austin, Texas, November 2016. Association for Computational Linguistics.

[Han *et al.*, 2018] Xu Han, Shulin Cao, Xin Lv, Yankai Lin, Zhiyuan Liu, Maosong Sun, and Juanzi Li. OpenKE: An open toolkit for knowledge embedding. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, November 2018.

[Jha and Mamidi, 2017] Akshita Jha and Radhika Mamidi. When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data. In *Proceedings of the second workshop on NLP and computational social science*, pages 7–16, 2017.

[Kingma and Ba, 2015] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[Kipf and Welling, 2017] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

[Kramer *et al.*, 2014] Adam DI Kramer, Jamie E Guillory, and Jeffrey T Hancock. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111(24):8788–8790, 2014.

[Kumar *et al.*, 2019] Abhishek Kumar, Asif Ekbal, Daisuke Kawahra, and Sadao Kurohashi. Emotion helps sentiment: A multi-task model for sentiment and emotion analysis. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2019.

[Lewis *et al.*, 2017] Ruth Lewis, Michael Rowe, and Clare Wiper. Online abuse of feminists as an emerging form of violence against women and girls. *British journal of criminology*, 57(6):1462–1481, 2017.

[Liu and Lane, 2016] Bing Liu and Ian Lane. Attention-based recurrent neural network models for joint intent detection and slot filling. *arXiv preprint arXiv:1609.01454*, 2016.

[Liu *et al.*, 2019] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[Mathew *et al.*, 2021] Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14867–14875. AAAI Press, 2021.

[Megarry, 2014] Jessica Megarry. Online incivility or sexual harassment? conceptualising women's experiences in the digital age. In *Women's Studies International Forum*, volume 47, pages 46–55. Elsevier, 2014.

[Poria *et al.*, 2021] Soujanya Poria, Navonil Majumder, Devamanyu Hazarika, Deepanway Ghosal, Rishabh Bhardwaj, Samson Yu Bai Jian, Pengfei Hong, Romila Ghosh, Abhinaba Roy, Niyati Chhaya, et al. Recognizing emotion cause in conversations. *Cognitive Computation*, 13(5):1317–1332, 2021.

[Rodríguez-Sánchez *et al.*, 2021] Francisco Rodríguez-Sánchez, Jorge Carrillo-de Albornoz, Laura Plaza, Julio Gonzalo, Paolo Rosso, Miriam Comet, and Trinidad Donoso. Overview of exist 2021: sexism identification in social networks. *Procesamiento del Lenguaje Natural*, 67:195–207, 2021.

[Samory *et al.*, 2021] Mattia Samory, Indira Sen, Julian Kohne, Fabian Flöck, and Claudia Wagner. "call me sexist, but...": Revisiting sexism detection using psychological scales and adversarial samples. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 573–584, 2021.

[Singh *et al.*, 2022] Gopendra Vikram Singh, Mauajama Firdaus, Shruti Mishra, Asif Ekbal, et al. Knowing what to say: Towards knowledge grounded code-mixed response generation for open-domain conversations. *Knowledge-Based Systems*, 249:108900, 2022.

[Vitis and Gilmour, 2017] Laura Vitis and Fairleigh Gilmour. Dick pics on blast: A woman's resistance to online sexual harassment using humour, art and instagram. *Crime, media, culture*, 13(3):335–355, 2017.

[Waseem and Hovy, 2016] Zeerak Waseem and Dirk Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93, 2016.

[Zhang *et al.*, 2018] Ziqi Zhang, David Robinson, and Jonathan Tepper. Detecting hate speech on twitter using a convolution-gru based deep neural network. In *European semantic web conference*, pages 745–760. Springer, 2018.