Exploring Social Bias in Slovenia: The EEC-SL Dataset

Jaya Caporusso[1,2,*] **Damar Hoogland**[1,*] **Boshko Koloski**[1,2]
Matthew Purver[1,3] **Senja Pollak**[1] **Špela Vintar**[1,4]
[1] Jožef Stefan Institute, Ljubljana, Slovenia
[2] Jožef Stefan International Postgraduate School, Ljubljana, Slovenia
[3] Queen Mary University of London, London, UK
[4] Faculty of Arts, University of Ljubljana, Ljubljana, Slovenia
{jaya.caporusso, damar.hoogland, boshko.koloski, matthew.purver, senja.pollak, spela.vintar}@ijs.si

Abstract

We introduce the EEC-SL dataset, an adaptation of the Equity Evaluation Corpus from English to Slovenian. Based on 11 sentence templates, the dataset contains 8,640 sentences, including pairs of minimally-distant sentences, varying with regard to one of two variables: gender (*female* or *male*), and ethnicity (*Slovenian* or *not-Slovenian*). In order to validate our selection of personal names, we create a localised version of the Implicit Association Test for ethnic bias, in which participants show a significant implicit bias favouring Slovenian over non-Slovenian names. We use the dataset to evaluate social bias in three computational language models (large language models and an encoder-only transformer) to perform sentiment analysis—specifically, valence. We analyse the results in terms of differences in sentiment between minimally-distant groups of sentences and inferential tests. We found limited evidence for social bias with regard to ethnicity, and no evidence for gender bias, in any of the employed models.

Keywords: dataset, social bias, bias detection, large language models, encoder-only models, gender bias, ethnic bias

1. Introduction

Bias is often conceptualised through three interrelated constructs: stereotypes, explicit bias, and implicit bias (Greenwald and Banaji, 1995). Stereotypes are generalised beliefs or cognitive schemas about the characteristics, traits, or behaviours of members of a particular social group. Explicit bias refers to consciously endorsed attitudes or evaluations toward a group, which individuals can deliberately report and may intentionally act upon. Implicit bias (Greenwald and Krieger, 2006; Holroyd et al., 2017) refers to unconscious associations or preferences that can influence decisions and judgments without conscious awareness. Such biases often stem from repeated exposure to societal norms, media portrayals, or cultural narratives.

Computational language models perpetuate the biases present in the training corpora (Bolukbasi et al., 2016; Lauscher and Glavaš, 2019). This is true also for large language models (LLMs), where their fairness and the mitigation of bias have become important considerations. While a great deal of effort has been put into studying bias (Shah et al., 2020; Delobelle et al., 2022; Chu et al., 2024), the vast majority of these studies focus on English. Several recent studies point to the importance of better understanding the potentially biased behaviour of multilingual models in non-English languages (Xu et al., 2025; Liu et al., 2025), and the effects of the dominant language(s) of pre-training data on bias in local use scenarios are still under-researched (Löhr et al., 2025).

Many approaches can be adopted to detect bias, such as vector similarity in static word embeddings (Bolukbasi et al., 2016), Word Embedding Analogy Test (WEAT) (Du et al., 2019), masked token predictions in LLMs (Kirk et al., 2021), a comparison of the probabilities assigned to minimally different sentences reflecting stereotypes vs. counter-stereotypes (Nangia et al., 2020), and sentiment analysis (Rawat and Vadivu, 2022).

For Slovenian, gender bias was explored by Derner and Batistič (2025), who used LLMs to generate short stories. They then used another LLM to find and label all gendered person references, checked these labels against human annotations, and calculated the masculine-to-feminine ratio to measure gender representation bias. Ulčar et al. (2021) applied the Word Embedding Association Test (WEAT) to Slovenian and Croatian word embeddings to measure gender bias, comparing different embedding models and analysing analogies between male and female occupations to assess how gendered associations are represented in the models. Caporusso et al. (2024) analysed news articles in Slovenian about the Syrian and Ukrainian migration crises, extending Mendelsohn et al. (2020) with word embeddings and sentiment analysis to measure how migrants were associated with moral disgust and vermin metaphors over time.

In this work, we present the creation of the Slovenian Equity Evaluation Corpus (EEC-SL; Vintar, 2025), a new dataset for evaluating bias in Slovenian translated and adapted from EEC (Kiritchenko and Mohammad, 2018), and some initial experi-

* These authors contributed equally to this work.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 4019–4030
11–16 May 2026.

ments in which it is employed. The selection of personal names in the localised dataset followed official statistical data to represent the ethnic situation in Slovenia as accurately as possible, and thus include female and male Slovenian names from various frequency ranges, and non-Slovenian names associated with former Yugoslavia, the European Union, and other countries. Once the dataset was created, the personal names selected for the representation of ethnic (and potentially gender) bias were evaluated through an Implicit Association Test by 60 human participants, and secondly, the new dataset was used to evaluate three computational language models.

We introduce related work in Section 2. In Section 3, we delineate the process of creating the EEC-SL dataset, and in Section 4 we describe the Implicit Association Test study we conducted. In Section 5, we present the methodology employed to assess bias through sentiment analysis using EEC-SL. Sections 6 and 7 go more into detail about such evaluation: while in the former we compare the role of gender and ethnicity variables in a coarse-grained sentiment analysis setting, using positive, negative, neutral class, in the latter we use continuous sentiment scores. Section 8 addresses future work, ethical considerations, and limitations.

The code is publicly available at: https://github.com/jayacaporusso/SocialBias-Slovenia-EEC-SL and the dataset EEC-SL 1.0 is publicly available on the CLARIN repository[1] (Vintar, 2025).

2. Related work

Bias in language models has become a central topic both in the design, alignment and benchmarking of LLMs (Gallegos et al., 2024). Apart from the notorious gender stereotypes related to vocation (Kotek et al., 2023), LLMs have been found to perpetuate ethnic and gender bias in personal names (Sakunkoo and Sakunkoo, 2025), political stance (Löhr et al., 2025), and even ageism, beauty and other subtle biases (Kamruzzaman et al., 2024). Substantially fewer studies were dedicated to the exploration of bias in non-English use cases, and to comparisons between monolingual, multilingual and multilingual-but-locally-fine-tuned models. An important finding by (Nie et al., 2024) shows that, in a controlled setting and equal-sized models, multilingual models show significantly less bias than monolingual ones. It has also been shown that biases learned on English data are transferred to other languages (Wendler et al., 2024).

2.1. The Equity Evaluation Corpus

If a model is given two sentences that are the same besides the variable of interest (e.g., gender or nationality) and it assigns them a different sentiment, then bias is present. Along this line, Kiritchenko and Mohammad (2018) created the Equity Evaluation Corpus (EEC), a benchmark resource designed to measure gender and race bias in sentiment and emotion analysis systems. The corpus consists of 8,640 English sentences automatically generated from 11 templates combining gender- and race-associated names or noun phrases with emotion words expressing four basic emotions (anger, fear, joy, and sadness). The authors evaluated 219 systems from the SemEval-2018 Affect in Tweets shared task. In their analysis, the authors found that between 75% and 86% of systems produced significantly different scores for one or more demographic contrasts, and that bias direction varied by emotion: sentences referring to female subjects were generally assigned higher intensity for joy and anger, while male subjects were associated with higher fear intensity. For racial bias, only 8–24% of systems showed no significant difference.

Building on this work, Goldfarb-Tarrant et al. (2023) extended the EEC framework beyond English by developing counterfactual evaluation corpora for four additional languages: Japanese, Simplified Chinese, German, and Spanish. Their study adapted the original English templates of Kiritchenko and Mohammad (2018) to account for cross-linguistic grammatical variation (e.g., gender agreement in German and Spanish, passive/active forms in Japanese), creating contrastive sentence pairs that differed only in a demographic variable such as gender or race. Using these resources, they compared baseline bag-of-words classifiers to pre-trained transformer models fine-tuned for sentiment polarity detection. The results showed that biases were present across all four languages, with non-English systems generally displaying stronger bias than English ones. Importantly, pre-training reduced the magnitude of bias overall but also changed its nature: non-pre-trained models exhibited a few extreme label flips under counterfactual swaps, whereas pre-trained models made many smaller but systematic shifts toward more negative sentiment for minoritised groups. This study demonstrates the transferability of EEC-style counterfactual evaluation to morphologically richer and typologically diverse languages and highlights that multilingual pre-training, while mitigating some extremes, does not eliminate bias altogether.

[1] https://www.clarin.si/repository/xmlui/handle/11356/2049

2.2. Sentiment analysis for bias detection with encoder-only transformers and large language models

Several studies have applied BERT-style encoder models to quantify social bias in sentiment and emotion prediction tasks. Jentzsch and Turan (2022) proposed a bias metric for IMDB sentiment classification by generating pairs of minimally-distant sentences (i.e., identical except for male or female names and pronouns) and comparing the resulting polarity scores across multiple BERT configurations. Although their approach followed the logic of the Equity Evaluation Corpus (EEC), it did not employ the EEC directly; instead, they created a gender-swapped evaluation set to analyse whether bias arises in contextual representations or in the classifier head. In contrast, Bhardwaj et al. (2021) explicitly used the EEC to test BERT embeddings in emotion and sentiment intensity prediction, training regression models for valence, anger, fear, joy, and sadness. Their results showed systematic sentiment shifts between male- and female-associated sentences, demonstrating that even pre-trained contextual encoders can encode and propagate social bias across affective dimensions.

Also LLMs have been used as sentiment classifiers for bias detection, usually using their zero-shot predictions. Recent examples are the work by Elbouanani et al. (2025), in which the authors found systematic positive/negative bias toward left or far-right politicians, stronger bias in Western languages, and that larger models show more consistent bias.

2.3. The Implicit Association Test

In social science, bias is often detected using the Implicit-Association Test (IAT). The IAT is a psychological tool designed to measure the strength of automatic associations in our minds, such as between concepts (e.g., male/female) and evaluations or stereotypes (e.g., science/arts, good/bad). Instead of relying on self-reports, it tracks how quickly and accurately people sort words or images into paired categories. Faster responses when certain categories are combined suggest a stronger unconscious association, while slower responses indicate weaker or conflicting associations. “The main idea is that making a response is easier when closely related items share the same response key” (Implicit, 2011), and the results of the test indicate the participant’s level of implicit stereotypes towards the investigated social group or concept. An example of IAT’s applications is the study of bias in the healthcare domain (Maina et al., 2018). The IAT score is essentially a D-score that reflects the strength and direction of the implicit association. Values around 0 indicate no implicit bias, while a positive score indicates the strength of the pairing between two compatible categories (usually the culturally stereotypical pairing, e.g. male=good, female=bad). A negative score indicates an implicit association between opposite categories.

In this paper, we present an adaptation of the EEC corpus to Slovenian, considering not only the language, but also the cultural context. In order to see whether ethnic bias is present among Slovenians, we create a localised version of IAT for ethnic bias. We test the usability of the new dataset by performing sentiment analysis with three different computational language models.

3. Creating the EEC-SL dataset

Our dataset[2] is a localised and adapted version of the EEC (Kiritchenko and Mohammad, 2018). Each of the 11 templates of the original corpus contains a reference to <person>, where the slot can be filled either by a name (female and male, African American or European American), or by a generic noun phrase (e.g., *she, my sister, this man, my dad*). The second and third variables present in 7 out of 11 templates are <emotional state word>and <emotional situation word>, which can be filled by words expressing four basic emotional states: Anger, Fear, Joy and Sadness. For each of the emotions, the authors selected five words that convey the emotion in varying intensities.

The original EEC was designed to test for race and gender bias in an American context, hence the authors selected female and male first names associated with being African American or European American, in line with the findings of (Caliskan et al., 2017). In contrast, the noun phrases are racially neutral and convey only the gender of the person.

Our aim was to create a Slovenian version of the dataset that would adhere to the methodological considerations of the original, but would at the same time be culturally and linguistically adapted to the current social reality in Slovenia. This entailed a number of decisions explained in more detail below.

3.1. Names

Slovenia’s multicultural reality is shaped by various historical, geographic, social and political factors, including its 45-year-long existence within the former Socialist Federal Republic of Yugoslavia, its relative prosperity compared to other Balkan regions, its location on the so-called Western Balkans migration path, and other more recent events. Since

[2]Since *corpus* usually refers to a collection of texts produced in authentic communicative settings, we believe *dataset* is more accurate for this resource.

race is not in the forefront of social bias in Slovenia, and since Slovenia is ethnically relatively homogeneous, we reframe the ethnicity-based bias as the juxtaposition between *Slovenian* and *non-Slovenian*, hence focusing on the perception of *foreignness* (however, we refer to this variable in the EEC-SL dataset as ethnicity).

The task is thus to find non-Slovenian names (10 female and 10 male) which roughly correspond to the current demographic situation in the country, and at the same time are perceived as non-Slovenian. According to the Statistical Office of the Republic of Slovenia[3], a large majority of foreigners come from former Yugoslavia (77%), followed by other countries (13%) and the EU (10%). To avoid having just a single name representing a group of citizens, we slightly shift the above ratios and selected 6 names for former Yugoslavia, and 2 for EU and other countries respectively, for each gender. Both Slovenian and non-Slovenian names are selected from the registry of names published by the above mentioned Slovenian Statistical Office, whereby we attempt to match the frequencies of the selected names with those in the original dataset. The selected names are presented in Table 1.

Female		Male	
Slo	Non-Slo	Slo	Non-Slo
Anja	Snježana	Anže	Hrvoje
Brigita	Zorana	Aljaž	Mirsad
Klara	Emina	Andrej	Stjepan
Ema	Sanela	Filip	Amir
Hana	Esma	Mihael	Ajdin
Katja	Jelena	Jan	Nenad
Klavdija	Jacqueline	Jurij	Pierre
Maja	Ingrid	Jakob	Gavin
Neža	Suki	Roman	Mohamed
Sara	Aisha	Rok	Kenji

Table 1: Slovenian and non-Slovenian names.

3.2. Sentence templates

The original 11 sentence templates for English contain three possible placeholders (*person*, *emotional state* and *emotional situation*) which are used in simple sentences such as:

The conversation with <person> was <emotional_situation_word>.

Since Slovenian is a morphologically rich language with six grammatical cases (1–6), three genders (female, male, neutral), and three numbers (singular, dual, plural), each of the placeholders needs to be adapted in each template, in order to appear in the correct case and number. In addition, since <person>may be male or female, and since Slovenian is a language with gender agreement for finite verbs, two variants are created from each template: one for female and one for male. Thus, the example above would translate to:

Pogovor <s_z> <person_F_5> je bil <emotional_situation_word_M_S_1>.

Pogovor <s_z> <person_M_5> je bil <emotional_situation_word_M_S_1>.

In addition, the English preposition *with* may be translated into Slovenian as either *s* or *z*, depending on whether the following phoneme is voiced or not. For all of the above reasons, the original 11 English templates are translated into 22 Slovenian variants, with each of them containing the grammatically appropriate form of the placeholder.

3.3. Emotional words

The translation of the emotional state and situation words into Slovenian followed the principle of matching the emotional intensity of the English words, so that the adjectives describing a particular emotion would vary in intensity and at the same time convey the intended meaning. To obtain emotion scores for English, we use the NRC Word-Emotion Association Lexicon (Mohammad and Turney, 2013), and for Slovenian the Slovenian Emotion Dimension and Emotion Association Lexicon (Brglez et al., 2024). An additional challenge is the fact that the expressions need to fit into the phraseological context of each of the templates, so each collocation is checked against a reference corpus of Slovenian. Below we list the selected emotional state words for each of the emotions.

- **Anger** jezen [angry], razdražen [annoyed], besen [enraged], razjarjen [furious], nataknjen [irritated]
- **Fear** zaskrbljen [anxious], pobit [discouraged], preplašen [fearful], prestrašen [scared], zgrožen [terrified]
- **Joy** vzhičen [ecstatic], vznemirjen [excited], zadovoljen [glad], srečen [happy], olajšan [relieved]
- **Sadness** potrt [depressed], obupan [devastated], razočaran [disappointed], nesrečen [miserable], žalosten [sad]

Once all of the elements are localised, the dataset is generated using the adapted sentence templates and in the same format as the original EEC.

[3]https://pxweb.stat.si/SiStatData/pxweb/en/Data/-/05E1014S.PX/

4. Implicit Association Test

To see whether ethnicity-based bias is present amongst Slovenian participants, we create a localised version of the well-known Implicit Association Test (IAT) for racial bias, as provided by Project Implicit.[4]

The experiment is implemented using the minnoJS library and hosted via GitHub, with real-time data collection managed through the Open Science Framework.[5] The experiment is designed as follows. After the introductory page, the participant is guided through some basic demographic questions including age, gender identity, nationality, country of residence, and education. Then, the instructions for the test are provided (in Slovenian):

You will use the keys 'E' and 'I' on your keyboard to categorize items you see on the screen. Below are the groups of items: Positive word, Negative word, Slovenian, non-Slovenian. There are seven subtasks with slightly different instructions, so pay attention.

Each subtask then starts with the instruction screen for the task (e.g., press 'E' for 'Slovenian' and 'I' for 'Non-Slovenian'), and 20 actual categorisation trials. The relevant tasks to measure implicit bias are the ones where participants are required to categorise Slovenian vs. non-Slovenian names with Positive vs. Negative words.

After the participants complete all seven subtasks, they are presented with four follow-up questions examining their conscious or explicit attitudes towards Slovenians vs. non-Slovenians. These questions are:

1. How warm or cold do you feel towards Slovenians? [0 - Extremely cold, 10 - Extremely warm]
2. How warm or cold do you feel towards non-Slovenians? [0 - Extremely cold, 10 - Extremely warm]
3. I try to appear nonprejudiced towards non-Slovenians. [-3 Strongly disagree, 3 - Strongly agree]
4. Which statement best describes you? [I strongly/moderately/slightly prefer Slovenians to non-Slovenians/non-Slovenians to Slovenians OR I like Slovenians and non-Slovenians equally.]

The IAT experiment is shared via social media with the primary aim of serving as a pilot test of the usability of the newly constructed dataset for the exploration of ethnic bias in Slovenia, so that in this initial stage a total of 51 complete responses have been collected. The gender of the participants is 33 female, 17 male and 1 non-binary, and their ages range from 16 to 61, with average age 31.6.

The mean IAT score was M = 0.69 with SD = 0.31. As the two-sided 95% confidence interval ranges from 0.61 to 0.78 (i.e., it does not include 0, the neutral score), we conclude that the mean score was significantly more positive than 0, indicating an implicit bias favouring Slovenian over non-Slovenian names. The difference in IAT scores between male participants (N = 17, M = 0.78, SD = 0.31) and female participants (N = 33, M = 0.64, SD = 0.33) was 0.14. This difference is not significant according to a Welch's independent-samples t-test, $t(34.30) = 1.48$, $p = .15$. Comparing the participants' explicit attitudes towards Slovenians vs. foreigners as expressed in the initial questionnaire with temperature scores, and their IAT scores, a weak negative, non-significant Pearson correlation is found ($r(49) = -0.14$, $p > .05$). Similarly, a weak negative, non-significant Pearson correlation is found between education level and IAT score ($r(49) = -0.11$, $p > .05$).

5. Methodology

We use the EEC-SL dataset to test the presence of social bias in computational models. Inspired by Kiritchenko and Mohammad (2018), we assess bias through sentiment analysis. In Section 6 we compare the role of gender and ethnicity variables in a coarse-grained sentiment analysis setting, using *positive, negative, neutral class*, while in Section 7, we use continuous sentiment scores, as in Kiritchenko and Mohammad (2018).

We employ two different kinds of language models focusing on Slovenian: large language models (LLMs; in particular, GaMS-27B-Instruct and GaMS-9B-Instruct LLMs) (Vreš et al., 2024), and an encoder-only transformer SloBERTa (Ulčar and Robnik-Šikonja, 2021), fine-tuned for sentiment-analysis[6]. For the LLMs we assign sentiment in a zero-shot, prompt-based classification setting, without fine-tuning; for the SloBERTa sentiment classifiers, the model's sentiment is also assigned without additional training. We first employ this three-point scale due to it being more in line with the specific way in which SloBERTa is trained for sentiment and to LLMs having difficulties with providing fine-grained ratings (see Section 6). To further investigate whether the results obtained are dependent on the coarse-grained sentiment analysis performed, we decide to prompt the models to provide continuous sentiment scores on a scale from -1 (negative) to 1 (positive; see Section 7 for an explanation of how continuous scores were obtained). We evaluate the bias via inferential tests

[4] https://www.projectimplicit.net

[5] https://osf.io

[6] cjvt/sloberta-sentinews-sentence

and for continuous predictions also via Δ scores (Kiritchenko and Mohammad, 2018).

6. Testing bias via coarse-grained sentiment analysis

6.1. Sentiment prediction

We employ instruction prompting to guide the model's behaviour, explicitly instructing it to act as a sentiment classifier and to classify each sentence as either *positive*, *neutral*, or *negative*. The prompt utilised is: "*You are a sentiment classifier. Classify this sentence as either positive, neutral, or negative: 'sentence'*".

The temperature parameter is fixed to 0.0 to ensure deterministic responses and minimise sampling variability.

The encoder-only transformer model `Sloberta-sentinews-sentence` operates in a zero-shot classification setting, also without fine-tuning. We use pre-trained sentiment model through Hugging Face's `transformers` library, applying its built-in sentiment-analysis pipeline to classify each sentence as *positive*, *neutral*, or *negative*. For each input, the model outputs label probabilities, and the label with the highest confidence score is selected as the final prediction.

For both types of models, predictions are generated in parallel batches of up to 32 sentences to optimise runtime efficiency.

Table 2 lists the percentage of ratings per model.

6.2. Analysis

6.2.1. Inferential tests of language model bias

We assess the degree of bias in each of the three models for Slovenian—SloBERTa, GaMS 9B and GaMS 27B. For each model separately, we test whether the probability with which it rates a sentence as negative, neutral or positive is affected by gender, ethnicity, and their interaction. We fit ordinal Bayesian regression models, $Sentiment \sim Gender * Ethnicity$, with standard minimally informative priors (Normal (0,2) for coefficients and Student's-T (3,0,2.5) for intercepts), using the `brms` package (Bürkner, 2017) in `R` (R Core Team, 2022). Gender is sum-coded. Because some predictors did not meet the proportional odds assumption (as indicated for a maximum-likelihood model fit to the same data using the package `ordinal`; (Christensen, 2019)) we include category-specific effects for the relevant predictors. We draw inferences based on the predictor's 95% Credible Interval (CI): when a predictor's 95% CI does not include zero, we interpret this as evidence that the predictor's effect is not zero, i.e., that the sentiment model is biased. We only discuss predictors for which this is the case and summarise the magnitude of other estimates. The full regression model estimates, as well as fit diagnostics and tests of the proportional odds assumption, are printed in the GitHub repository.

We find no evidence that SloBERTa is biased for gender or ethnicity. The coefficients of the ordinal regression model predicting SloBERTa's sentiment ratings for gender and ethnicity and their interaction are all small (-0.04 to 0.06 in log-odds, meaning that the largest difference in odds between any two categories was 5.8%). The CI of all predictors include zero.

We also find no evidence that GaMS 9B is biased for gender or ethnicity. The coefficients of the ordinal regression model predicting GaMS 9B's sentiment ratings from gender and ethnicity and their interaction were all small (-0.08 to 0.09 in log-odds, meaning that the largest difference in odds between any two categories was 7.1%). The CI of all predictors included zero.

In the regression analysis we find some evidence that GaMS 27B has a slight bias towards rating sentences with foreign names as more positive than sentences with no ethnicity implied, although this effect only holds robustly for the threshold between negative to neutral ratings. Specifically, the odds of a sentence being rated as neutral compared to negative when it mentions a foreign-sounding name is 1.17 times higher than when no ethnicity is implied. The 95% CI is 0.05 to 0.27, which is wide but does not include zero. For comparison, the odds of a sentence being rated as neutral compared to negative if it mentions a Slovenian name is 1.03 times higher than if it implies no ethnicity. All other estimates are small (-0.05 to 0.06 in log-odds).

To conclude, by prompting the models trained on Slovenian—SloBERTa, GaMS 9B, and GaMS 27B—to rate the sentiment of sentences from this dataset on a 3-point scale, we find little evidence for either gender-based or ethnicity-based bias. The only exception is that GaMS 27B shows a slight inclination to rate sentences with foreign names as neutral rather than negative.

6.2.2. Qualitative analysis

A qualitative analysis of the predictions is performed to examine potential differences between individual names, that is, whether any of the names would evoke significantly more non-neutral predictions than other names, for any of the models. Firstly, we examine the sentences containing a name but no emotion word (N=80). Out of the three models, only SloBERTa does not assign exclusively neutral sentiment to such sentences. This would imply that for GaMS models, the sentiment of the sentence cannot be altered solely by the mention of a name, regardless of ethnicity. SloBERTa predicts

Gender	Ethnicity	Rating	SloBERTa	GaMS 27B	GaMS 9B
Female	/	Negative	76.2	67.2	74.9
Female	/	Neutral	4.3	7.8	1.8
Female	/	Positive	19.4	25.1	23.3
Female	Slovenian	Negative	75.8	67.0	75.0
Female	Slovenian	Neutral	4.0	9.0	3.4
Female	Slovenian	Positive	20.1	24.0	21.6
Female	Non-Slovenian	Negative	76.0	64.2	73.7
Female	Non-Slovenian	Neutral	4.0	10.4	4.7
Female	Non-Slovenian	Positive	20.0	25.4	21.6
Male	/	Negative	75.9	66.6	75.0
Male	/	Neutral	4.9	9.2	2.4
Male	/	Positive	19.2	24.2	22.6
Male	Slovenian	Negative	75.3	65.5	74.8
Male	Slovenian	Neutral	5.4	10.5	3.7
Male	Slovenian	Positive	19.3	24.0	21.5
Male	Non-Slovenian	Negative	75.8	62.2	72.8
Male	Non-Slovenian	Neutral	5.1	14.2	5.5
Male	Non-Slovenian	Positive	19.1	23.6	21.7

Table 2: Percentage of ratings (negative, neutral, positive) per model (coarse-grained sentiment analysis).

a positive sentiment for 14 sentences featuring foreign and 15 featuring Slovenian names, with no negative predictions.

The comparison of sentiment scores assigned to individual names by different models reveals no relevant differences between names, as the ratio between positive and negative predictions for an individual name generally follows the ratio between positive and negative emotion words in the dataset (approximately 1:2.5), with a slight tendency of GaMS 27B to rate fewer non-Slovenian names as negative than Slovenian.

7. Testing bias via fine-grained sentiment analysis

7.1. Sentiment prediction

Because a three-point scale may not be sensitive enough to detect small effects of gender and ethnicity, we subsequently also prompt the LLMs to provide ratings on a continuous scale from -1 (negative) to +1 (positive). As SloBERTa was trained to provide three-class sentiment classifications, we obtain continuous sentiment scores from the underlying probabilities (softmaxed logits) assigned by the model to each of the three classes by subtracting the probability of being negative from the probability of being positive.

7.2. Analysis

7.2.1. Inferential tests of language model bias

We test whether the models reveal any bias when rating the sentences on this scale by fitting Bayesian beta regression models, *Sentiment rating* $\sim$ *Gender * Ethnicity*, with standard minimally informative priors (Normal(0,1) for coefficients, Student's-t(3,0,2.5) for intercepts, and Exponential(1) for ϕ), using the `brms` package (Bürkner, 2017) in R (R Core Team, 2022). Gender is sum-coded. To fit a Beta regression model to the sentiment rating scale, we re-scale the sentiment rating from -1–1 to 0–1. Because the Beta distribution is not defined for the edge values 0 and 1, we then clip values 0 and 1 by a small amount by replacing any value = 0 with 1×10^{-4} and any value = 1 with $1 - 1 \times 10^{-4}$. We draw inferences based on the predictor's 95% Credible Interval (CI): when a predictor's 95% CI does not include zero, we interpret this as evidence that the predictor's effect is not zero, i.e., that the sentiment model is biased. We only discuss predictors for which this is the case and summarise the magnitude of other estimates.

We find no evidence that SloBERTa is biased for gender or ethnicity. The coefficients of the regression model predicting SloBERTa's sentiment ratings from gender and ethnicity and their interaction are all small (β = -0.03 to 0.07). The CI of all predictors included zero.

We find some evidence that GaMS 9B is biased for ethnicity. Specifically, when a Slovenian name is mentioned, the model rates the sentence as more negative than if no name is mentioned (β = -0.09, CI = -0.16 to -0.02). When a non-Slovenian name is mentioned, the model rates the sentence as more negative than if no name is mentioned (β = -0.19, CI = -0.26 to -0.12). Sentences with Slovenian names are rated as slightly more positive than those containing non-Slovenian names (β = 0.1, CI = 0.03

to 0.17; this corresponds to around 0.05 points or 2.53% on the scale from -1 to 1, CI = 0.05 to 0.05).

We find no evidence that GaMS 27B is biased for gender or ethnicity. The coefficients of the regression model predicting GaMS 27B's sentiment ratings from gender and ethnicity and their interaction are all small (β = 0 to 0.01). The CI of all predictors includes zero.

7.2.2. Computation of bias scores

Having continuous scores allows us to perform an additional bivariate analysis using aggregate data following a procedure similar to that of Kiritchenko and Mohammad (2018). For each model, we group the sentences by sentence template and emotion word. There are 144 sentence template $\times$ emotion word groups, but note that not each group contains the same number of sentences. To assess the degree of gender bias, we then calculate the mean sentiment score of female vs male sentences in each group. To assess the degree of ethnicity bias, we calculate the mean sentiment scores of sentences mentioning Slovenian vs non-Slovenian names vs no names in each group. Figure 1 illustrates the difference between mean sentiment score per group.

We additionally calculate 'bias scores' as the difference Δ between female-male or Slovenian-non-Slovenian mean sentiment score per group. Table 3 lists the mean Δ and SD Δ for each model and both comparisons (Gender and Ethnicity) to summarise the direction and magnitude of potential bias. The mean bias scores are small, ranging from Δ = -0.0035 to 0.0148.

We also perform paired t-tests to assess the difference in mean sentiment score between male-female and Slovenian-Non-Slovenian sentence groups. Since the Slovenian-Non-Slovenian comparison is done over a subset of the same data as the female-male comparison, we adopt a = 0.025 significance threshold (Bonferroni-corrected from the conventional 0.05). We do not perform corrections to account for the analyses of the same data in Section 7, so the test statistics should be interpreted with caution. The results can be found in Table 3. The only statistically significant difference (in bold in Table 3) is that in SloBERTa, sentences with Slovenian names seem to have a positive bias compared to those with non-Slovenian names.

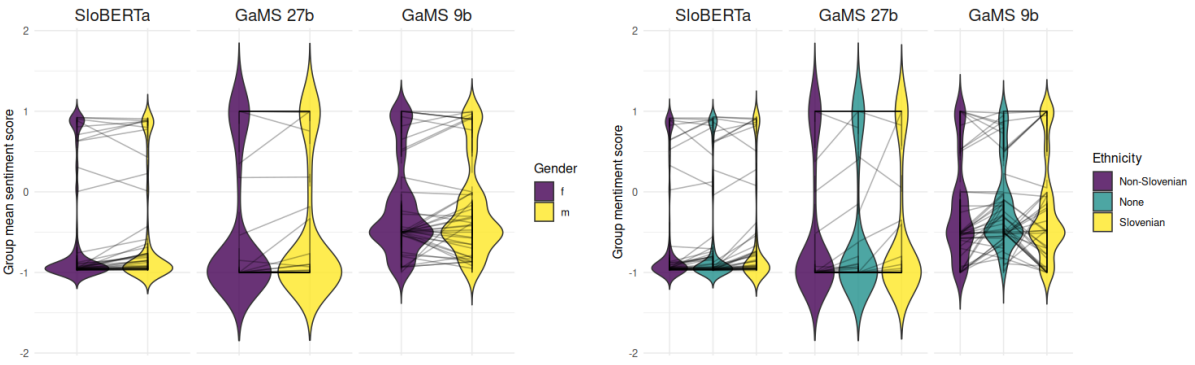
8. Conclusion and future work

We present a new dataset for evaluating ethnic and gender bias for Slovenian, which was created on the basis of the EEC corpus (Kiritchenko and Mohammad, 2018) with all the necessary linguistic and cultural adaptations. The linguistic aspect involved a modification of all sentence templates to accommodate the syntactic and morphological structure of Slovenian, a careful selection of adjectives to span the entire range of emotional valences and to fit into the collocational contexts, and the cultural adaptation involved selecting native versus foreign names to represent the ethnic situation in Slovenia. We further conducted a IAT on a small sample of participants to validate the selection of names, and show that implicit bias against non-Slovenian names is present, while correlations between the IAT score and gender, education or explicit bias were not statistically significant.

We then present experimental results testing two language models and one encoder-only transformer, for sentiment analysis. Using two different rating scales (3-point categorical and -1 to 1 continuous), we found inconsistent evidence regarding the bias of the three models. Using a 3-point scale and a Bayesian regression analysis, GaMS 27B showed a slight bias in favour of sentences with a non-Slovenian name compared to sentences with no ethnicity implied. Using a continuous scale, we found some evidence that GaMS 9B has a slight bias against both Slovenian and non-Slovenian names, but more so against non-Slovenian names. In a second analysis using aggregate data (bias scores), we found that in SloBERTa, Slovenian names are slightly more positively-biased than non-Slovenian names. This non-uniform evidence suggests that a language model produces differently biased predictions depending on the prompt, specifically on the granularity of the labels requested, and suggests that the predictions our prompts produced may not correspond to biased behaviour in other tasks in a straightforward way. Importantly, recent studies show that LLMs' predictions are not robust under variation of scoring methods.

Social bias can be considered part of the value system in a particular culture and, as many studies have shown (Tao et al., 2024; Karinshak et al., 2024), LLMs used in non-English scenarios frequently mirror the cultural and social knowledge from a dominant language to non-dominant ones. When it comes to the conceptualisation of ethnicity in Slovenia, none of the employed models have probably seen enough local or regional data to encode ethnic bias in the same way we believe it to be present in society. Together with the trend to build local and more culture-aware LLMs, we may see more ethnic bias in the future.

In future work, we would like to compare our results with human annotations of the dataset, and to extend the dataset to other South Slavic languages. Furthermore, we plan to produce other datasets to explore social bias in Slovenian.



(a) Gender bias

(b) Ethnicity bias

Figure 1: For each Language Model, the distribution of mean sentiment scores per sentence template × emotion word group is compared for (a) female vs male and (b) non-Slovenian, none (no name mentioned), and Slovenian. The black lines connect each unique template × emotion word group.

Model	F − M		Slo − Non-Slo	
	Mean (SD)	p-value	Mean (SD)	p-value
SloBERTa	-0.0035 (0.0293)	0.1526	**0.0138 (0.0363)**	**1.047e-05**
GaMS 9B	0.0052 (0.0561)	0.2612	0.0148 (0.0801)	0.0282
GaMS 27B	0.0018 (0.0509)	0.6753	-0.0006 (0.0343)	0.8460

Table 3: Mean, standard deviation (SD), and p-values of bias scores (Δ) for gender and ethnicity. If $\Delta < 0$, the score is more negative towards female (gender) or Slovenian (ethnicity). If p-value < 0.025, then the result is statistically significant.

9. Acknowledgements

We acknowledge the financial support from the Slovenian Research Agency ARIS via the projects EMMA (Embeddings-based techniques for Media Monitoring Applications; L2-50070), LLM4DH (Large Language Models for Digital Humanities; GC-0002), SOVRAG (Hate speech in contemporary conceptualisations of nationalism, racism, gender and migration; J5-3102), CroDeCo (Cross-Lingual Analysis for Detection of Cognitive Impairment in Less-Resourced Languages; J6-60109), and research core funding for the programme Knowledge Technologies (P2-0103).

This work was also partially supported by the European Union's Horizon Europe research and innovation programme under grant agreement No 101214398 (ELLIOT). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them.

JC is supported by the Young Researcher Grant PR-13409. BK is funded by Young Researcher Grant PR-12394. MP is also supported by UKRI through the grant Responsible AI UK (EP/Y009800/1) keystone project AdSoLve (Addressing Socio-technical Limitations of LLMs for Medical and Social Computing; KP0016).

Finally, we thank students of the Digital Linguistics Joint Master Programme at the University of Ljubljana, in particular Vanessa Sobočan and Veronika Durn, for their involvement in the implementation of the IAT test.

9.1. Ethical considerations and limitations

Although in the present work we address gender as a binary variable, following the structure of the EEC corpus (Kiritchenko and Mohammad, 2018), we are aware that gender identity is a spectrum (Matsuno and Budge, 2017) and that this should be addressed in NLP studies (Dev et al., 2021). Similarly, in our evaluation we looked at ethnicity as a binary variable (Slovenian, Non-Slovenian), but considering it in a more fine-grained manner might bring to further identification of racial bias, which is often group-specific. We selected only three models for detailed analysis, but we initially obtained three-point prompts from 9 models, which had a mean agreement (calculated as the sum of numeric representation of the categories per sentence, divided by 9) of 0.745 (SD = 0.235) in all combinations of variables. Importantly, our findings only apply to the specific versions of the language models we used. For the time being, contrary to

Kiritchenko and Mohammad (2018), we did not calculate Δ between each pair of minimally distant sentences (which can be done with generic nouns like *my sister* or *my brother*, but not with personal names—e.g., *Ema* or *Suki*—since they cannot be coupled with precision), but only within sentences with the same template and emotion word. Conversely, we did not account for the effect of emotion word in our regression analyses. Since the effect of the emotion words on the valence scores assigned to each sentence by the LLMs is likely to be large, our analysis may not have been sensitive enough to detect the (likely much smaller) effect of bias on the valence scores assigned to each sentence by the LLMs. Further work could include random intercepts to model the effect of the valence of emotion words explicitly. Additionally, the results for ethnic bias may be unreliable due to the limited number of foreign names (n = 20) and possible tokenisation issues affecting non-Slovene names. Furthermore, although the GaMS models were fine-tuned on Slovenian data, they are derived from English pre-trained models. It is therefore unlikely that a systematic bias against ex-Yugoslavian names was introduced during fine-tuning if it was not already present in the original English model. Additional analysis would be required to investigate the reasons of discrepancy between the IAT and the models' results, as well as discrepancies across models. Employing additional bias-detection methods would be needed to provide a more comprehensive overview of the presence of bias in the analysed models. Future work will involve comparing the amount and quality of local fine-tuning and instruction-tuning data to better understand the sources of bias, as previous research in non-English contexts shows that this relationship is not straightforward.

10. Bibliographical references

Rishabh Bhardwaj, Navonil Majumder, and Soujanya Poria. 2021. Investigating gender bias in BERT. *Cognitive Computation*, 13(4):1008–1018.

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.

Mojca Brglez, Jaya Caporusso, Damar Hoogland, Boshko Koloski, Senja Pollak, and Matthew Purver. 2024. Slovenian emotion dimension and emotion association lexicon SloEmoLex 1.0. Slovenian language resource repository CLARIN.SI.

Paul-Christian Bürkner. 2017. brms: An R package for Bayesian multilevel models using Stan. *Journal of statistical software*, 80:1–28.

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

Jaya Caporusso, Damar Hoogland, Mojca Brglez, Boshko Koloski, Matthew Purver, and Senja Pollak. 2024. A computational analysis of the dehumanisation of migrants from Syria and Ukraine in Slovene news media. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 199–210, Torino, Italia. ELRA and ICCL.

Rune Haubo Bojesen Christensen. 2019. ordinal—regression models for ordinal data. *R package version*, 10(2019):54.

Zhibo Chu, Zichong Wang, and Wenbin Zhang. 2024. Fairness in large language models: A taxonomic survey. *ACM SIGKDD explorations newsletter*, 26(1):34–48.

Pieter Delobelle, Ewoenam Kwaku Tokpo, Toon Calders, and Bettina Berendt. 2022. Measuring fairness with biased rulers: A comparative study on bias metrics for pre-trained language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 1693–1706. Association for Computational Linguistics.

Erik Derner and Kristina Batistič. 2025. Gender representation bias analysis in LLM-generated Czech and Slovenian texts. In *Proceedings of the 10th Workshop on Slavic Natural Language Processing (Slavic NLP 2025)*, pages 124–135, Vienna, Austria. Association for Computational Linguistics.

Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff Phillips, and Kai-Wei Chang. 2021. Harms of gender exclusivity and challenges in non-binary representation in language technologies. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1968–1994, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yupei Du, Yuanbin Wu, and Man Lan. 2019. Exploring human gender stereotypes with word association test. 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 6133–6143.

Akram Elbouanani, Evan Dufraisse, and Adrian Popescu. 2025. Analyzing political bias in LLMs via target-oriented sentiment classification. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 15476–15505, Vienna, Austria. Association for Computational Linguistics.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and fairness in large language models: A survey.

Seraphina Goldfarb-Tarrant, Adam Lopez, Roi Blanco, and Diego Marcheggiani. 2023. Bias beyond English: Counterfactual tests for bias in sentiment analysis in four languages. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4458–4468, Toronto, Canada. Association for Computational Linguistics.

Anthony G Greenwald and Mahzarin R Banaji. 1995. Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychological review*, 102(1):4.

Anthony G. Greenwald and Linda Hamilton Krieger. 2006. Implicit bias: Scientific foundations. *California Law Review*, 94(4):945–967.

Jules Holroyd, Robin Scaife, and Tom Stafford. 2017. What is implicit bias? *Philosophy Compass*, 12(10):e12437. E12437 PHCO-1075.R1.

Project Implicit. 2011. About the iat.

Sophie Jentzsch and Cigdem Turan. 2022. Gender bias in BERT - measuring and analysing biases through sentiment rating in a realistic downstream classification task. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 184–199, Seattle, Washington. Association for Computational Linguistics.

Mahammed Kamruzzaman, Md. Minul Islam Shovon, and Gene Louis Kim. 2024. Investigating subtler biases in LLMs: Ageism, beauty, institutional, and nationality bias in generative models.

Elise Karinshak, Amanda Hu, Kewen Kong, Vishwanatha Rao, Jingren Wang, Jindong Wang, and Yi Zeng. 2024. LLM-GLOBE: A benchmark evaluating the cultural values embedded in LLM output. *arXiv preprint arXiv:2411.06032*.

Svetlana Kiritchenko and Saif Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. pages 43–53.

Hannah Rose Kirk, Yennie Jun, Filippo Volpin, Haider Iqbal, Elias Benussi, Frederic Dreyer, Aleksandar Shtedritski, and Yuki Asano. 2021. Bias out-of-the-box: An empirical analysis of intersectional occupational biases in popular generative language models. *Advances in neural information processing systems*, 34:2611–2624.

Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in large language models. In *Proceedings of The ACM Collective Intelligence Conference*, CI '23, page 12–24. ACM.

Anne Lauscher and Goran Glavaš. 2019. Are we consistently biased? multidimensional analysis of biases in distributional word vectors. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 85–91, Minneapolis, Minnesota. Association for Computational Linguistics.

Qianying Liu, Katrina Qiyao Wang, Fei Cheng, and Sadao Kurohashi. 2025. Assessing large language models in agentic multilingual national bias. *arXiv preprint arXiv:2502.17945*.

Konrad Löhr, Shuzhou Yuan, and Michael Färber. 2025. The hidden bias: A study on explicit and implicit political stereotypes in large language models.

Ivy W Maina, Tanisha D Belton, Sara Ginzberg, Ajit Singh, and Tiffani J Johnson. 2018. A decade of studying implicit racial/ethnic bias in healthcare providers using the implicit association test. *Social science & medicine*, 199:219–229.

Emmie Matsuno and Stephanie L Budge. 2017. Non-binary/genderqueer identities: A critical review of the literature. *Current Sexual Health Reports*, 9(3):116–120.

Julia Mendelsohn, Yulia Tsvetkov, and Dan Jurafsky. 2020. A framework for the computational linguistic analysis of dehumanization. *Frontiers in artificial intelligence*, 3:55.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages

1953–1967, Online. Association for Computational Linguistics.

Shangrui Nie, Michael Fromm, Charles Welch, Rebekka Görge, Akbar Karimi, Joan Plepi, Nazia Afsan Mowmita, Nicolas Flores-Herr, Mehdi Ali, and Lucie Flek. 2024. Do multilingual large language models mitigate stereotype bias? *arXiv preprint arXiv:2407.05740*.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. [Computer software].

Sachin Rawat and G Vadivu. 2022. Media bias detection using sentimental analysis and clustering algorithms. In *Proceedings of international conference on deep learning, computing and intelligence: ICDCI 2021*, pages 485–494. Springer.

Annabella Sakunkoo and Jonathan Sakunkoo. 2025. Name of thrones: Evaluating how LLMs rank student names, race, and gender in status hierarchies.

Deven Santosh Shah, H. Andrew Schwartz, and Dirk Hovy. 2020. Predictive biases in natural language processing models: A conceptual framework and overview. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5248–5264, Online. Association for Computational Linguistics.

Yan Tao, Olga Viberg, Ryan S. Baker, and Rene F. Kizilcec. 2024. Cultural Bias and Cultural Alignment of Large Language Models. 3(9):pgae346.

Polina Tsvilodub, Hening Wang, Sharon Grosch, and Michael Franke. 2024. Predictions from language models for multiple-choice tasks are not robust under variation of scoring methods. *arXiv preprint arXiv:2403.00998*.

Matej Ulčar and Marko Robnik-Šikonja. 2021. SloBERTa: Slovene monolingual large pretrained masked language model. *Proceedings of Data Mining and Data Warehousing, SiKDD*, pages 17–20.

Matej Ulčar, Anka Supej, Marko Robnik-Šikonja, and Senja Pollak. 2021. Slovene and Croatian word embeddings in terms of gender occupational analogies. *Slovenščina 2.0: empirične, aplikativne in interdisciplinarne raziskave*, 9(1):26–59.

Špela Vintar. 2025. Slovenian equity evaluation corpus EEC-SL 1.0. Slovenian language resource repository CLARIN.SI.

Domen Vreš, Martin Božič, Aljaž Potočnik, Tomaž Martinčič, and Marko Robnik-Šikonja. 2024. Generative model for less-resourced language with 1 billion parameters.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in English? on the latent language of multilingual transformers.

Yuemei Xu, Ling Hu, Jiayi Zhao, Zihan Qiu, Kexin Xu, Yuqi Ye, and Hanwen Gu. 2025. A survey on multilingual large language models: Corpora, alignment, and bias. *Frontiers of Computer Science*, 19(11):1911362.