Journal of Computational Social Science (2025) 8:57
https://doi.org/10.1007/s42001-025-00382-y

RESEARCH ARTICLE

'Slightly disappointing' vs. 'worst sh** ever': tackling cultural differences in negative sentiment expressions in AI-based sentiment analysis

Franziska Sofia Hafner[1,2] · Lena Hafner[3] · Roberto Corizzo[1]

Received: 15 June 2024 / Accepted: 31 March 2025 / Published online: 21 May 2025

Abstract

Advertisers, politicians, and social scientists alike have an interest in the current zeitgeist. With people expressing their sentiments in online comments, the Internet has become a great source for 'reading' the zeitgeist via artificial intelligence-based sentiment analysis (SA). At this, negative sentiments are of special concern. They can serve as early indicators for events that require action, such as dropping customer satisfaction, discontent amongst potential voters, or a threat of social unrest. However, due to cultural differences in how negative sentiments are expressed, conventional SA methods typically classify such texts with higher accuracy for some ethnicities compared to others. In this paper, we demonstrate this using a large real-world corpus of Google Maps reviews. Across eight ethnic groups, linguistic patterns vary more starkly in negative (1 star) reviews than in neutral (2–4 star) and positive (5 star) reviews. Consequently, ethnicity-blind SA methods 'struggle' to classify negative reviews correctly. To mitigate this problem, we propose a novel SA method, based on balanced training and subsequent ethnicity-conscious fine-tuning. Our approach is simultaneously able to mitigate bias and enhance overall model performance. Thus, we hope to contribute to a more equal appreciation of negative

✉ Roberto Corizzo
rcorizzo@american.edu
https://scholar.google.com/citations?hl=en&user=vAp8J1wAAAAJ

Franziska Sofia Hafner
franziska.hafner@oii.ox.ac.uk
https://scholar.google.com/citations?user=v661CCgAAAAJ&hl=en&oi=ao

Lena Hafner
lh623@cam.ac.uk
https://scholar.google.com/citations?user=zPcdffsAAAAJ&hl=en&oi=ao

[1] Department of Computer Science, American University, 4400 Massachusetts Avenue NW, Washington, DC 20016, USA

[2] Oxford Internet Institute, University of Oxford, 1 St Giles, Oxford OX1 3JS, UK

[3] Department of Politics and International Studies, University of Cambridge, Cambridge CB2 1TN, UK

sentiments of ethnically diverse customer-, voter-, or research populations, and, consequently, a more nuanced approximation of the overall zeitgeist.

Keywords Polarity detection · Sentiment analysis · Algorithmic fairness · Bias

1 Introduction

'With the greatest respect,…', - if a British person addresses you with these words you might think 'great, they respect me'. However, a quick glance into the Anglo-EU Translation Guide [1] reveals that a more accurate translation is 'I think you are an idiot'. This is just one example from 'popular science', but it showcases a central tenet of the 'real' science of cultural studies: A nuanced understanding of intercultural linguistic differences is necessary to correctly interpret the sentiment behind words. A long tradition of ethnographic work in 'real-world' settings across the globe, and more recently also digital ethnographies of Internet platforms, has shown that these linguistic differences are especially stark when it comes to antagonistic contexts, such as conflicts and complaints.

However, this insight has not yet arrived in the field of machine learning that attempts to 'read' sentiments on scale, namely sentiment analysis (SA). Traditional SA models treat language as uniform [2, 3]. Recently, critical SA scholars have shown that this can lead to distorted sentiment predictions when explicit identity terms (e.g. gay, jew) are mentioned in the text, and proposed solutions to mitigate this kind of 'explicit' bias [4–9]. Furthermore, a few authors have pinpointed that implicit linguistic choices might also lead to skewed accuracy rates [10–12]. However, no efficient mitigation strategy for implicit bias has been presented yet [2]. Furthermore, to the best of our knowledge, no work in SA so far has strived to account for the additional nuance that linguistic patterns not only vary by culture but also by context.

In this paper, we fill this gap by proposing a novel SA method, based on balanced training and subsequent ethnicity-conscious fine-tuning. In our approach, ethnicity detection is used to group reviews by ethnicity. This information is first used to balance data and train a general model. Subsequently, the general model is fine-tuned using only training data pertaining to a specific ethnicity group, resulting in ethnicity-specific models. This approach is simultaneously able to mitigate bias and enhance overall model performance.

The rest of the paper is structured as follows. Section 2.1 dives deeper into the cultural studies background, and, Sect. 2.2 into the SA background. Thirdly, Sect. 3 brings both spheres together by presenting our own SA work. In Sect. 3.1 we connect the two spheres by showing that in our large corpus of Google Maps reviews, we can really trace the linguistic differences that cultural studies lead us to expect. Section 3.2 describes our methodology for developing a novel, ethnicity-aware SA model. Section 4.5 discusses the experimental results, which also confirm that it is the negative (1-star) Google Maps reviews in which ethnicity-aware SA makes the largest improvements in accuracy rates. Section 4.6 makes suggestions for further research. Section 5 wraps up the paper.

2 Background

2.1 Bias in cultural linguistics

Feeling is universal. A large body in anthropological and psychological research suggests that people across the globe hold feelings, which they attribute both to themselves as well as others. The social sciences have thus integrated the concept of 'feeling' as part of the universal folk model of a person and deem it 'safe' to be used for the investigation of human experience everywhere [13].

In contrast, emotions and sentiments are culture-specific. In the words of Shouse, they are the mechanisms with which we 'broadcast our feelings to the world' [14]. This broadcasting can sometimes reflect our inner state, and sometimes social expectations, but it always follows the culturally dependent 'script' internalized through the upbringing within specific cultural norms. Mesquita et al. trace how sentiments and emotions are actively constructed to meet the demands of the respective cultural environment - starting with children's books (e.g. depiction of children with calm expressions in Japan, children with emotional expressions in the US), parental guides, and religious texts [15]. Recursively, these culturally normative patterns become entrenched as people benefit from experiencing them. Emotional 'fallouts' are socially sanctioned, whereas emotional conformity enables individuals to navigate their social environment in a coordinated fashion. On a societal level, this emotional mainstreaming is argued to be key for achieving 'collective intentionality' in cohesive cultural groups [16].

Then, how exactly do emotions and sentiments differ across cultural groups? This is the central question in cultural studies. As emotions are 'the most revealing indicators of cultural similarities, and of cultural differences' [17], they can serve as visible flags for other underlying cultural dimensions. The most widely accepted framework of cultural dimensions was developed by Hofstede. He distinguishes cultures along six axes: individualism vs. collectivism, high vs. low power distance, masculinity vs. femininity, uncertainty avoidance vs. embracement, long vs. short-term orientation, and indulgence vs. restraint [18, 19]. On which side of these dimensions cultures are located is influenced by their philosophical or religious traditions. For instance, Protestantism leads cultures to hinge on the side of individualism, as it promotes individual salvation and personal accountability. In contrast, Confucianism correlates with collectivism, as it promotes group harmony and the subordination of personal desires to societal needs. Whereas these spiritual creeds play a pivotal role in embedding cultural dimensions, worldly structures, such as family-, educational- or economic systems, reinforce them. For instance, Protestantism has been linked to early industrialization as individualist cultures create individuals suitable for the competitive and independent nature of industrial work. In contrast, Confucianism relates to long histories of agricultural reliance, as farming required cooperation in collective family units [20]. In short, interlocking social systems lay the groundwork for the cultural dimensions' broad geography.

The manifestations of these dimensions were charted on the basis of, and successively tested in, 'real world' ethnographies. Hofstede himself distilled the dimensions from observations of IMB employees across the globe. Later

research sites range from international space missions [21, 22], and multinational businesses [23, 24], to foreign exchange programs in universities [25]. Emblematic observations from these studies are, for instance, Japanese students expressing their utmost disagreement with American teachers by answering 'probably', which is misunderstood by an American audience as a lack of knowledge instead of disagreement [25]. From such interactions, ethnographers developed meso-level ethnolinguistic theories that bridge the gap between Hofstede's macro-level cultural dimensions and the way those are operationalized in concrete linguistic expressions. Three such theories stand out.

Firstly, Hall's theory of high- and low-context communication. 'High-context' means that communication relies heavily on implicit messages, which depend more on the context than the words themselves. Linguistically, this traces through to more indirect speech, circumlocution, and hedging. This linguistic style tends to grow out of collectivist cultures as their emphasis on group harmony mandates avoiding direct confrontation to maintain social cohesion [26, 27]. In contrast, low-context communication values clarity and directness. This is reflected in the language by more direct speech, 'I'-perspective, and speech acts (e.g. questions, imperatives,...). This style of communication is characteristic of individualistic cultures. Given that those cultures have less of a social net that already holds part of the meaning through implicit, shared understandings, they need more verbal precision to ensure individual understanding [28].

Secondly, Brown and Levinson's politeness theory. This theory holds that in collectivist, high-context cultures 'negative' politeness strategies prevail. This means that politeness consists of not mentioning or only indirectly expressing negative sentiments, to avoid imposing on others. In contrast, in individualist, low-context cultures positive politeness strategies are the socially desired standard. Linguistically this is expressed through more vocabulary pertaining to an open declaration of emotions, flanked by phrases of civility [29].

Thirdly, the Sapire-Whorf theory of linguistic relativity. It adds the perspective that language itself shapes thoughts and perceptions. For instance, collectivist, high-context cultures tend to regard emotions as fluid. Thus, they have produced a nuanced, mediated vocabulary for emotions. An example is the Japanese concept of 'amae', which encapsulates varied feelings of dependence and affection. In contrast, in individualist, low-context cultures the need for clarity has produced a categorical understanding of emotions (e.g., 'happy' and 'angry', as one-dimensional labels of emotions). This is recursive, as the available vocabulary constraints or enables the emotions and behaviors that are cognitively possible [30]. Thus, the Sapire-Whorf theory ties back to Hofstede's macro-level of cultural dimensions, by explaining not only how cultural dimensions shape linguistic features, but also how these linguistic features in turn perpetuate the distinct cultural dimensions.

While all these theories have been derived from 'real-world' ethnographies, more recent cultural studies have exploited the vast amounts of user-generated text on the Internet to put these earlier theories to test in 'digital ethnographies'. For instance, Fang et al. [31] and Feng and Ren [32] compare reviews on Amazon US and Amazon China, Fong and Burton [33] online discussion boards in the US to those in China, Tsang and Prendergast [34] computer game reviews by Chinese and

American customers, and Cenni and Goethals [35] hotel reviews on TripAdvisor written in English, Dutch, and Italian.

All these accounts conclude that Hofstede's dimensions, as well as the ethnolinguistic theories tied to them, have traveled well into the digital age: They seem to be equally applicable to the impersonal social environment of the Internet as to the face-to-face context they were developed in. For instance, Chinese as representatives of a collectivist, high-context culture were found to provide fewer negative reviews than their individualistic, low-context counterparts, in line with the high value of 'giving face to others' to maintain collective harmony [31, 34]. Furthermore, those reviews that were negative differed markedly: Chinese negative reviews contained less emotional expressions [36, 37] and more euphemisms, hedges, negations, and down-toners [32, 38]. In contrast, Western consumers' online complaints contained more on-record impoliteness and sarcasm [32] and co-occurred more frequently with speech acts [39]. These linguistic patterns are in line with the high value of the 'self' in individualistic cultures.

The 'real world' observations as well as the digital ethnographies converge in the observation that the greatest intercultural linguistic differences occur in negative contexts. This is the case as in positive or neutral situations, less 'masking' of feelings is necessary. However, when it comes to regulating the antagonism inherent in negative situations, cultural scrips kick in. These alter significantly how negative emotions are broadcast around the world.

2.2 Bias in sentiment analysis

This distinction between negative and positive emotions has not traveled through to computerized methods of sentiment analysis. The digital ethnographies that have corroborated this distinction rely on manual coding of user-generated text. This approach is labor intensive and thus only feasible for a small fraction of the massive repository of emotional content that is the Internet. However, computer scientists have developed algorithms that can apply linguistic rules to electronically 'read' sentiments. This so-called sentiment analysis (SA) has thus far enabled wide-scale analysis in multiple fields. For instance, in business studies, it is applied to understand people's inclinations for targeted marketing and brand monitoring [40]. In the field of politics, Smith [41] use SA on Tweets to predict election results, Georgiadou et al. [42] to analyze public sentiment on Brexit, and Rozado and al-Gharbi [43] to measure political sentiment polarization in left- and right-leaning news outlets. In the social sciences, Balahur et al. [44] use SA for opinion summarization, Bollen et al. [45] and Mogilner et al. [46] for the expression of collective moods, and Saito and Haruyama [47] to estimate social sentiments in response to COVID-19 incident rates.

Despite the uptake in so many areas, the exchange between cultural studies and computer science has been scant. The distance seems to be mutual, with cultural scientists being criticized for still overly relying on manual 'close reading' [48], while computer scientists have thus far largely failed to integrate insights from cultural

studies into their SA tools, such as context-dependent cultural-linguistic differences in sentiment expression patterns.

This is especially true for early advances in SA, which followed a 'one-size-fits-all' mindset. SA models became increasingly far-reaching - from identifying sentiments at the aspect level (e.g. the word 'terrible'), to sentence level (e.g. the sentence 'I am not pleased with the result'), to document-level analysis (e.g. the book 'Brave New World') [49]. Also, their underlying structure became increasingly complex - from lexicon-based approaches, that rely on sets of words annotated with their sentiment score (e.g. WorldNet, AFINN, and SentiWordNET) to machine-learning and artificial intelligence methods, (e.g. based on Word2Vec, Doc2Vec, GloVe, and BERT), to user-friendly web-based applications [50]. However, all these approaches have in common that they treat language as uniform [2].

More recently, however, computer scientists started to challenge one aspect of the notion of language as a 'homogeneous plain'. They realized that some words, namely identity terms of certain demographic groups, elicit disproportionately strong sentiment scores. Since their work is restricted to terms explicitly mentioned in the text, Liu et al. [12] coined the term 'explicit bias'. The most common method to tease out explicit bias in SA models is counterfactual analysis. This analysis consists of strategically swapping only the identity terms in otherwise identical sentences (e.g. replacing 'this *man* made me feel angry' with 'this *woman* made me feel angry') and testing how the predicted sentiment score varies. In this way, explicit bias has been extensively proven to exist (e.g. [4, 5, 7, 8]), with the most comprehensive study based on over 200 SA tools, finding that most provide consistently higher sentiment intensity predictions for one race or one gender [6]. Goldfarb-Tarrant et al. [9] extend counterfactual analysis, which has previously only been done on English-language SAs, to Japanese, Chinese, German, and Spanish. They show that while all language models suffer from some explicit bias, the German model is the most extreme, changing its sentiment prediction from very positive to very negative when the identity variable is changed from privileged to minoritized.

Attempts to remove explicit bias from SA models have been equally numerous. Since one of the main modes through which bias is imported into SA models is through training algorithms on (historical) texts that discriminate against minoritized groups, the first step to iron it out is to balance the training data. This has been done either by down-sampling, i.e. reducing the number of training entities to that available for the protected group [12, 51], or by up-sampling, i.e. creating new training entities by systematically swapping identity terms within training texts [52]. Furthermore, for lexicon-based SAs it has been shown that fairness can be improved by removing words from the lexicon that hold different sentiment scores for different identities (e.g. 'soft' being positively connotated for women, but negatively connotated for men) [53]. Lastly, for AI-based SAs, approaches have been suggested to modify the underlying algorithms, either through identifying and neutralizing gender-subspaces in world embeddings in Word2Vec [4], through quantifying social bias related to location names and removing it from BERT and ELMo embeddings [54], or through applying mutation strategies in the post-processing phase [55].

However, explicit bias is only the tip of the iceberg. While it is important to identify bias in language describing different demographic groups, we know from

cultural studies that the disparities in language generated by different demographics might be a source of bias, too. This less tangible but more perversive form of bias is referred to as implicit bias. As noted by Liu et al. [12], works studying implicit bias are very limited.

The few works investigating the topic show that implicit bias encroaches into SA models when the training data features stylistic differences in the word choices of writers from varying backgrounds [10]. For instance, a Standard American English (SAE) comment may be 'Can't wait to visit your new home. Yes, I'm going to be a great guest!', while the equivalent in African American English (AAE) is 'Can't wait to visit your new home. Yup, I goin to be a great guest!'. Even though the words 'yup' and 'goin' are irrelevant to the sentiment, their common use amongst African Americans hints the SA model to classify the first sentence as positive, but the second as negative [12]. This illustrates how a training set having more positive examples from white authors and more negative ones from black authors will lead the model to learn the 'shortcut' [56] of indiscriminately associating the language style of white people with positive sentiments and that of back people with negative sentiments.

The existence of such implicit bias has also been proven through counterfactual analysis. In this case, however, finding the paired synonyms for term swapping is less straightforward than for explicit bias. Nevertheless, Shen et al. [10] have developed a methodology to find implicit paired synonyms. They divide the training corpus by author demographic (e.g. only text written by white Americans, only text written by black Americans) and construct a world frequency distribution to distill the top 1000 differing words for each corpus segment. This analysis reveals world pairs like 'huge' (SAE) and 'big' (AAE), and 'darling' (SAE) and 'babygirl' (AAE). Performing substitutions such as changing 'Oh anytime darling. I want to see you soon!' to 'Oh anytime babygirl. I want to see you soon!' decreases the sentiment score. Both words are similar terms of endearment, but 'babygirl' is five times as likely to appear in the AAE segment than in the SAE segment. Thus, the model associates it with a more negative sentiment.

More recent studies have further corroborated the dialect-based prejudice against AAE-speakers. Groenwold et al. have created a dataset of 'intent-equivalent' SAE/AAE Tweet pairs. Even though those pairs are designed to convey the same core message, their test on the NLP model GPT-2 showed that the AAE Tweets were classified as more negative than their SAE-counterparts [57]. Hofmann et al. extend these findings by investigating four further language models (RoBERTa, T5, GPT−3.5, and GPT-4). They conclude that all these models embody covert racism by exhibiting racio-linguistic stereotypes against AAE-speakers. For instance, whereas the models rank a person who says 'I am so happy when I wake up from a bad dream because they feel so real' high on 'brilliance' and 'intelligence', they rank a person who says the same in AAE dialect ('I be so happy when I wake up from a bad dream cus they be feelin too real') high on 'dirtiness' and 'laziness' [58]. Lastly, Resende et al. test sentiment and toxicity scoring tools (Vader, TextBlob, Flair, Google's Perspective, and Detoxity) on YouTube-,Twitter- as well as on interview-based datasets, finding that the usage of AAE expressions causes the speaker to be considered substantially more toxic than non-AAE speakers, even when talking about the same topic [59].

While all these studies on implicit bias focus on AAE, Zhiltsova et al. [11] expand the field by comparing the performance of SA tools on native vs. non-native speakers of English. In this case, the substitution pairs are cognates, i.e. English words non-native speakers tend to use more since they have an origin in the speaker's native language (e.g. French English speaker 'fatigue' for SAE 'tiredness'). SA models discriminated against non-native speakers, as their sentiments were misclassified more often.

Mitigation strategies against implicit bias are still in their infancy. Most are lexicon-based. This means that they try to overcome bias by creating better-annotated lexica. Zhiltsova et al. [11], for instance, suggest adding cognates to the lexicon. Das [60] proposes an interactive game to collect sentimental polarity for intercultural lexica, in which people across the globe are presented with terms that they give a sentiment rating. Lin et al. have hired AAE speakers to re-write seven popular benchmarks, such as HumanEval and GSM8K, into AAE. The result is a dialectical benchmark called ReDial, which comprises more than 1.2 thousand parallel query pairs in SAE and AAE [61].

For AI-biased SAs, only one attempt to remove implicit bias was found: Hovy [2] uses the international customer review platform Trustpilot to collect texts that they can annotate with the reviewers' profile information. This enables them to construct sub-corpora for reviews by Danish, French, German, British English, and SAE-speakers. Then, they train different Word2Vec models: One on the entire unlabeled review texts and another one on each of the sub-corpora separately. They test how the latter training method can reduce bias in three language classification tasks: age classification, topic classification, and SA. Hovy themselves state that, while they made advances towards more equalized performances in the former two tasks, they did not do so for SA. They speculate that this might be due to their approach being too simplistic and not task-specific.

In the following, we suggest an improvement over this status quo. We offer an AI-based approach to tackle implicit bias specifically for the task of SA. At this, we, first, overcome the trade-off between fairness and overall accuracy which is often reported in AI de-biasing works [62]. Secondly, we bring in knowledge from cultural studies that goes beyond the simple recognition that differences exist between the linguistic patterns of different cultures, to the realization that these differences also depend on the negative or positive nature of the context. This allows us to provide a more nuanced understanding of why and in which cases ethnicity-aware SA is better suited for recognizing the sentiments of diverse populations.

3 Method

Our method analyzes reviews that include: *i*) the text written by the user, i.e., the natural language content of the review; *ii*) numerical ratings (e.g., star ratings from 1 to 5), which enable sentiment classification; *iii*) personal names, which enable us to infer the ethnicity of the reviewer. In the following, we describe our method in detail, focusing on its two main stages: ethnicity detection, and model training.

3.1 Ethnicity detection

In order to perform ethnicity bias mitigation in sentiment classification, we need to accurately identify the ethnicity of the authors of the Google Maps reviews. To this end, we group the reviews by ethnicity, leveraging the deep learning classification model proposed in [63]. To the best of our knowledge, this model is the most accurate and least biased model for name ethnicity classification to date, judged by its minimal deviation in sensitivities across age, gender, and ethnicity.

The adopted model in [63] has a CNN+LSTM architecture with dropout layers trained via a Negative-Log-Likelihood loss function (NLLLoss) with CompaniesHouse, a public dataset comprehensive of names and ethnicities of United Kingdom (UK) business owners. To deal with class imbalance, data augmentation was performed across the ethnicities to provide them with equal representation in the model.

Now, to segment our training data, we define $R = \{(x, y)\}$ as the global set of user reviews, where x is a text review, and $y \in \{1, 2, 3, 4, 5\}$ is its corresponding star rating. The ethnicity detector acts as a function $\phi : R \rightarrow E$, assigning each review $r \in R$ to one of the target ethnicity groups $E = \{$ Anglo-American, European, African, Arabic, East Asian, Hispanic, South Asian, Scandinavian $\}$. This function allows reviews to be grouped by ethnicity group as:

$$R_G = \left\{ \bigcup_{e \in E} R_e \subset R \mid \forall r \in R_e \ \ \phi(r) = e \right\}, \tag{1}$$

where each subset R' includes reviews for a specific group.

Segmenting our training data by ethnicity in this way allows us to put in place bias mitigation measures. This is the case as it allows us to take into account ethnicity in the sentiment classification workflow, e.g., by balancing training data based on ethnicity groups and implementing our bias-aware fine-tuning stage. Moreover, it allows us to test and compare the accuracy of sentiment classification for each group.

3.2 Model training

Model training takes place in two main stages (see Fig. 1). First, a general model m is trained on all reviews. More formally:

$$m = fit(m_0, R, \ \mathcal{L}_{base}), \tag{2}$$

$$\mathcal{L}_{base} = \frac{1}{|N|} \sum_{i=1}^{N} (\hat{y}_i - y_i), \tag{3}$$

where *fit* is a function that trains an initial model m_0 via stochastic gradient descent using the loss $\mathcal{L}_{base}$. The loss computation considers all reviews equally, i.e., considering, for each review, the predicted star rating $\hat{y}_i$ and the corresponding actual

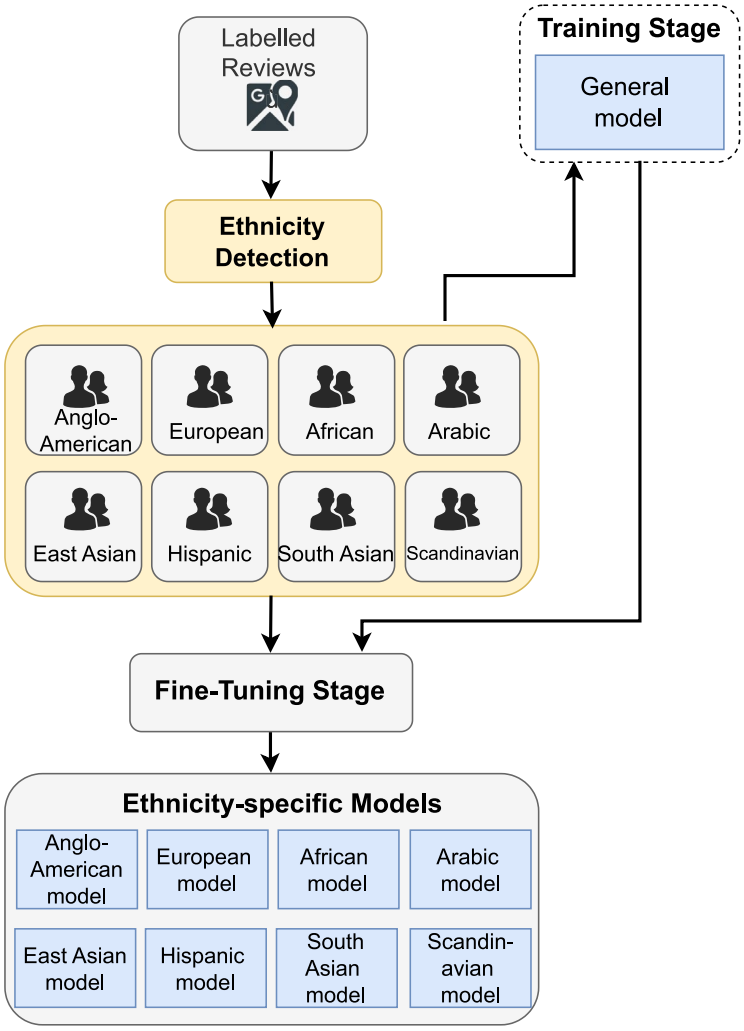

Fig. 1 Overview of the proposed method (training stage). Ethnicity detection is used to group reviews by ethnicity. This information is first used to balance data and train a general model. Subsequently, the general model is fine-tuned using only training data pertaining to a specific ethnicity group, resulting in eight ethnicity-specific models. This process leads to debiased sentiment classification models that handle ethnicity-specific language more appropriately

(ground truth) star rating y_i, without explicitly taking into account the ethnicity group of each reviewer.

The following stage involves fine-tuning ethnicity-specific models. To this end, a pool M of models is built, where each model is fine-tuned with the set of reviews of a specific ethnicity group R_e:

$$M = \left\{ \bigcup_{\forall e \in E} m_e \mid \quad m_e = fit(m, R_e, \mathcal{L}_{base}), \quad R_e \subset R_G \right\} \tag{4}$$

During the training stage, the model receives training data in batches. Each batch contains text reviews and the ground truth label (star rating) associated with them. Then, the model tries to predict the star rating (as a proxy for sentiment value) of each review. In order to evaluate how successful the model was at this task, and to update it accordingly, a standard loss function calculates the loss as the average difference between the predicted and the true value of the star rating for each batch member. Throughout the model training, the model performance is improved by minimizing the loss. A pseudo-code of our model training approach is described in algorithm 1.

Our approach for de-biasing SA is designed to take two aspects into consideration: (i) improve performance overall, (ii) reduce performance inequalities between ethnicity groups. Thus, in addition to reviews and their gold standard sentiment value (as in the standard loss), we also provide the ethnicity group associated with each review in the batch.

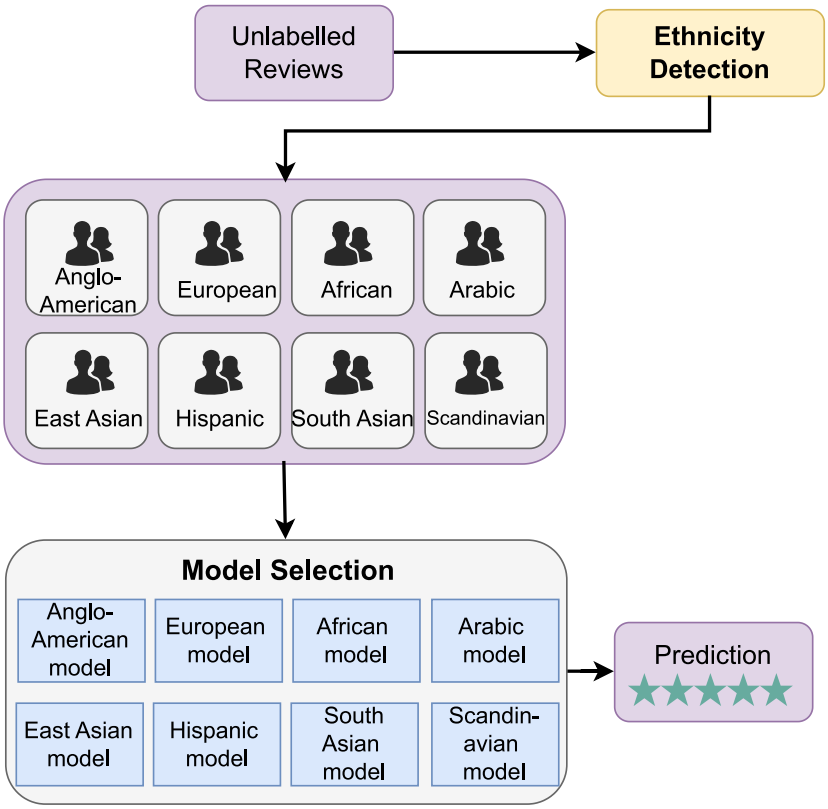

Fig. 2 Overview of the proposed method (inference stage). Given a set of unlabelled reviews, name-to-ethnicity detection is performed to infer their ethnicity group. Grouped reviews allow us to select the most appropriate ethnicity-specific pre-trained model for the classification task. The selected model extracts predictions (5-star ratings) for each group of unlabelled reviews

Given an unlabelled review r_u, the inference workflow that leads to a prediction $\hat{r_u}$ (see Fig. 2) can be formalized as:

$$\hat{r_u} = predict(\gamma(M,\ \phi(r_u)),\ r_u), \tag{5}$$

where ϕ infers the ethnicity group from the review r_u, γ is a gating operator that selects the correct ethnicity-specific model m_e from the pool of models M, and *predict* is a function that extracts a star rating prediction given model m_e and review r_u.

Our approach is coherent with the observation that models with multiple branches, such as multi-head attention in transformers, typically correspond to different grammatical functions [64] (e.g. co-reference and noun modifiers), linguistic properties, or downstream tasks [65]. Our pool of models serves the function of 'storing' various linguistic properties pertaining to linguistic patterns that are more common for some ethnicity groups compared to others. Considering this aspect should, in principle, allow us to mitigate bias across ethnicities and maximize the performance of each model for each respective ethnic group. This way, the accuracy achieved for each ethnicity-specific model should constitute an upper bound regarding models trained with the given amount of data and the presented model architecture for each respective ethnicity.

Algorithm 1 Ethnicity-aware model training approach

Input: R - User reviews
Result: M - Pool of trained ethnicity-specific models

Initialize set of grouped reviews $R_G \leftarrow \emptyset$
Initialize model pool $M \leftarrow \emptyset$

$B \leftarrow$ Split R into mini-batches $\{B_1, B_2, \ldots, B_N\}, B_i \subset R$
for $B_i \in B$ **do**
  $m \leftarrow \text{train}(m, B_i)$ `// Train general model on all reviews`
end

$R_G \leftarrow \phi(R)$ `// Group reviews via Ethnicity Detection (Eq. 1)`

for $R_e \in R_G$ **do**
  $B_e \leftarrow$ Split R_e into mini-batches $\{B_{e1}, B_{e2}, \ldots, B_{eN}\}, B_{ei} \subset R_e$
  $m_e \leftarrow m$ `// Copy model: initial weights from general model`
  for $B_{ei} \in B_e$ **do**
    $m_e \leftarrow \text{train}(m_e, B_{ei})$ `// Ethnicity-specific Fine Tuning`
  end
  $M \leftarrow M \cup m_e$
end
return M

4 Experiments

4.1 Research questions

We aim to answer the following research questions:

- *RQ1*: Do ethnicity groups' linguistic differences make the adoption of ethnicity-aware sentiment classification necessary?
- *RQ2*: Is our method able to achieve a competitive overall sentiment classification performance when compared with popular baselines?
- *RQ3*: Is our method able to effectively mitigate bias by improving performance metrics separately for each ethnicity group?
- *RQ4*: How are specific sentiment scores (i.e., negative and positive contexts) impacted by the adoption of our ethnicity-conscious approach?

4.2 Setup

To implement our approach, we adopt a *bert-base-cased* model fine-tuned on our review corpus using the Adam optimizer. The model has an input size of 256, an intermediate layer of size 512, and an output layer of size 5 corresponding to 5-star ratings.

The selected values for the hyperparameters are motivated by works in the literature providing effective heuristics. Specifically, for learning rate, [66] suggests starting from a default value of 0.01 and experimenting with a decreasing factor (negative power of 10), where 10^{-6} is considered an extremely small value. Similarly, for batch size, powers of 2, i.e., 8, 16, 32 are recommended values.

Following these guidelines, the learning rate in our experiments is optimized on the validation data (10% of the training data) using a scheduler starting from 1e−5 with a decay of 1e−6. For all other hyperparameters, we set the batch size to 16 and epochs to 10. Larger values for batch size significantly increased computational requirements, whereas an increased number of epochs did not yield significant improvements.

Our method is implemented using Python and TensorFlow 2. Experiments are performed on a workstation equipped with an Intel Xeon W-2145 (3.7GHz) CPU, 64GB of RAM (DDR4-2666), 512GB SSD drive, and an NVIDIA RTX 4090 GPU.

4.3 Dataset

We obtained our dataset by extracting reviews from Google Maps business pages of UK-based companies. For this purpose, we leverage the list of companies on the UK government webpage *CompaniesHouse*, which provides company names and addresses. For each company identified with a valid Google Maps entry, we downloaded up to 100 reviews, which include the reviewer's name, the text content of the review, and the assigned star-rating. In our experiments, we leverage

our dataset containing 40,000 reviews, 8000 for each star rating, with an equal distribution of reviews for ethnicities, resulting in 1000 reviews for each ethnicity group and star rating.

The training set for the initial model is balanced across all ethnicity groups by down-sampling to the number of reviews from the smallest ethnicity group. Moreover, the training set used for fine-tuning each ethnicity-specific model is in turn balanced across star ratings, resampling reviews of each star rating to match the smallest group size.

Similarly, for the testing set, we conducted sampling with replacement to ensure equal representation of the smallest ethnicity groups, resulting in a total of 240,000 reviews: 6000 for each star rating and 30,000 for each of the eight ethnicity groups. The substantial size of our testing set enables a more detailed examination of the performance within individual ethnic groups, allowing us to determine the statistical significance of our findings more accurately.

4.4 Metrics

We utilize conventional metrics within machine learning-based classification, including Precision (P), Recall (R)-also recognized as Sensitivity-and the F-Measure ($F1$), which are defined as:

$$P = \frac{T_p}{T_p + F_p}; \quad R = \frac{T_p}{T_p + F_n}; \quad F1 = 2 \times \frac{P \times R}{P + R},$$

where T_p represents the count of true positives, and F_p denotes the count of false negatives. To account effectively for class imbalance, we employ weighted versions of Precision, Recall, and F-Measure. These metrics involve individual calculations for each label, and their average is weighted by support, signifying the number of true instances for each label.

4.5 Discussion

Firstly, we evaluate whether significant linguistic disparities among ethnic groups can be identified in our review corpus. We do this by comparing the performance of the ethnicity-specific sub-corpora in statistical tests commonly used to measure linguistic features, namely punctuation count and sentence length, repetitiveness, readability, as well as the top 10 most distinct words in each ethnic group.

Start with punctuation count and sentence length. Both metrics varied notably across ethnicities. We consider each pairwise combination of the eight ethnicity groups, resulting in 28 unique combinations (e.g. Anglo-American compared to East-Asian, Anglo-American compared to African, etc.). 24 out of these 28 cases showed notable dissimilarities in punctuation count, and 14 out of 28 cases displayed variance in sentence length. For instance, the ethnicity that uses the most punctuation is European (with 1.6 punctuations per sentence), whereas the ethnic group that uses the least punctuation is Arabic (with only 1.3 punctuations per

sentence). The ethnicity that uses the longest sentences is East Asian (on average 13.1 words per sentence), compared to less wordy Arabs (which use only 12.3 words per sentence).

Secondly, repetitiveness, measured by bi-gram overlap, also exhibited considerable variations across ethnicities, with 17 out of 28 pairwise comparisons indicating statistically significant differences. Bi-gram overlap refers to the repetition of two consecutive words and is used as a measure of redundancy and predictability of language. With a Bi-gram overlap of 0.13, Arabs are found to have the highest repetitiveness. On the other side of the spectrum, East Asians are the least repetitive (0.10).

Thirdly, readability metrics showed distinct differences among ethnicity groups. Readability was quantified using the SMOG index score, which is a formula based on the number of polysyllabic words (words with three or more syllables) in a set number of sentences [67]. Comparing these SMOG indices between the ethnic groups indicated highly significant contrasts. For example, the Anglo-American and East Asian comparison displayed a remarkably low p-value of 2.79e−14. Apart from being highly significant, the actual SMOG values also present a large spread amongst the ethnic groups: With a SMOG index of 1.71, Anglo-Americans were ranked to be the most 'readable', compared to 'least-easy-to-read' East Asians with a SMOG index of 2.24.

Lastly, the top 10 distinct words. To establish a word-frequency ranking for all the eight ethnicity-specific sub-corpora, our initial step is to remove stop-words. To be able to further distinguish between positive and negative contexts, we create separate dictionaries for terms found in 1-star and 5-star reviews. Afterward, we calculate the disparity in TF-IDF values between 1-star and 5-star reviews for each term, identifying the most prevailing terms within both sets of reviews. The results (see Table 1) reveal a noticeable variation in commonly used terms among different ethnicity groups. Furthermore, we observe that this variation is smaller in 5-star reviews (for which only three terms are uniquely used by only a single ethnicity) than in 1-star reviews (with 8 unique terms). For instance, some ethnicities use very specific vocabulary in negative reviews (e.g. 'whatsoever' in AAE, and 'unprofessional' in comments from South Asians), whereas the term 'great' consistently holds the top position in 5-star reviews across all ethnicities—as cultural studies let us to expect, also in Google Maps reviews a 'smile' might be the same in all languages, but complaints and curses are culturally distinct. Consequently, we can confirm inherent cross-cultural variations in language characteristics. This motivates the need to represent ethnic linguistic traits in AI models. Such 'ethnicity-conscious' approaches will be better equipped to understand the underlying sentiments of text-even of underrepresented ethnicities-and thus provide more accurate predictions (RQ1).

Now, we turn to assessing the performances of distinct sentiment analysis AIs. Table 2 shows the overall performance (in terms of Precision, Recall, and F1-Score) of all methods considered in our study. These are four conventional SA tools, namely Doc2Vec + RF, BERT + RF, BERT + SVM, and BiLSTM, as well as two SA tools custom-trained by us for this study. One of these custom-trained models is ethnicity-blind BERT, the other is our proposed ethnicity-conscious BERT. The

Table 1 Most distinctive 1-star review terms (top) and 5-star review terms (bottom) grouped by ethnicity and sorted by decreasing TF-IDF score

Anglo-American	European	African	Arabic	East Asian	Hispanic	South Asian	Scandinavian
Rude	Company	*Brookson*	Worst	Company	Avoid	Rude	Avoid
Company	Worst	Company	Rude	Avoid	Company	Company	Rude
Avoid	Avoid	Customer	Bad	Rude	Terrible	Worst	Company
Terrible	Rude	Poor	Company	Poor	Poor	Avoid	Worst
Poor	Money	*Asking*	Poor	Told	Rude	*Unprofessional*	Phone
Told	Don	Rude	Terrible	Worst	Worst	Told	Poor
Don	*Did*	*Pay*	Order	Customer	*Just*	Terrible	Told
Customer	Order	Don	Don	Don	Don	Poor	Customer
Awful	Bad	*Ve*	Money	Terrible	Order	Don	Order
worst	terrible	*Whatsoever*	Awful	Phone	Bad	Awful	Terrible
Great	Great	Great	Great	Great	Great	Great	Great
Friendly	Friendly	Friendly	Good	Friendly	Good	Friendly	Friendly
Excellent	Good	Recommend	Friendly	Recommend	Friendly	Excellent	Good
Recommend	Professional	Excellent	Best	Good	Helpful	Good	Recommend
Highly	Excellent	Good	Highly	Helpful	Amazing	Helpful	*Fantastic*
Service	Amazing	Highly	Amazing	Excellent	Excellent	Highly	Highly
Helpful	Highly	*Staff*	Professional	Highly	Highly	Recommend	Excellent
Lovely	Helpful	Helpful	Excellent	Lovely	Recommend	Best	Lovely
Professional	Service	Professional	Recommend	*Really*	Best	Professional	Helpful
Good	Best	Best	Helpful	Amazing	Professional	Amazing	Professional

Unique words are marked in italics

Table 2 Summary experimental results (Precision, Recall, F1-Score) for all methods considered in our study (balanced data)

Method	Precision	Recall	F1-Score
Doc2Vec + RF	0.510	0.499	0.504
BERT + RF	0.594	0.595	0.594
BERT + SVM	0.538	0.541	0.539
BiLSTM	0.616	0.612	0.614
Ethnicity-Blind BERT	0.775	0.762	0.768
Ethnicity-Conscious BERT	**0.783**	**0.781**	**0.782**

Results are averaged across all star ratings

'blind' version is a fine-tuned BERT model with standard categorical cross-entropy. We argue that this is a strong baseline given that this model has been trained using a version of the target dataset that has been balanced by both ethnicity and star ratings. This approach is considered a best practice for bias mitigation since it facilitates the model in giving equal importance to all social groups [52].

Results show that the custom-trained models, i.e., ethnicity-blind and ethnicity-conscious BERT, significantly outperform all other conventional models. Specifically, Doc2Vec+RF only has an F1-Score performance of 0.504. Comparing other methods with this baseline, BERT+SVM improves Doc2Vec+RF results by 6.94%, whereas BERT+RF improves them by 17.86%. BiLSTM provides significantly better results, improving the baseline by 21.83%. The custom-trained ethnicity-blind BERT model outperforms all other conventional models. It improves the Doc2Vec+RF baseline by 52.38%. However, the proposed ethnicity-conscious model achieves the highest performance in terms of Precision (0.783), Recall (0.781), F1-Score (0.782), improving the Doc2Vec+RF baseline even further, by 55.16%. Comparing results in terms of accuracy (see Fig. 4) shows that our proposed ethnicity-conscious BERT approach outperforms the ethnicity-blind BERT as well as all other considered baselines, i.e., BiLSTM, BERT, and Doc2Vec (RQ2).

We now know that our proposed method achieves a competitive overall sentiment classification—but what about biases between ethnicity-specific sentiment classification? Therefore, we proceed to investigate if our proposed ethnicity-conscious BERT is capable of further reducing bias compared to the standard approach of balancing input data by demographic group, as done by ethnicity-blind BERT, as well as some other conventional models. In order to investigate this aspect, we zoom into the disaggregated performance by ethnicity group. Results in Table 3 show the performance in terms of Precision, Recall, and F1-Score grouped by ethnicity. Additional results in Fig. 3 show the accuracy of all methods, corresponding to the percentage of accurately predicted reviews by ethnicity. Overall, we observe that our proposed ethnicity-conscious BERT model has an accuracy rate of 80% across all eight ethnic groups. Differences in accuracy rates of ethnicity-conscious BERT between the ethnic groups are negligible (under 2 percentage points). This minor deviation makes it the least biased model out of all the considered SAs. Conventional SAs were found to classify some ethnicities' reviews with more than four percentage points higher accuracy than others (for instance, BERT + RF gets the reviews of

Table 3 Summary experimental results (Precision, Recall, F1-Score) for all methods considered in our study and for all ethnicities

Ethnicity-Blind BERT	Precision	Recall	F1-Score
African	0.784	0.772	0.778
Anglo-American	0.776	0.761	0.768
East Asian	0.781	0.766	0.773
European	0.778	0.766	0.772
Hispanic	0.774	0.761	0.767
Arabic	0.766	0.751	0.758
Scandinavian	0.775	0.761	0.768
South Asian	0.772	0.757	0.764
Ethnicity-Conscious BERT	Precision	Recall	F1-Score
African	0.792	0.789	0.791
Anglo-American	0.785	0.784	0.784
East Asian	0.790	0.788	0.789
European	0.789	0.789	0.789
Hispanic	0.780	0.779	0.779
Arabic	0.774	0.771	0.773
Scandinavian	0.783	0.781	0.782
South Asian	0.785	0.784	0.785

Results are averaged across all star ratings (balanced dataset)

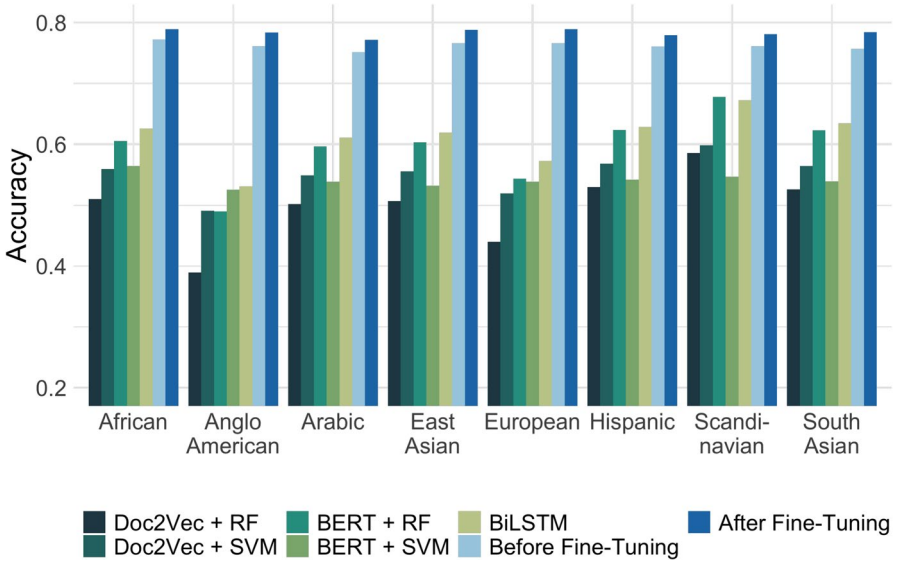

Fig. 3 Accuracy of all methods across all ethnicities (African, Anglo-American, East Asian, European, Hispanic, Arabic, Scandinavian, South Asian)

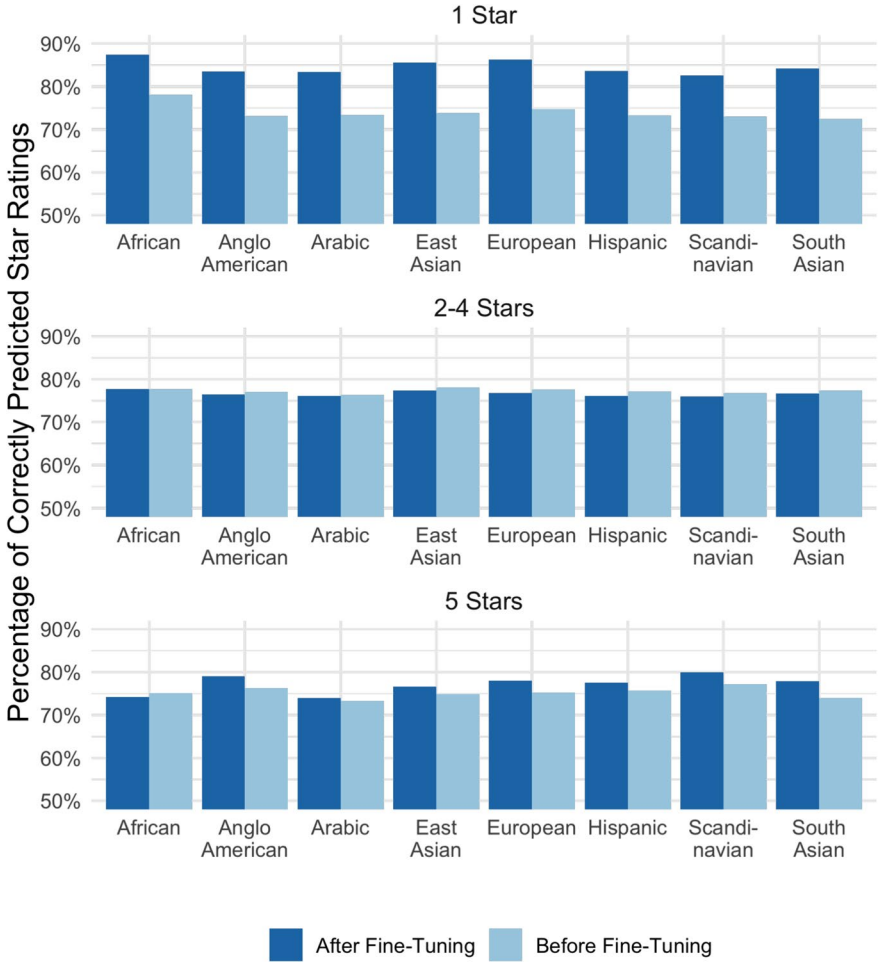

Fig. 4 Comparison between the Ethnicity-Blind BERT (before fine-tuning, light blue) and Ethnicity-Conscious BERT (after fine-tuning, dark blue) model accuracy by star rating across ethnicities

Scandinavians right in 65% of cases, but those of Hispanics only in 61% of cases). Furthermore, Fig. 3 shows that this increased fairness does not occur on the back of lower overall accuracy. On the contrary, the fine-tuned models that resulted from our ethnicity-conscious approach outperform ethnicity-blind models for all ethnicities. Specifically, the improvements over the ethnicity-blind model range from 2.22 percentage points (African) to 3.58 percentage points (South Asian). These results show that our ethnicity-conscious approach yields consistent results across all ethnicities, and does not penalize high-performing ethnicities to decrease overall performance disparities. Table 3 further highlights that, for all ethnicity groups, the ethnicity-conscious approach is also capable of outperforming ethnicity-blind models

in terms of Precision, Recall, and F1-Score, showcasing the effectiveness of our bias mitigation approach (RQ3).

After having discarded bias along ethnic lines, we now disaggregate accuracy by star ratings, to get a better picture of potential biases along the lines of negative vs. positive contexts. Results in Fig. 4 reveal that the ethnicity-conscious approach improves upon the ethnicity-blind approach dramatically for 1-star reviews, where the improvement is close to 10% for most ethnicities. For instance, the ethnicity-conscious approach correctly identified East Asians' reviews as 1-star in 81% of the cases, compared to the ethnicity-blind approach only getting their discontentment right in 74% of the cases. The performance also improves for 5-star reviews, for all ethnicities (except for African), albeit at a lower rate of 1–3%. For middle-range reviews, i.e. two to four stars, ethnicity-conscious BERT achieves similar accuracy rates to ethnicity-blind BERT. A more detailed view of model misclassifications through confusion matrices that show ethnicity-specific and star-rating-specific patterns is available in the appendix.

To illustrate this with concrete reviews, Table 4 compares reviews from different ethnicities whose star ratings were either over- or underpredicted. For example, a typical case is a 1-star review in SAE mistaken by ethnicity-blind BERT for a five-star review as it contains the word 'wow': 'Wow... What a Fawlty Towers Experience. Came with my elderly parents and my husband and once seated we were given one menu to share. I asked for two menus but this was refused. [...]' 'Wow' is generally positive, but in SAE it can also be used ironically to denote negative astonishment. Similarly, the following review, written by a European, is accompanied by a 1-star rating, but ethnicity-blind BERT predicted 5-stars: 'If your Dad Grandad and Great Grandad don't work here you won't last 5 min.' It is possible that in cultures placed on the individualist side by Hofstede (such as the US) such terms associated with the nuclear family prompted the model to deduce a positive sentiment, whereas in 'old-world' Europe they might as well be connotated with nepotism, such as in this review. As a last example, a review by an Arabic person ('[...] Always something is wrong with the order and a few times I called about it and answered bring it back') might have been misclassified due to faulty English. As also found by Zhiltsova et al. [11], this might be a further case indicating that non-native speakers might suffer from accuracy rates biased against them. In these three cases, ethnicity-conscious BERT correctly predicted the actual star ratings. Thus, we hypothesize that ethnicity-blind BERT is less susceptible to sarcasm, ethnicity-specific terminology, and errors in non-native language, which may be more common for some ethnic groups than others.

To conclude, these results align with the theory proposed by cultural studies. As cultural studies found in the 'real world' as well as the 'digital world' using qualitative ethnographic methods, we can confirm via the computerized methods of SA that the expression of strongly negative sentiments differs most across ethnicities. Consequently, ethnicity-blind SAs are not equipped to grasp these culture-dependent, context-specific nuances, resulting in the starkest biases in negative contexts. This reinforces the need for ethnicity-aware methods when examining data from heterogeneous social groups, especially in a context where accurately detecting strongly negative opinions is important (RQ4).

Table 4 Examples of reviews where Ethnicity-Blind BERT makes severe misclassifications while Ethnicity-Conscious BERT is accurate

Ethnicity	Overpredicted Predicted: 5 stars–Actual: 1 star	Underpredicted Predicted: 1 star–Actual: 5 stars
Anglo-American	Wow... What a Fawlty Towers Experience. Came with my elderly parents and my husband and once seated we were given one menu to share. I asked for two menus but this was refused. After an hour sat in full sun my husband had to get out	We purchased a beautiful car from GCS. I needed same day delivery the delivery team was working to capacity so Heba (sales) delivered it after work. She refused to take cash as a thank you. It was 8pm. She is amazing—excellent service fantastic cars. Thank you GCS
European	If your Dad Grandad and Great Grandad don't work here you won't last 5 min	Jigsaw pulled out all the stops for our company after we had to try to save an order with an incredibly difficult customer. They have a CAN DO attitude which is very rare in this industry
African	Anna and Krishna are really bad people. Lying and leaving our bathroom not complete after months of waiting. Beware	Have used Sharman Burgess to sell my late in laws bungalow. Hard to know where to begin in terms of writing a review. Of all estate agents contacted in the area they were the only ones that had a specific policy for Covid-19
Arabic	After ordering several times from Fernandez I have learned my lesson eventually owners have no clue about customer service as well. Always something is wrong with the order and a few times I called about it and answered bring it back	FEW ISSUES WITH THE CAR BUT JAMEY SORTED IT OUT EVERYONE NOW HAPPY THANKS JAMEY
EastAsian	So I wasn't going to write a written review about my poor experience with new lease but after just rating new lease I received a very unprofessional email asking "why have you just provided a Google review?"	Excellent service. Quick efficient and effective. Carpets look great—and no I have absolutely no connection with the business beyond being a customer
Hispanic	I've bought a phone from here and it worked only a month !!	My second review for Crane Bank Garage:The garage is brilliant ever helpful and willing to help and get you back on the road. Sarah on reception was as previously friendly organised and helpful. I was kept well informed of work
SouthAsian	Messed up my tariff the phones wouldn't work for data. They don't answer their landlines or its incorrect on google! Great advert for a telecoms company	I worked for caringcrew 2 years ago the training you receiveis so robust and they won't let you care until you complete your trainingBut if like eunica p you overdose a client they will deal with the issue even if it's a minor

Table 4 (continued)

Ethnicity	Overpredicted Predicted: 5 stars–Actual: 1 star	Underpredicted Predicted: 1 star–Actual: 5 stars
Scandinavian	Beware of prices of many items they put many items on SPECIAL OFFER but at till they simply scan and take full amount. The lady at till said to the man at the next tillthey need to type the offer amountthen only the offer price comes in	Good roads. No one dare to avoid rules and regulations. Strict rules are protect the valuable life of lots of people. Always roads are very clear

4.6 Open issues and limitations

There are multiple limitations of our proposed approach, which simultaneously open up paths for future research. One set of limitations arises from our reliance on personal names as a proxy for ethnicity. Such ethnicity detection might raise ethical concerns. One ethical school of thought decries that ethnicity itself is a potentially harmful social construct as it serves to 'single out' groups, which are put at risk of discrimination by the very process of labeling them as 'ethnic'. Thus, this school advocates an 'ethnicity blind' approach. In contrast, an 'ethnicity conscious' approach posits that without collecting information on ethnicity we can neither raise awareness of ethnicity-based injustices nor enable affirmative action to combat them [68]. The latter ethic of ethnicity consciousness permeates our research approach. While acknowledging the dangers around the construct of ethnicity, we think that our approach to detecting ethnicity is justified as it enabled us to uncover the inequalities in sentiment analysis' accuracy rates along ethnic lines, as well as to pursue affirmative action to make sentiment analysis work better for marginalized ethnic groups.

Another issue that comes along with name-to-ethnicity classification is that it will likely make some misclassifications (e.g., mistake a European name for an Anglo-American). However, this should not impact the overall patterns discovered, as long as errors are distributed equally across groups. To ensure the highest and most equal accuracy across groups, we used the name-to-ethnicity classifier N2E, which was specifically designed to ensure parity in error rates between ethnicity groups [63]. Thus, even if some reviews are misclassified, the overall patterns discovered in this research should still hold. Moreover, misclassification likely leads to an underestimation of the effect of ethnicity in our results, as it dilutes the distinct linguistic features attributed to each group in the analysis. Therefore, it is reasonable to assume that the true underlying patterns might be even more significant than our findings suggest.

An additional limitation is that name-to-ethnicity classification necessitated us to pre-define distinct labels (i.e., African, Anglo-American, Arabic,...), which might obscure within-group variation. Some sub-groups within our ethnicity categories might, for example, use a dialect or vocabulary that is specifically difficult for SA tools to classify (e.g., within the label of 'African', Congolese comments might be more difficult to classify than South African ones). Again, this leads to an underestimation of the total effect of ethnicity-related linguistic features on SA. Future research might work on more nuanced ethnicity labeling. This would enable a more accurate estimation of the effect of ethnicity, as well as a more targeted identification of groups that are specifically impacted.

A further set of limitations arises from our usage of Google Maps reviews. As our dataset consists of reviews of UK businesses written in English, our results do not necessarily generalize to other cultural or language contexts. Future research might extend the scope, for instance, by scraping Google Maps reviews from other countries. Such a multilingual and multinational perspective might discover further important patterns. Moreover, the dataset resulting from such an approach would likely be more balanced by ethnic group. In our UK-only dataset, we had to

significantly down-sample to ensure comparability between groups (i.e., to have only as many reviews from UK nationals as from the other groups). Thus, an extended scope might improve overall accuracy.

Furthermore, Google Maps comments are relatively homogeneous, as most are short, relate to a business, and use a five-star metric. Here, future research might test the generalizability of our approach on data from other online platforms using different types of reviews and sentiment scores, such as a three-class setup (positive, neutral, and negative), or a continuous sliding scale. While cultural studies suggest that similar patterns to those we observed should also persist in other scenarios in which negative sentiments are expressed by a diverse set of people, research on other platforms could confirm this.

Lastly, our approach is limited by using ethnicity as the only axis of differentiation. However, future research could benefit from an intersectional approach that also considers social factors such as gender or age, and how these interact with ethnicity in the context of SA. In many other instances of bias in machine learning, such a perspective was able to uncover that social factors can have intersecting effects on performance metrics. Such intersectional disparities were found, for example, in the realm of facial recognition. Here, gender and race intersected, so that technologies worked best on white men, and worst on black women [69]. It would be interesting to find out if SA also gets the sentiments of white men better than those of black women.

5 Conclusion

This paper has married cultural studies and computer science by introducing the concept of context-dependent cultural-linguistic differences to the field of SA. It has shown that between eight cultural groups the vocabulary used in positive (five-star) Google Maps reviews is relatively homogeneous, but distinct in negative (one-star) reviews. To handle this linguistic variance, the paper then proposed an ethnicity-conscious SA model. This novel SA model was trained on the corpus of Google Maps reviews, annotated with the reviewers' ethnicity via a name-to-ethnicity classifier. This enabled splitting the corpus to fine-tune ethnicity-specific models, which were then re-integrated in a post-processing layer. We demonstrated that this ethnicity-conscious SA improves overall model performance as well as fairness between the ethnic groups. More specifically, for positive (five-star) reviews, it improves accuracy rates for all but two ethnic groups by an average of 3 percentage points, for middle-range (two- to four-star) reviews it leaves accuracy relatively unaltered, and for negative (one-star) reviews it makes the greatest leap by improving accuracy by an average of 10 percentage points for all eight ethnicities. This evidence might guide all advertisers, politicians, and social scientists who care about negative sentiments brewing amongst their customers, voters, or research populations, so they can take timely action to address resentments. Using conventional SAs is likely to misclassify negative sentiments, especially amongst minoritized populations. And, with the greatest respect, SA should start to treat everyone's negative feelings equally.

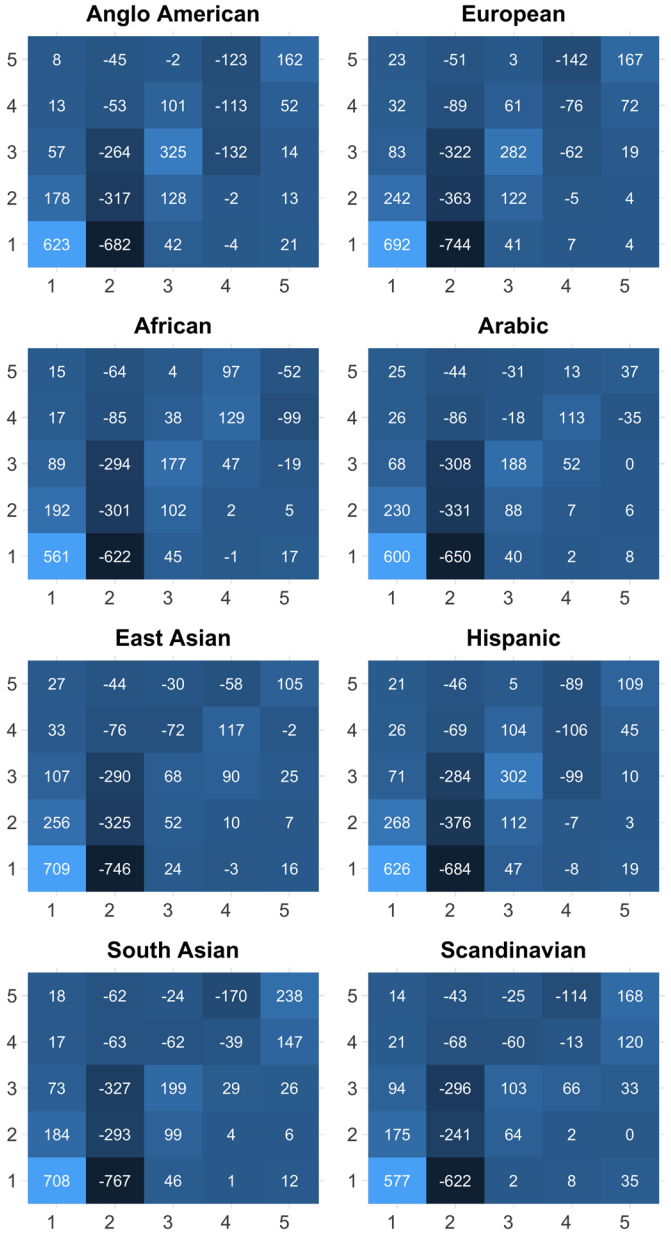

Fig. 5 Entry-wise difference between Ethnicity-Conscious BERT and Ethnicity-Blind BERT's confusion matrices. Rows indicate ground truth star ratings. Columns indicate model predictions. A positive entry indicates a higher value of the corresponding entry in the Ethnicity-Conscious BERT's confusion matrix than the Ethnicity-Blind BERT

Appendix

Focusing on confusion matrices in Fig. 5 allows us to zoom into a specific comparison between Ethnicity-Blind and Ethnicity-Conscious BERT. In this visualization, rows represent ground truth star ratings, whereas columns represent predicted star ratings. Each entry in the matrices is computed as the difference between the corresponding entries in the confusion matrices of Ethnicity-Conscious BERT and Ethnicity-Blind BERT, respectively. As a result, a positive value in a single entry indicates that a larger value was found in Ethnicity-Conscious BERT's confusion matrix. A positive entry on the main diagonal means that Ethnicity-Conscious BERT presented more correct predictions than Ethnicity-Blind BERT. While positive entries that lie close to the main diagonal show a small deviation between predictions made by Ethnicity-Conscious BERT and Ethnicity-Blind BERT, positive entries that lie further away from the main diagonal indicate a large deviation in predictions made by the two approaches. For example, an entry of 101 in row 4, column 3 for the Anglo-American ethnicity group indicates that 101 additional reviews were predicted with a 3-star rating for Ethnicity-Conscious BERT compared to Ethnicity-Blind BERT, which corresponds to 4-star ratings in the ground truth.

In general, our confusion matrices provide insights at a finer level of granularity, highlighting that small differences in accuracy can correspond to a significant difference in terms of the number of predicted reviews between the two approaches, depending on the size of the test split in our dataset. From a model calibration viewpoint, we observe that for all ethnicities Ethnicity-Conscious BERT predicts a larger number of reviews with a 1-star or 3-star rating than Ethnicity-Blind BERT, as shown by entries in columns 1 and 3 which systematically present large numbers. Conversely, Ethnicity-Conscious BERT predicts a lower number of reviews with 2-star and 4-star ratings compared to Ethnicity-Blind BERT, as shown by the negative entries in columns 2 and 4. Interestingly, the largest amount of misclassified reviews by Ethnicity-Blind BERT can be identified in 2-star reviews (-682 for Anglo American, -744 for European, -622 for African, -650 for Arabic, -746 for East Asian, -684 for Hispanic, -767 for South Asian, -622 for Scandinavian).

Differently than 1-star predictions, where column-wise entries are systematically positive, and 2-star predictions, where entries are systematically negative, 3, 4, and 5-star predictions present some differences across ethnicities. For instance, Hispanic, Anglo American, and European are similar, in that entries are mainly negative for 4-star ratings and positive for 3-star and 5-star ratings. On the other hand, African, Arabic, East Asian, South Asian, and Scandinavian predictions present a more mixed behavior, without a clear pattern. A small difference in misclassifications between Ethnicity-Blind and Ethnicity-Conscious BERT can be observed for 1 and 2-star reviews misclassified as 4 and 5 stars, as evident by the small entries in the bottom right corner of the confusion matrices. Notably, there is a larger difference between the two approaches in terms of misclassification

of 4 and 5-star reviews as 1 and 2-stars, as shown by the top left corner of the heatmaps.

Data Availability Our code implementation and models are publicly available at our Github repository: https://github.com/rcorizzo/debiased-sa-fine-tuning/. All data sources used for our experiments are public and disclosed in the paper. Reviews are available on Google Maps UK: https://www.google.com/maps, a list of companies registered in the UK to systematically access business pages on Google Maps is available on Companies House: https://www.gov.uk/government/organisations/companies-house. We can share the pre-processed version of the dataset upon reasonable request.

Declarations

Funding Not applicable.

Conflict of interest The authors declare that there are no financial or non-financial interests directly or indirectly related to the work submitted for publication.

Ethical approval Not applicable.

References

1. ANGLO-EU TRANSLATION GUIDE (2024). Retrieved 16 may 2024 from https://www.labourmobility.com/anglo-eu-translation-guide/
2. Hovy, D.: Demographic factors improve classification performance. In: Zong, C., & Strube, M. (2015). (eds.) In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing* (Volume 1: Long Papers), pp. 752–762. Association for Computational Linguistics, Beijing, China. Retrieved 30 Apr 2024 fromhttps://doi.org/10.3115/v1/P15-1073 . https://aclanthology.org/P15-1073
3. Li, J. (2024). Advances in Sentiment Analysis - Techniques, Applications, and Challenges. In: Advances in Sentiment Analysis—Techniques, Applications, and Challenges. IntechOpen. https://doi.org/10.5772/intechopen.111293. Retrieved 16 may 2024 from https://www.intechopen.com/chapters/undefined/chapters/87589
4. Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., & Kalai, A. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. Retrieved 16 may 2024 from arXiv. arXiv:1607.06520
5. Dixon, L., Li, J., Sorensen, J., Thain, N., & Vasserman, L. (2018) Measuring and Mitigating Unintended Bias in Text Classification. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, And Society, pp. 67–73. ACM, New Orleans LA USA. Retrieved 16 may 2024 from https://doi.org/10.1145/3278721.3278729 . https://dl.acm.org/doi/10.1145/3278721.3278729
6. Kiritchenko, S., & Mohammad, S.M. Examining gender and race bias in two hundred sentiment analysis systems. Retrieved 09 may 2022 from arxiv: 1805.04508

7. Park, J.H., Shin, J., & Fung, P. (2018). Reducing Gender Bias in Abusive Language Detection. Retrieved 16 may 2024 from arXiv. arXiv:1808.07231 [cs]. http://arxiv.org/abs/1808.07231
8. Zhang, G., Bai, B., Zhang, J., Bai, K., Zhu, C., & Zhao, T. (2020) Demographics Should Not Be the Reason of Toxicity: Mitigating Discrimination in Text Classifications with Instance Weighting. Retrieved 16 may 2024 from arXiv. arXiv:2004.14088 [cs, stat]. http://arxiv.org/abs/2004.14088
9. Goldfarb-Tarrant, S., Lopez, A., Blanco, R., Marcheggiani, D. (2023). Bias Beyond English: Counterfactual Tests for Bias in Sentiment Analysis in Four Languages. Retrieved 29 Apr 2024 from arXiv. arXiv:2305.11673 [cs]. http://arxiv.org/abs/2305.11673
10. Shen, J., Fratamico, L., Rahwan, I., & Rush, A.M. Darling or Babygirl? Investigating stylistic bias in sentiment analysis. Retrieved 16 may 2024 from https://www.semanticscholar.org/paper/Darling-or-Babygirl-Investigating-Stylistic-Bias-in-Shen-Fratamico/fea6c14d6769dffc04d18d344921536b14a04b61
11. Zhiltsova, A., Caton, S., & Mulwa, C. (2019). Mitigation of unintended biases against non-native english texts in sentiment analysis.
12. Liu, H., Jin, W., Karimi, H., Liu, Z., & Tang, J. (2021). The Authors Matter: Understanding and Mitigating Implicit Bias in Deep Text Classification. In: Zong, C., Xia, F., Li, W., Navigli, R. (eds.) Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pp. 74–85. Association for Computational Linguistics, Online. Retrieved 29 Apr 2024 from https://doi.org/10.18653/v1/2021.findings-acl.7 . https://aclanthology.org/2021.findings-acl.7
13. Wierzbicka, A. (1999). Emotions Across Languages and Cultures: Diversity and Universals, Digital printing [der ausg.] 1999 edn. Studies in emotion and social interaction Second series. Cambridge University Press [u.a.], Cambridge.
14. Shouse, E. (2005). Feeling emotion affect. *M/C Journal*. https://doi.org/10.5204/mcj.2443
15. Mesquita, B., Boiger, M., & De Leersnyder, J. (2016). The cultural construction of emotions. *Current Opinion in Psychology, 8*, 31–36. https://doi.org/10.1016/j.copsyc.2015.09.015
16. Schweikard, D.P., Schmid, H.B. (2013). Collective Intentionality, Fall 2021 edn. Metaphysics Research Lab, Stanford University, Stanford. https://plato.stanford.edu/archives/fall2021/entries/collective-intentionality/
17. Matsumoto, D.R. (1996). Unmasking Japan : Myths and Realities About the Emotions of the Japanese. Stanford, Calif. : Stanford University Press, Stanford. Retrieved 16 may 2024 from http://archive.org/details/unmaskingjapanmy0000mats
18. Hofstede, G. (1980). Culture's Consequences : International Differences in Work-related Values / Geert Hofstede. Cross-cultural research and methodology series ; v.5. Sage, Beverly Hills ; London Publication Title: Culture's consequences: International differences in work-related values.
19. Hofstede, G. (2017). Lokales Denken, Globales Handeln. Interkulturelle Zusammenarbeit und Globales Management. Retrieved 16 may 2024 from https://www.springerprofessional.de/lokales-denken-globales-handeln-interkulturelle-zusammenarbeit-u/4574168
20. Adair-Toteff, C. (2014). Max weber on confucianism versus protestantism. *Max Weber Stud, 14*(1), 79–96.
21. Bluth, B. J. (1984). The benefits and dilemmas of an international space station. *Acta Astronlogy, 11*(2), 149–53.
22. Tomi, L., Rossokha, K., & Hosein, J. (2002). The role of cross-cultural factors in long-duration international space missions: Lessons from the SFINCSS-99 study. *Space Technology (Oxford, England), 22*, 137–44.
23. Bargiela-Chiappini, F., & Nickerson, C. (2003). Intercultural Business Communication: A rich field of studies. *Journal of Intercultural Studies, 24*(1), 3–15. https://doi.org/10.1080/07256860305789. Accessed 2024-05-16.
24. Okoro, E. (2019). Intercultural communication competence in multinational competitiveness: Literature review and synthesis. *Journal of Business Economic Policy*. https://doi.org/10.30845/jbep.v6n2a1
25. Kobayashi, J., & Viswat, L. (2010). Cultural expectations in expressing disagreement: Differenceerzs between Japan and the United States.
26. Ting-Toomey, S. (1985). Toward a theory of conflict and culture. In: KIm, Y.Y., Gudykunst, W.B. (eds.) Communication, culture and organizational processes, pp. 71–86. Sage, Beverly Hills.
27. Hammer, M. R. (2005). The intercultural conflict style inventory: A conceptual framework and measure of intercultural conflict resolution approaches. *International Journal of Intercultural Relations, 29*(6), 675–695. https://doi.org/10.1016/j.ijintrel.2005.08.010. Accessed 2024-04-15.

28. Hall, E. (1976). Beyond Culture. Anchor Press/Doubleday, Garden City. Retrieved 16 may 2024 from https://monoskop.org/images/6/60/HallspsEdwardspsTspsBeyondspsCulture.pdf
29. Brown, P., & Levinson, S. C. (1987). *Politeness: Some Universals in Language Usage*. Cambridge University Press.
30. Panocová, R. (2020). *Theories of Intercultural Communication*. Košice: Univerzita Pavla Jozefa Šafárika v Košiciach.
31. Fang, H., Zhang, J., Bao, Y., & Zhu, Q. (2013). Towards effective online review systems in the Chinese context: A cross-cultural empirical study. *Electronic Commerce Research and Applications, 12*(3), 208–220. https://doi.org/10.1016/j.elerap.2013.03.001. Accessed 2024-04-15.
32. Feng, W., & Ren, W. (2020). Impoliteness in negative online consumer reviews: A cross-language and cross-sector comparison. *Intercultural Pragmatics, 17*(1), 1–25. https://doi.org/10.1515/ip-2020-0001. Accessed 2024-04-15.
33. Fong, J., & Burton, S. (2008). A cross-cultural comparison of electronic word-of-mouth and country-of-origin effects. *Journal of Business Research, 61*(3), 233–242. https://doi.org/10.1016/j.jbusres.2007.06.015. Accessed 2024-04-15.
34. Tsang, A., & Prendergast, G. (2009). Does culture affect evaluation expressions? A cross-cultural analysis of Chinese and American computer game reviews. *European Journal of Marketing, 43*(5/6), 686–707. https://doi.org/10.1108/03090560910947007. Publisher: Emerald Group Publishing Limited. Accessed 2024-04-15.
35. Cenni, I., & Goethals, P. (2017). Negative hotel reviews on TripAdvisor: A cross-linguistic analysis. *Discourse, Context & Media, 16*, 22–30. https://doi.org/10.1016/j.dcm.2017.01.004
36. Hong, Y., Huang, N., Burtch, G., & Li, C. (2016). Culture, Conformity, and Emotional Suppression in Online Reviews. Journal of the Association for Information Systems. https://doi.org/10.17705/1jais.00443
37. Brand, B. M., Kopplin, C. S., & Rausch, T. M. (2022). Cultural differences in processing online customer reviews: Holistic versus analytic thinkers. *Electronic Markets, 32*(3), 1039–1060. https://doi.org/10.1007/s12525-022-00543-1. Accessed 2024-04-15.
38. Ren, W. (2018). Mitigation in Chinese online consumer reviews. *Discourse, Context & Media, 26*, 5–12. https://doi.org/10.1016/j.dcm.2018.01.001
39. Vásquez, C. (2011). Complaints online: The case of TripAdvisor. *Journal of Pragmatics, 43*(6), 1707–1717. https://doi.org/10.1016/j.pragma.2010.11.007. Accessed 2024-05-16.
40. Ahmed, A., Agarwal, S., Kurniawan, I., Anantadjaya, S. P., & Krishnan, C. (2022). Business boosting through sentiment analysis using artificial intelligence approach. *International Journal of System Assurance Engineering and Management, 13*(Suppl 1), 699–709.
41. Smith, N.A. (2010). A mixture model of demographic lexical variation. Proceedings of NIPS Workshop on Machine Learning in Computational Social Science. Retrieved 16 may 2024
42. Georgiadou, E., Angelopoulos, S., & Drake, H. (2020). Big data analytics and international negotiations: Sentiment analysis of Brexit negotiating outcomes. *International Journal of Information Management, 51*, 102048. https://doi.org/10.1016/j.ijinfomgt.2019.102048
43. Rozado, D., & Gharbi, M. (2022). Using word embeddings to probe sentiment associations of politically loaded terms in news and opinion articles from news media outlets. *Journal of Computational Social Science, 5*(1), 427–448. https://doi.org/10.1007/s42001-021-00130-y
44. Balahur, A., Lloret, E., Boldrini, E., Montoyo, A., Palomar, M., & Martínez-Barco, P. (2009). Summarizing Threads in Blogs Using Opinion Polarity. In: Orasan, C., Hasler, L., Forascu, C. (eds.) Proceedings of the workshop on events in emerging text types, pp. 23–31. Association for Computational Linguistics, Borovets, Bulgaria. Retrieved 16 may 2024 from https://aclanthology.org/W09-4304
45. Bollen, J., Pepe, A., & Mao, H. (2009). Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. Retrieved 16 may 2024 from arXiv. arXiv:0911.1583 [cs]. https://doi.org/10.48550/arXiv.0911.1583. http://arxiv.org/abs/0911.1583
46. Mogilner, C., Kamvar, S. D., & Aaker, J. (2011). The shifting meaning of happiness. *Social Psychological and Personality Science, 2*(4), 395–402. https://doi.org/10.1177/1948550610393987
47. Saito, R., & Haruyama, S. (2023). Estimating time-series changes in social sentiment @Twitter in U.S. metropolises during the COVID-19 pandemic. *Journal of Computational Social Science., 6*(1), 359–388. https://doi.org/10.1007/s42001-022-00186-4
48. Gunter, B., Koteyko, N., & Atanasova, D. (2014). Sentiment analysis: A market-relevant and reliable measure of public feeling? *International Journal of Market Research, 56*(2), 231–247. https://doi.org/10.2501/IJMR-2014-014. Accessed 2024-04-29.

49. Birjali, M., Kasri, M., & Beni-Hssane, A. (2021). A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems, 226*, 107134. https://doi.org/10.1016/j.knosys.2021.107134
50. Zhao, X., & Wong, C.-W. (2023). Automated measures of sentiment via transformer- and lexicon-based sentiment analysis (TLSA). *Journal of Computational Social Science*. https://doi.org/10.1007/s42001-023-00233-8. Accessed 2024-04-17.
51. Han, X., Baldwin, T., & Cohn, T. (2022). Balancing out Bias: Achieving Fairness Through Balanced Training. Retrieved 16 may 2024 from arXiv. arXiv:2109.08253 [cs]. http://arxiv.org/abs/2109.08253
52. Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K.-W. Gender bias in coreference resolution: Evaluation and debiasing methods. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pp. 15–20. Association for Computational Linguistics. Retrieved 11 may 2022 from https://doi.org/10.18653/v1/N18-2003 . https://aclanthology.org/N18-2003
53. Thelwall, M. Gender bias in sentiment analysis **42**(1), 45–57 https://doi.org/10.1108/OIR-05-2017-0139 . Publisher: Emerald Publishing Limited. Accessed 2022-05-09
54. Wu, F., Du, M., Fan, C., Tang, R., Yang, Y., Mostafavi, A., & Hu, X. (2022). Understanding Social Biases Behind Location Names in Contextual Word Embedding Models. *IEEE Transactions on Computational Social Systems, 9*(2), 458–468. https://doi.org/10.1109/TCSS.2021.3106003. Accessed 2024-05-16.
55. Yang, Z., Jain, H., Shi, J., Asyrofi, M.H., & Lo, D. (2021). BiasHeal: On-the-Fly Black-Box Healing of Bias in Sentiment Analysis Systems. In: 2021 IEEE International Conference on Software Maintenance and Evolution (ICSME), pp. 644–648. IEEE, Luxembourg. https://doi.org/10.1109/ICSME52107.2021.00073. Retrieved 16 may 2024 from https://ieeexplore.ieee.org/document/9609175/
56. Mahabadi, R., Belinkov, Y., & Henderson, J. (2020). End-to-End Bias Mitigation by Modelling Biases in Corpora. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J. (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 8706–8716. Association for Computational Linguistics, Online. https://doi.org/10.18653/v1/2020.acl-main.769. Retrieved 16 may 2024 from https://aclanthology.org/2020.acl-main.769
57. Groenwold, S., Ou, L., Parekh, A., Honnavalli, S., Levy, S., Mirza, D., & Wang, W.Y. (2020). Investigating African-American Vernacular English in Transformer-Based Text Generation. In: Webber, B., Cohn, T., He, Y., Liu, Y. (eds.) Proceedings of the 2020 Conference on Empirical Methods in Natural anguage Processing (EMNLP), pp. 5877–5883. Association for Computational Linguistics, Online. https://doi.org/10.18653/v1/2020.emnlp-main.473. Retrieved 30 Dec 2024 from https://aclanthology.org/2020.emnlp-main.473
58. Hofmann, V., Kalluri, P. R., Jurafsky, D., & King, S. (2024). AI generates covertly racist decisions about people based on their dialect. *Nature, 633*(8028), 147–154. https://doi.org/10.1038/s41586-024-07856-5. Accessed 2024-12-30.
59. Resende, G.H., Nery, L.F., Benevenuto, F., Zannettou, S., & Figueiredo, F. (2024). A Comprehensive View of the Biases of Toxicity and Sentiment Analysis Methods Towards Utterances with African American English Expressions. arXiv. arXiv:2401.12720 [cs]. https://doi.org/10.48550/arXiv.2401.12720. Retrieved 30 Dec 2024 from http://arxiv.org/abs/2401.12720
60. Das, A.: PsychoSentiWordNet. In: Petrovic, S., Selfridge, E., Pitler, E., Osborne, M., & Solorio, T. (2011). (eds.) Proceedings of the ACL 2011 Student Session, pp. 52–57. Association for Computational Linguistics, Portland, OR, USA. Retrieved 20 Apr 2024 from https://aclanthology.org/P11-3010
61. Lin, F., Mao, S., Malfa, E.L., Hofmann, V., Wynter, A.d., Yao, J., Chen, S.-Q., Wooldridge, M., & Wei, F. (2024) One Language, Many Gaps: Evaluating Dialect Fairness and Robustness of Large Language Models in Reasoning Tasks. Retrieved 02 Jan 2025 from arXiv. arXiv:2410.11005 [cs]. https://doi.org/10.48550/arXiv.2410.11005. http://arxiv.org/abs/2410.11005
62. Wang, Y., Wang, X., Beutel, A., Prost, F., Chen, J., & Chi, E.H. (2021). Understanding and Improving Fairness-Accuracy Trade-offs in Multi-Task Learning. In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. KDD '21, pp. 1748–1757. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3447548.3467326. Retrieved 16 may 2024 from https://dl.acm.org/doi/10.1145/3447548.3467326
63. Hafner, L., Peifer, T.P., & Hafner, F.S. (2023). Equal accuracy for andrew and abubakar-detecting and mitigating bias in name-ethnicity classification algorithms. AI & society, 1–25

64. Liu, X., Zheng, Y., Du, Z., Ding, M., Qian, Y., Yang, Z., & Tang, J. (2023). Gpt understands, too. AI Open.
65. Jo, J.-y., & Myaeng, S.-H. (2020). Roles and utilization of attention heads in transformer-based neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 3404–3417.
66. Bengio, Y. (2012). Practical recommendations for gradient-based training of deep architectures. In: Neural Networks: Tricks of the Trade: Second Edition, pp. 437–478. Springer.
67. Laughlin, G. H. M. (1969). Smog grading-a new readability formula. *Journal of Reading, 12*(8), 639–646.
68. Aspinall, P. J. (2009). The future of ethnicity classifications. *Journal of Ethnic and Migration Studies, 35*(9), 1417–1435. https://doi.org/10.1080/13691830903125901
69. Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Friedler, S.A., Wilson, C. (eds.) Proceedings of the 1st Conference on Fairness, Accountability and Transparency. Proceedings of Machine Learning Research, vol. 81, pp. 77–91. PMLR. https://proceedings.mlr.press/v81/buolamwini18a.html