STATISTICS IN TRANSITION new series, March 2026
Vol. XX, No. X, pp. xx–XX, https://doi.org/10.59139/stattrans-2026-xxx
Received – xx.xx.xxxx; accepted – xx.xx.xxxx

Is the GPT model suitable for sentiment analysis? Testing for geographical, political and gender bias

Agnieszka Choczyńska[1]

Abstract

The new generation of Large Language Models, based on Generative Pre-trained Transformers (GPT) can be useful for automatic text annotation and sentiment analysis. However, they tend to learn the bias from training data, which can lead to distorted results. In this paper, the GPT-4o-mini model by OpenAI is tested for the presence of geographical, political and gender bias in the case of Polish economic news headlines. It has been found that the model consistently differs in sentiment scores for the same sentence, depending on the country mentioned. A remedy to this problem is proposed, which masks the references to countries and nationalities using the GPT model. Some differences in sentiment scores resulting from explicit references to gender or political parties are also identified, although these types of bias are considerably weaker than geographical bias.

Key words: large language models, geographical bias, gender bias, political bias, sentiment analysis.

1. Introduction

Large Language Models (LLMs) based on Generative Pre-trained Transformers (GPT) are increasingly used in scientific research. Text annotation is one of the applications that can benefit from the GPT models (Kheiri and , 2023). On the one hand, they are faster and more affordable than human annotators. On the other, they are more versatile than the language models trained for a narrow purpose and do not require a training dataset (Kocoń *et al.*, 2023). Given that most of the information created by the human population is in natural language, having a universal, ready-to-use text-mining tool would be beneficial for in social science.

There are, however, some obstacles to overcome. LMMs tend to absorb the biases found in the texts on which they are trained (Rozado, 2020). A growing amount of research finds gender (Radaideh, Kwon and Radaideh, 2025; Lee *et al.*, 2024; Zhu, Wang and Liu, 2024), nationality (Manvi *et al.*, 2024; Aslan 2024), political (Retzlaff, 2024; Rozado, 2023), and other types of bias (Huang *et al.*, 2020) in the text generated by the GPT models. Other studies focus on word embeddings (vector representations of words created during model training) and find that models pick up stereotypical associations from the natural language (Garg *et al.*, 2018; Rozado, 2020).

[1] AGH University of Krakow, Poland. E-mail: aghachocz@agh.edu.pl. ORCID: https://orcid.org/0000-0001-7134-567X.

One could suspect that the bias learned by the generative AI also impacts their abilities in text annotation. For example, they would assign lower sentiment to the sentences mentioning a country or demographic group the training datasets were biased against. In their analysis of sentences related to the energy industry, Radaideh, Kwon and Radaideh (2025) found that the GPT-2 model assigned a lower score to the sentences mentioning nuclear energy, male gender, old age or conservative political ideology.

However, the topic is still understudied. Firstly, most of the existing literature focuses on the English language, while one of the benefits of the GPT models is their multilingualism. The studies uncovering algorithmic bias in GPT models typically analyze the word embeddings or the impact of bias on text generation. Despite sentiment analysis being a popular application for this kind of models, it is still not well known how the bias can distort its outcomes.

This paper analyzes the bias in GPT-based text annotations, focusing on the Polish language and the texts broadly related to economics. The issue is approach from a new angle, using a set of fictional economic news headlines with positive, negative, or neutral implications for the mentioned country. Headlines are generated with different country names each time and prompt the GPT-4o-mini model to assign the sentiment to the sentence. The results show that countries significantly impact the sentiment score (p-value < 0.001), even though the headline does not change.

The same framework is used to test if the GPT models exhibit political bias in sentiment analysis. A set of headlines mentioning political orientation (left-wing or right-wing party), power dynamics (ruling party or the opposition), and the names of the main political parties in Poland are generated. As a control, there are also provided headlines where no political party, position or orientation is mentioned.

The results show that the sentences without any mention of a party tend to have the highest sentiment score, though the differences are generally small. A positive sentence gets, on average, a higher sentiment score if it mentions the ruling party, but negative sentences about the ruling party get lower scores than if they mention the opposition. No bias was found with regard to political orientation or particular party names.

Similarly, the paper assesses the presence of gender bias in sentiment analysis. The generated sentences include either a) direct mention of gender (e.g. men, women, male, female), b) mention of a fictional male or female name, or c) gendered grammatical forms. In Polish, verbs, nouns and adjectives have gendered forms, so it is impossible to change the gender of the subjects by just replacing names and pronouns.

The analysis shows very small effect of gender bias. Among the positive sentences, the ones mentioning female names received higher sentiment scores than the ones mentioning male names. Among negative sentences, the model assigned slightly lower scores to the ones that mentioned the female gender. The successes of particular women may be perceived more positively, as they are typically framed as a bigger breakthrough. On the other hand, if the problem is related to women (negative sentences with direct mention of gender), it is perceived as a bigger problem - and assigned a lower sentiment - than if it was related to men. However, no other configurations yielded significant differences. In particular, there is no evidence of gender grammatical forms impacting the sentiment score.

Finally, there is proposed a method of dealing with inherent geographical bias by censoring country names from the text. A dataset of economic news headlines from the public TV portal is used for this purpose. The GPT-4o-mini model is prompted to assign a sentiment score from -5 (strongly negative) to 5 (strongly positive) towards each country mentioned in the text. Next, there are create anonymized sentences by replacing all references to countries with codes and run sentiment analysis for that modified dataset.

The model consistently overstates the sentiment for Poland. For Germany and Russia, it tends to produce more neutral scores (e.g. less positive for positive news and less negative for negative news), while the references to the US receive more extreme values. However, the model's performance in country recognition and anonymization is unsatisfactory, leaving a room for improvement.

Although the researchers have been long aware of the problem of algorithmic bias, this paper expand the current knowledge by showing how it can impact the outcomes of sentiment analysis in a non-English language. Tests for geographical, gender and political bias reveal that they are all present, with the first one being by far the strongest. This study may be helpful for those looking to apply the GPT model to sentiment analysis, especially for the case of news analysis.

2. Literature review

2.1. Applications of the GPT models in sentiment analysis

With their natural language processing abilities, the GPT models could be used for sentiment analysis and other text-mining tasks. They are many times faster and more affordable than human annotators. Unlike specialized machine learning models, they can perform a wide variety of tasks on different forms of text. Some of the existing solutions are available through API, meaning that the researcher does not need to have the computational power or storage needed to train a large model.

These models can outperform untrained text annotators (Gilardi, Alizadeh and Kubli, 2023) and, in some cases, the state-of-the-art solutions (Kheiri and Karimi (2023); Fatouros *et al.*, 2023). However, their performance varies by task and dataset, and they should not be treated as a universally good solution (Curry, Baker and Brookes, 2024). Most research finds the GPT models to underperform, compared to the high-tuned, specialized language models (Kocoń *et al.*, 2023; Liyanage, Gokoni and Mago, 2024; Krugmann and Hartmann, 2024; Kristensen-McLachlan *et al.*, 2023).

However, there are a few caveats to this research. First, in the case of OpenAI, we do not know the full list of language corpora these models were trained on. If researchers perform the tests using publicly available datasets, the model may have already seen them, meaning that its performance on new data may be overestimated (Kocoń *et al.*, 2023; Ahuja *et al.*, 2023). Secondly, comparing a specialized model trained for a specific task with a GPT model in a zero-shot approach may underestimate the latter's abilities with additional training. It is still difficult to determine the extent of additional training this model would need to match the performance of a specialized solution, and if it would be indeed substantially smaller than preparing a model from scratch.

Finally, with the fast pace of AI development, it is difficult to assess their performance. Most of the aforementioned studies are not yet published, and the models they test will likely be obsolete before they do. Overall, researchers call for caution and additional validation before using GPT models for text annotation (Pangakis, Wolken and Fasching, 2023; Kristensen-McLachlan *et al.*, 2023; Ollion *et al.*, 2023; Curry, Baker and Brookes, 2024).

Most of the research on LLMs in text analysis is focused on the English language. A Common Crawl corpus, widely used in model training, has 45% of English texts, so one could expect models will be the most proficient in this language (Dac Lai *et al.* 2023). In a comprehensive evaluation of several LLMs (including GPT-3.5 and 4) in 70 languages, Ahuja *et al.*, (2023) noticed a worse performance for prompts written in non-English language. For the same reason, Etxaniz *et al.*, (2023) proposed translating the problem to English before analysis. Similar results were found by Dac Lai *et al.* (2023). However, there are some studies that counter these findings, not finding the benefits of English prompts (e.g. Debess, Simonsen and Einarsson, 2024). Both model and task-specific aspects may interfere with the results, although most of the research seems to find the benefit of English prompts.

This study focuses on bias in sentiment analysis and not on the accuracy of the model. Based on the research above, the authors decided to write the prompts in English and set the temperature parameter to 0.25, which gave the most consistent results in the previous experiments with the GPT model.

2.2. Bias in sentiment analysis

The GPT models are based on embeddings, which are vector representations learned from a large corpus of natural language. Closely related words in natural language should end up relatively close in the vector space (Garg *et al.*, 2018). If the corpus contains stereotypical associations between words, the model will likely incorporate that information and express human-like bias (Caliskan, Bryson and Narayanan, 2017). Moreover, the bias will not necessarily be diminished by more training, if the additional training dataset contains bias as well (Radaideh, Kwon and Radaideh, 2025).

Rozado (2020) found that most of the research on bias in word embedding models considered gender bias (93% of analyzed papers) and racial bias (54%). They performed a wider analysis of associations between positive words and terms related to gender, age, race, religiosity, affluence, and political orientation in several natural language corpora used in LLMs training. They found that positive words were more associated with women and femininity, youth, beauty, affluence and liberal political orientation. Typical African-American names and religiosity held negative associations, but the results for direct mentions of race or sexual orientations were mixed, with different directions, depending on the corpora. Garg *et al.*, 2018 found that associations between gender/race and certain occupations were correlated with the factual proportions of employees. The additional bias was largely explained by stereotypes held by the population.

The research on bias in the GPT models mainly focused on the stereotypical or toxic elements in the generated text. Huang *et al.*, (2020) asked the model to finish sentences with notions of different genders and occupations, finding that it can produce more negative outputs in certain contexts. A similar framework was used by Lee *et al.,* 2024 in the South Korean context and Zhu, Wang and Liu (2024) in Chinese, finding bias related to gender and nationality.

Manvi *et al.*, (2024) performed a study of geographical bias, prompting the model to rank the countries, according to objective facts (e.g. population density), objective facts uncorrelated with geographic position (e.g. solar flux), and subjective opinion (e.g. attractiveness of the citizens). They found that the model systematically underestimates or overestimates the ranks of objective facts, despite being able to provide precise numbers when prompted. There is also a bias against the regions of lower socioeconomic conditions in rankings of subjective opinions. The authors tested 5 models and the GPT-4 exhibited the lowest bias.

Another form of bias is of a political nature. When asked questions from the political compass test, the GPT model leaned towards liberalism (Retzlaff, 2024). These findings are supported in the analysis of word embeddings by Rozado, 2020, but the political bias is not as well studied as that related to gender, race or nationality.

To what extent the bias built in the word embeddings or present in the generated text would impact the sentiment analysis? This topic is not yet well studied. Radaideh, Kwon and Radaideh (2025) studied the impact of bias on the sentiment scores assigned by five LLMs, including the GPT-2 model, in the case of sentences related to the energy industry. They generated sentences in which they switched terms related to energy source, politics, gender, age and ethnicity, and prompted the models to assign sentiment scores to each configuration. They found that the mentions of nuclear energy, conservative ideology, male gender, old age and white race usually lowered the sentiment score, however, with some variations between models. A similar approach is applied in this analysis.

2.3. Bias mitigation

Strategies to mitigate bias include a) creating more fair and balanced datasets, b) fairness-aware model training, and c) algorithmic debiasing (Srinivasan *et al.*, 2024, Liu, 2025). The first strategy may be done by balancing the dataset to obtain equal number of observations for majority and minority group (Han, Baldwin and Cohn, 2022). In an unbalanced dataset, minority groups may be classified with a higher error, due to their limited representation in the dataset. Another strategy is to embed fairness into the training process, designing loss-function so that it takes the bias into account.

Specifically in LLMs, it is possible to mitigate biases by manipulating word embeddings. Zhao *et al.*, (2018) used this approach to neutralize the gender connotations of (by definition) gender-neutral occupations. For example, a word "nurse" may refer to any gender, but its vector representation appears closer to "female" in the embedding space, as historically most nurses were women. Zhao *et al.*, (2018) captured the distances between occupations and genders and used them as weights in training of a gender-neutral model. Ravfogel *et al.*, (2020) presented an Iterative Null-Space Projection, a method of removing certain properties from neural representations. Liang *et al.*, (2021) tested their approach on the embeddings of the GPT-2 model.

The first problem is that these methods require *a priori* knowledge of all underprivileged groups and biases. Utama, Moosavi and Gurevych (2020) proposed a framework in which the first "shallow" model is trained on a limited dataset to pick up existing stereotypes and biases, which are further used to down-weight biased observations, lowering their impact on the final model. This is based on the assumption that biases represent the most superficial knowledge, that would be learned first by the model presented with limited data. A similar approach was tested by Orgad and Belinkov (2023).

The second problem is that these methods require either access to the data or repeating the training process. In the case of large, pre-trained models this may not be feasible. An alternative approach was proposed by Liu (2021), who attempted to mitigate political bias. They obtained the hidden states from the GPT-2 model and transformed them so that a gender neutral embedding was of equal distance to the two options, in this case liberal and conservative. However, they noticed a trade of between fairness and fluency and accuracy (see also: Nadeem, Bethke and Reddy, 2021, Liang *et al.*, 2021).

2.4. Hypotheses development

In this analysis, three types of sentiment bias are considered: geographical, political and gender bias. As public TV covers international news, one should expect the bias against particular countries could make a big difference in sentiment. The first hypothesis is based on the results obtained by Manvi *et al.*, (2024) and Rozado (2020):

H1: The GPT model is biased against countries of low socio-economic status.

Economic and business news may often reference political parties as well when they report government economic policy, investments, state-owned companies or corruption affairs. If the training data corpora contain positive associations with liberal and progressive political ideology (Rozado, 2020), one could expect the second hypothesis to be true:

H2: The GPT model is biased against right-wing parties.

Finally, the study considers the aspect of gender bias. In Polish, most parts of the speech have gendered forms. Every time a news headline mentions a person or a group of people, their gender is revealed through grammar. If the GPT model associates female names or grammatical forms with more positive sentiment, it could distort the analysis. Hence the third hypothesis:

H3: The GPT model is biased against men.

Although researchers note other dimensions of bias, they are not likely to impact the sentiment analysis in this case. Poland is a rather racially homogenous country and mentions of race are not common in economic news articles. Due to the economic focus, headlines do not generally mention physical appearance, sexual orientation or disabilities of the subjects, so these aspects were ommited as well.

3. Testing the GPT models for text annotation

3.1. Data

The data used in this analysis are news headlines from the business section of the Polish public TV internet portal. The dataset spans from 2012-06-13 to 2024-09-13 and consists of 17,554 pieces. Each news piece is composed of a headline and a one- or two-sentence description, that introduces a longer video material.

The first part of this study uses generated headlines similar to these articles but constructed in a way suitable for bias testing. For geographical bias sentences have to mention exactly one country. They had to include one party or party members for the political test. In the case of the gender bias test, each sentence had to either directly mention gender or a fictional person of a specified gender. In the second part of the study, the original headlines are used to test how geographical bias impacted sentiment analysis in a real-life scenario.

Similarly to human annotators, the GPT model will not always return the same output for the same prompt. First, there is a check of replicability of the GPT text-annotation task results. A random sample of 1000 headlines is selected and the model is prompted to perform two text-mining tasks. The first is to assign a sentiment score from -5 (strongly negative) to 5 (strongly positive). The second one is to extract all countries mentioned in the text. Each analysis is performed 10 times in different sessions to assess the consistency of the results.

The results are fairly consistent. In 66.4% of cases, the model returned the same sentiment in each round, and only four times (0.4%) the difference was 3 points or more. For the country recognition task, the Jaccard similarity index was applied. For each pair of outputs, it counts the number of countries provided in both outputs (intersection of sets), divided by the overall number of countries that appeared in them (union of sets). The average score is 0.955.

All tasks are carried out with the GPT-4o-mini model by OpenAI using the Batches interface. Batches enable scheduling of a larger portion of API requests for asynchronous processing. As each text is to be analyzed independently, it is a suitable option for text-mining tasks.

3.2. Geographical bias test

The test uses 30 sentences (10 positive, 10 negative, and 10 neutral in sentiment). The sentences are similar to the news headlines in the TVP dataset, but constructed in a way that only one country is mentioned in a headline, and the country name is interchangeable. English translations of example headlines are provided below:

Positive: Prices no longer on the rise. Inflation in XXX falls quicker than expected.

Neutral: XXX is struggling with drought. The government is implementing special support programs for farmers.

Negative: In XXX, the problem of unemployment is getting worse. Every fifth adult is looking for a job.

The full set of headlines can be found in the repository (https://github.com/agachocz/SiT_GPT_bias_appendix.git). After generating the sentences, the XXX placeholder is replaced with one of the 194 country names in a correct grammatical form. Then the GPT-4o-mini model is prompted to assign the sentiment score to each sentence.

If the model's sentiment analysis abilities are not impaired by geographical bias, each sentence should be assigned the same score across countries. The sentiment analysis is repeated 10 times to test if potential differences in the outputs occur consistently.

Table 1 presents the test results: minimum, maximum and average score obtained for each group of sentences, along with the Kruskal-Wallis test statistic. Kruskal-Wallis test is a non-parametric alternative for ANOVA, more suitable for analyzing differences in rankings distributions. Under H_0 hypothesis, there are no statistically significant differences withing groups of sentences, therefore the model assigns the same sentiment score consistently, regardless of a country mentioned.

Table 1. The results of geographical bias test. Significance codes: * < 0.05, ** < 0.01, *** < 0.001

Group	Min score	Max score	Average score	Kruskal-Wallis test	Correlation with GDP PC
Positive	2	5	4.15	855.98***	-0.003
Neutral	-3	-3	0.534	602.36***	-0.055
Negative	-5	-1	-2.9	399.93***	-0.057
All	-5	5	0.593	77.209	

As presented in Table 1, the differences were significant in all sentiment categories, with the highest Kruskal-Wallis statistic for positive sentences. However, there is no significant effect for all categories taken together. One reason may be that the bias is not linear - some countries may score more positively in positive sentences and more negatively in negative.

This effect can be seen in Figure 1, presenting the average sentiment scores for all countries within categories. Note that the maps have independent colour scales so the differences are better visible. In positive sentences, Poland scores the highest among all countries, because the good news may seem the best for the Polish language users if it is related to their own country. However, among negative sentences, the sentiment for Poland is at the bottom of the scale, since the bad news may seem worse when they hit close to home.

Similarly, Russia scores the lowest in positive sentences and is about in the middle of the scale for negative ones. Since it is responsible for aggression on Ukraine near the Polish border, positive news about the Russian economy may be perceived more negatively, and negative ones - more positively.

To directly test Hypothesis 1, the correlation is computed between the average sentiment and the GDP per capita. The data on GDP expressed in purchasing power parity is sourced from the Worldometer database (Worldometer, 2024), excluding 14 countries for which there was no data. The correlation coefficients are presented in the last column of Table 1.

In all sentiment categories, correlations between sentiment scores and GDP per capita are small and insignificant. Contradictory to the hypothesis, the sentiment does not seem to depend on the economic development of the countries. Looking at Figure 1, it is clear that the sentiment scores are the lowest for the countries that may be perceived as a threat by Polish citizens. The most notable being Russia, but also North Korea, Afghanistan and China.

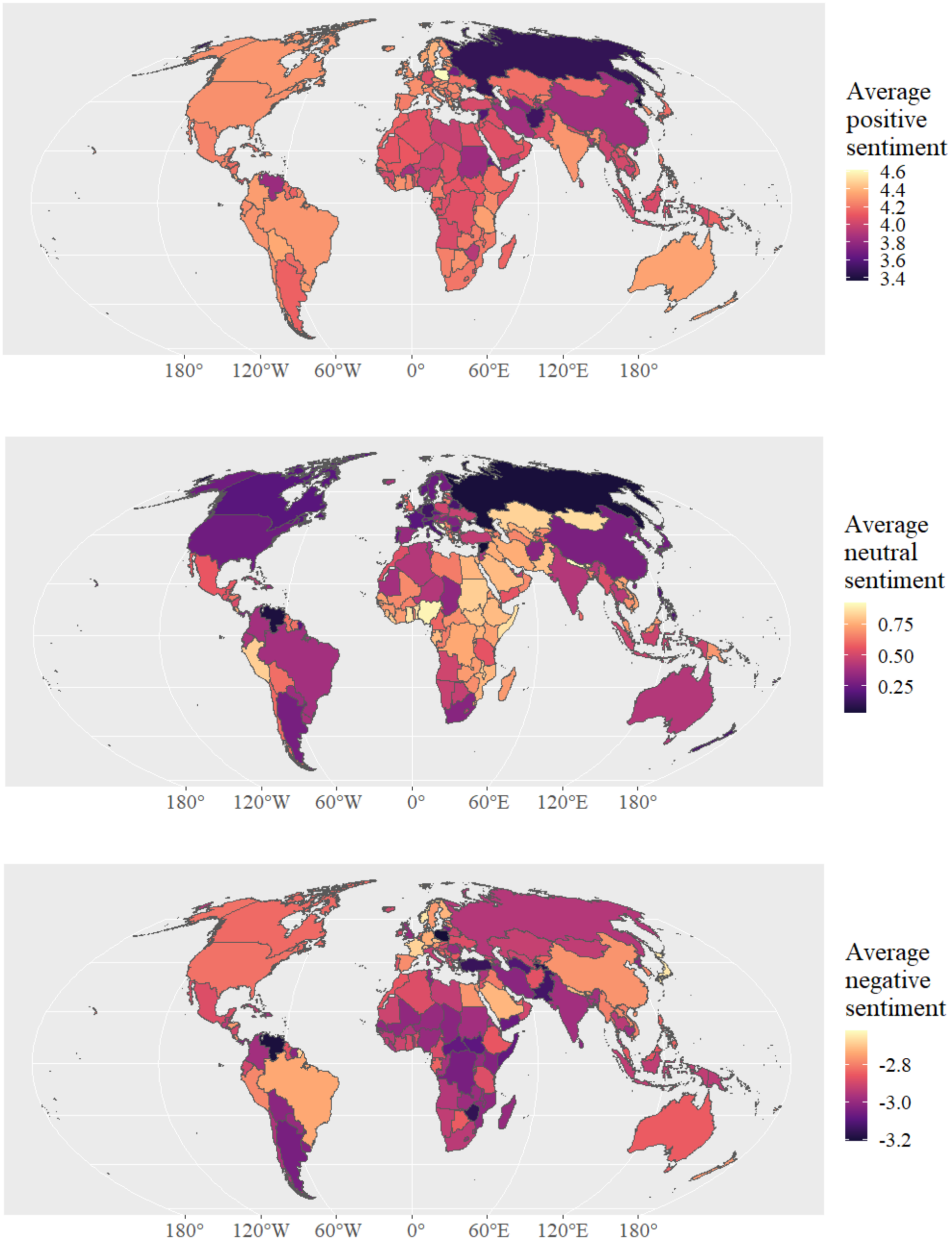

Figure 1. Average sentiment of positive, neutral, and negative sentences. Note that each map has an independent colour scale.

3.3. Political bias test

The same framework is used in the political bias test. Instead of country names, there are two terms for the party's position (ruling party, opposition), two terms for its political orientation (left-wing, right-wing), and eight names of the main political parties in Poland. The example sentences are as follows:

Positive: Renowned economist praises XXX's proposition. "This action is long overdue."

Neutral: XXX has published its new programme. What does it plan for seniors?

Negative: Millions in grants, zero results. The foundation linked to XXX MPs will come under scrutiny.

Additionally, there are generated the same sentences without any political terms, or with anonymous ones, such as "one of the parties" or "this party". This will provide the benchmark for the political terms. Again, the model is used to assign sentiment scores on a scale of -5 to 5 and repeat this task 10 times.

Table 2 provides the average scores for each term and the Kruskall-Wallis test statistics within groups. In general, any mention of politics lowers the sentiment of the sentence. Headlines without a term related to a party have the highest scores among positive sentences.

Table 2. The results of the political bias test. Significance codes: * < 0.05, ** < 0.01, *** < 0.001

Political term	Positive		Neutral		Negative	
	Average	Test	Average	Test	Average	Test
No term	3.470	-	0.340	-	-2.840	-
Position (ruling/opposition) + no term						
Ruling	3.32	11.889	0.31	0.067	-3.06	18.924
Opposition	2.97	**	0.38		-2.71	***
Orientation (left-wing/right-wing) + no term						
Left-wing	3.06	12.005	0.2	2.358	-2.78	0.944
Right-wing	2.98	**	0.26		-2.85	
Party names + no term						
KO	3.20	11.149	0.44	4.779	-2.780	6.901
Konfederacja	3.15		0.39		-2.770	
Nowa Lewica	3.20		0.40		-2.740	
PiS	3.26		0.21		-2.920	
PO	3.19		0.45		-2.810	
Polska2050	3.28		0.46		-2.770	
PSL	3.09		0.39		-2.780	

The Kruskal-Wallis test statistic is significant in the political orientation group for positive sentences. However, the pairwise Wilcox Rank Sum test reveals that only a difference between left/right orientation and no political term is significant. There is no difference between scores for left-wing and right-wing sentences, but they both score lower than sentences with no adjective related to political orientation. As there are no significant effects of bias toward specific party names, the Hypothesis 2 is rejected.

As for the position of the party in the political system, the ruling party tends to get higher scores in positive sentences, and lower for the negative ones, compared to the opposition. This may be a sign of more polarized opinions toward ruling parties, or assigning them a higher responsibility for the positive or negative outcomes.

3.4. Gender bias

Finally, the gender bias is tested. There are constructed sentences with mentions of gender, either explicitly (man/woman or male/female), through a fictional male or female name (Jan Nowak and Anna Kowalska), or just as a grammatical form (because nouns, verbs and adjectives have gendered alterations in Polish). The examples of headlines are provided below:

Positive: Infrastructure Minister Jan Nowak at the opening of the new power plant. "A milestone towards green energy."

Neutral: Polish men are innovative, but few of their discoveries live to see patents. Conclusions of a new CSO report.

Negative: Company founded by Jan Nowak in huge financial trouble. It is likely to declare bankruptcy.

The positive sentence above has a feminized version: «Infrastructure Minister **Anna Kowalska** at the opening of the new power plant. "A milestone towards green energy."» Then the name is removed to create an indirectly gendered sentence. Although in English this sentence would not imply the gender of the Minister, in Polish the word "Minister" would have two forms.

The example of a neutral sentence does not contain a name but a direct reference to the male gender, which can be switched to "women" for the female version. The indirect sentences are "Poles are innovative, but few of their discoveries live to see patents. Conclusions of a new CSO report.", where "Poles" has gendered form ("Polki" or "Polacy").

The results of the analysis are provided in the Table 3. In general, gendered grammatical forms do not differentiate sentiment scores. Among positive sentences, the ones including a female name received significantly higher scores than the ones including a male name. Possibly, the individual success of a woman is perceived as a bigger breakthrough (and more impressive by that) than the same success of a man. However, negative sentences mentioning the female gender received lower scores than the same sentences related to the male gender. This may be because problems seem more grave if they are related to women.

Overall, the results support Hypothesis 3, that the sentiment analysis with the GPT-4 model is biased against men. However, the effects are miniscule, compared to the scale of geographical bias, and the bias is related only to two specific cases.

Table 3. The results of gender bias test. Significance codes: * < 0.05, ** < 0.01, *** < 0.001

Type	Women avg	Men avg	Kruskall-Wallis test
		Positive	
Explicit gender	4	4	-
Gendered name	3.81	4.09	4.067*
Grammatical gender	4.05	3.90	2.620
		Neutral	
Explicit gender	0.1	-0.15	0.639
Gendered name	1.45	1.35	0.514
Grammatical gender	0.64	0.42	1.123
		Negative	
Explicit gender	-2.93	-2.73	4.248*
Gendered name	-3.16	-3.2	0.435
Grammatical gender	-3.15	-3.17	0.148

4. Anonymization

In the previous section, it was found that the GPT-4o-mini model is vulnerable to strong geographical bias. A method to remedy this problem is proposed by masking all references to country or nationality in the source text. This section is dedicated to a sentiment analysis of economic media headlines with and without masking and compare the results. The prompts used for particular tasks are provided in the repository (https://github.com/agachocz/SiT_GPT_bias_appendix.git).

First, the GPT model is prompted to modify the headlines by masking all countries or nationalities with codes: AAA, BBB, CCC, and so on. An example of this transformation could be:

Input: Hundreds of Estonian companies still trade with Russia. Estonian exports to Russia fell drastically after the latter invasion of Ukraine, but over 300 companies registered in Estonia kept trading with this country.

Output: Hundreds of AAA companies still trade with BBB. AAA exports to BBB fell drastically after the latter invasion of CCC, but over 300 companies registered in AAA kept trading with this country.

In the same prompt, the model is also asked to return the dictionary in the form "Country:code", in this case: "Estonia:AAA;Russia:BBB;Ukraine:CCC", to easily retrieve country names behind the codes after sentiment analysis.

Next, the model is asked to separately assign a sentiment score to the original and masked headlines. Finally, there is a comparison of the results to see to what extent the bias related to country names impacted the sentiment analysis. For this purpose, there were selected four countries with the biggest coverage. Poland has the highest number of mentions (5,733 news pieces), followed by Russia (1,399), Germany (1,235) and the US (1,017).

Table 4. Differences between sentiment assigned by the GPT model for sentences with and without anonymization. A positive average means that anonymized mentions receive higher scores than the ones revealing country names. N refers to the number of mentions

	Poland	Germany	Russia	US
n	3,403	865	1178	824
Correlation	0.854	0.768	0.709	0.773
		Positive		
n	2,363	348	375	454
average diff	-0.066	0.856	1.570	-1.040
std deviation	1.510	2.080	2.800	2.370
		Neutral		
n	348	164	171	139
average diff	-0.891	-0.445	0.474	0.583
std deviation	2.050	1.700	1.880	2.080
		Negative		
n	692	353	632	231
average diff	-1.240	-1.050	-0.231	0.342
std deviation	2.250	1.440	1.440	1.680

The correlation coefficients between sentiment scores for original and anonymized sentences are presented in Table 4. The scores are quite similar, as the correlation varies from 0.709 for Russia to 0.854 for Poland. The impact of bias related to country names is not very high.

The articles are split into sentiment categories based on the sentiment score from anonymized sentences: positive for scores higher than 2, negative for lower than -2, and neutral for scores between. Most mentions of Poland and the US were positive, while mentions of Germany and Russia were largely negative. Table 4 presents the difference between scores from anonymized and original sentences and present the averages and standard deviations.

In all categories, the averages for Poland are negative, meaning that mentions of Poland resulted in higher scores than thise that would have been obtained from the anonymized sentence. In the case of Russia and Germany, the average difference for positive sentences is positive, so mentions of these countries make the sentence seem less positive, compared to an anonymized sentence. However, for the negative sentences, there is the same pattern as for Poland, where the sentiment with country mention tends to be less negative than the sentiment without it. The reverse is true for the mentions of the United States. Here, the positive sentences typically receive higher scores, and negative sentences get even lower scores than the ones with masked country names.

Overall, the model seems to consistently overstate the sentiment, when Poland is mentioned. For Germany and Russia, it tends to produce more neutral scores (e.g. less positive for positive news and less negative for negative news), while the mentions of the US receive more extreme values.

However, the model does not handle these tasks perfectly. There are cases, where the outputs do not match, either because the model does not recognize all countries mentioned in the original sentence, makes mistakes in anonymizing, or fails to provide sentiment for all codes from masked sentences. The cases with multiple countries, nationality adjectives, or mentions of geographical regions, institutions, companies and other entities tend to produce mismatched outputs. These problems can be somewhat remedied by providing more precise prompts or cleaning the data afterwards, but the method still leaves room for improvement.

5. Conclusions

One of the drawbacks of automatic text annotations with Large Language Models is algorithmic bias. The models tend to learn the stereotypical associations present in human-created data used to train them and it may distort the results.

In the case of Polish economic news, this study tests for the presence of geographical, political and gender bias in sentiment scores assigned by the GPT-4o-mini model by OpenAI. The method uses generated sentences with interchangeable terms related to country names, political parties or gender, and prompt the model to assign sentiment scores on the scale from -5 (strongly negative) to 5 (strongly positive). If the model was unbiased, there should be no difference in sentiment scores.

The results show a significant effect of geographical bias. The GDP model tends to judge the same sentences as more or less positive, depending on the country mentioned. Contrary to other studies, sentiment bias was not correlated with the GDP of countries. For the positive sentences, the lowest scores were obtained for countries violating international law and human rights, such as Russia or North Korea.

Political and gender bias were not that strong. Mentions of Polish parties did not differentiate sentiment and references to both left-wing and right-wing political orientations resulted in lower sentiment than no mention at all. Sentiment related to the ruling party was more polarized, compared to the opposition. Positive sentences mentioning the former were more positive, and the negative ones - were more negative. Positive sentences were also judged slightly more positively if they mentioned a female name, but negative sentences received more negative scores when the female gender was mentioned.

A remedy to the problem of geographical bias is proposed by masking all countries mentioned in the text. For this purpose, a dataset of economic news headlines from the public TV portal is used. The GPT model is prompted to assign a sentiment score from -5 (strongly negative) to 5 (strongly positive) towards each country mentioned in the text. Next, the anonymized sentences are created by replacing all references to countries with codes and run sentiment analysis for that modified dataset.

The study finds that the impact of bias depends on the country. For Poland, the model consistently provided higher scores for original sentences, compared to the anonymized headlines. For Germany and Russia, it tended to give less positive scores for positive news and less negative for negative news. On the contrary, the references to the US received more extreme values than the same sentences with masked country mentions. However, the model's performance in country recognition and anonymization has a room for improvement.

From the practical point of view for the GPT model users, testing for social bias is a necessity. The model can exhibit stereotypical tendencies when assigning sentiment, and the types of bias may depend on the specific data and use case. Recognizing and mitigating such biases should be a standard procedure in any sentiment analysis using large language models. This study shows that anonymization of the inputs may be a simple solution to deal with geographical bias, although in a larger scale, it requires a more reliable model to remove geographical markers from the text.

These findings could be useful for social science researchers interested in using Large Language Models for text analysis, especially if the source text is in Polish. The problem of algorithmic bias, especially related to countries and nationality, significantly affects the outcomes of sentiment analysis. However, the analysis is limited to Polish and short news articles about the economy. There may be cases where gender or political bias plays a greater role, or where sentiment analysis is distorted in other kinds of bias that have not been tested for here.

Additional data

Tables with generated sentences and all prompts used in the analysis are available in the repository: https://github.com/agachocz/SiT_GPT_bias_appendix.git.

References

Ahuja, K., Diddee, H., Hada, R., Ochieng, M., Ramesh, K., Jain, P., Nambi, A., Ganu, T., Segal, S., Axmed, M., Bali, K. and Sitaram, S., (2023). MEGA: Multilingual Evaluation of Generative AI, in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Singapore, pp. 4232–4267. https://aka.ms/MEGA.

Aslan, F., (2024). *Bias assessment in Large Language Models*, PhD thesis, Tilburg University, Tilburg

Caliskan, A., Bryson, J. J. and Narayanan, A., (2017). Semantics derived automatically from language corpora contain human-like biases Science, 356, pp. 183–186. https://doi.org/10.1126/science.aal4230.

Curry, N., Baker, P. and Brookes, G., (2024). Generative AI for corpus approaches to discourse studies: A critical evaluation of ChatGPT, *Applied Corpus Linguistics*, 4(1). https://doi.org/10.1016/j.acorp.2023.100082.

Dac Lai, V., Trung Ngo, N., Pouran Ben Veyseh, A., Man, H., Dernoncourt, F., Bui, T. and Huu Nguyen, T., (2023). ChatGPT Beyond English: Towards a Comprehensive Evaluation of Large Language Models in Multilingual Learning, in 'Findings of the Association for Computational Linguistics: EMNLP 2023', *Association for Computational Linguistics, Singapore*, pp. 13171–13189 .

Debess, I. N., Simonsen, A. and Einarsson, H., (2024). Good or Bad News? Exploring GPT-4 for Sentiment Analysis for Faroese on a Public News Corpora, Technical report, *ELRA Language Resource Association*. https://huggingface.co/datasets/hafsteinn/.

Etxaniz, J., Azkune, G., Soroa, A., Lopez De Lacalle, O. and Artetxe, M., (2023). Do Multilingual Language Models Think Better in English?, in *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, vol. 2, Association for Computational Linguistics, Mexico City, pp. 550–564

Fatouros, G., Soldatos, J., Kouroumali, K., Makridis, G. and Kyriazis, D., (2023). Transforming sentiment analysis in the financial domain with ChatGPT, *Machine Learningwith Applications*, 14, 100508.

Garg, N., Schiebinger, L., Jurafsky, D. and Zou, J., (2018), Word embeddings quantify 100 years of gender and ethnic stereotypes, *Proceedings of the National Academy of Sciences of the United States of America*, 115(16), pp. 3635–3644.

Gilardi, F., Alizadeh, M. and Kubli, M., (2023). ChatGPT outperforms crowd workers for text-annotation tasks, *Proceedings of the National Academy of Sciences of the United States of America*, 120(30) .

Han, X., Baldwin, T., and Cohn, T., (2022). Balancing out Bias: Achieving Fairness Through Balanced Training. In Y. Goldberg, Z. Kozareva, and Y. Zhang (Eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, EMNLP 2022, pp. 11335–11350. https://doi.org/10.18653/v1/2022.emnlp-main.779.

Huang, P.-S., Zhang, H., Jiang, R., Stanforth, R., Welbl, J., Rae, J., Maini, V., Yogatama, D. and Kohli, P., (2020). Reducing Sentiment Bias in Language Models via Counterfactual Evaluation, *arXiv*. http://arxiv.org/abs/1911.03064.

Kheiri, K., Karimi, H., (2023). SentimentGPT: Exploiting GPT for Advanced Sentiment Analysis and its Departure from Current Machine Learning, *arXiv*. http://arxiv.org/abs/2307.10234.

Kocoń, J., Cichecki, I., Kaszyca, O., Kochanek, M., Szydło, D., Baran, J., Bielaniewicz, J., Gruza, M., Janz, A., Kanclerz, K., Kocoń, A., Koptyra, B., Mieleszczenko-Kowszewicz, W., Miłkowski, P., Oleksy, M., Piasecki, M., Radliński, , Wojtasik, K., Woźniak, S. and Kazienko, P., (2023). ChatGPT: Jack of all trades, master of none, *Information Fusion*, 99 101861. https://doi.org/10.1016/j.inffus.2023.101861.

Kristensen-McLachlan, R. D., Canavan, M., Kardos, M., Jacobsen, M. and Aarøe, L., (2023). Chatbots Are Not Reliable Text Annotators, *arXiv* . http://arxiv.org/abs/2311.05769.

Krugmann, J. O. and Hartmann, J., (2024). 'Sentiment Analysis in the Age of Generative AI', *Customer Needs and Solutions*, 11(1).

Lee, S., Kim, D., Jung, D., Park, C. and Lim, H., (2024). Exploring Inherent Biases in LLMs within Korean Social Context: A Comparative Analysis of ChatGPT and GPT-4, in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 4, pp. 93–104.

Liang, P. P., Wu, C., Morency, L.-P. and Salakhutdinov, R., (2021). Towards Understanding and Mitigating Social Biases in Language Models, *ICML*. https://arxiv.org/abs/2106.13219.

Liu, R., Jia, C., Wei, J., Xu, G., Wang, L. and Vosoughi, S., (2021). Mitigating Political Bias in Language Models Through Reinforced Calibration, in *The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21)*. https://doi.org/10.48550/arXiv.2104.14795.

Liu, Z., (2025). Cultural Bias in Large Language Models: A Comprehensive Analysis and Mitigation Strategies, *Journal of Transcultural Communication*, 3(2), pp. 224–244. https://www.degruyterbrill.com/document/doi/10.1515/jtc-2023-0019/html.

Liyanage, C. R., Gokani, R. and Mago, V., (2024). GPT-4 as an X data annotator: Unraveling its performance on a stance classification task, *PLoS ONE*, 19 .

Manvi, R., Khanna, S., Burke, M., Lobell, D. and Ermon, S., (2024), Large Language Models are Geographically Biased, in *Proceedings of the 41st International Conference on Machine Learning*, 1409, pp. 34654 - 34669.

Nadeem, M., Bethke, A. and Reddy, S., (2021). StereoSet: Measuring stereotypical bias in pretrained language models, in 'Proceedings ofthe 59th Annual Meeting ofthe Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing', *Association for Computational Lingusitics*, pp. 5356–5371.

Ollion, É, Shen, R., Macanovic, A. and Chatelain, A., (2023). ChatGPT for Text Annotation? Mind the Hype! https://doi.org/10.31235/osf.io/x58kn.

Orgad, H. and Belinkov, Y., (2023). BLIND: Bias Removal With No Demographics, in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, pp. 8801–8821. https://doi.org/10.18653/v1/2023.acl-long.490.

Pangakis, N., Wolken, S. and Fasching, N., (2023). Automated Annotation with Generative AI Requires Validation, *arXiv*. http://arxiv.org/abs/2306.00176.

Radaideh, M. I., Kwon, H. and Radaideh, M. I., (2025). Fairness and Social Bias Quantification in Large Language Models for Sentiment Analysis, Knowledge-based Systems 319, 113569. https://doi.org/10.1016/j.knosys.2025.113569.

Ravfogel, S., Elazar, Y., Gonen, H., Twiton, M. and Goldberg, Y., (2020). Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection, in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7237–7256.

Retzlaff, N., (2024). Political Biases of ChatGPT in Different Languages, Preprints.org, URL: www.preprints.org.

Rozado, D., (2020). Wide range screening of algorithmic bias in word embedding models using large sentiment lexicons reveals underreported bias types, *PLoS ONE*, 15(4).

Rozado, D., (2023). The Political Biases of ChatGPT, textitSocial Sciences, 12(3), 148. https://doi.org/10.3390/socsci12030148.

Srinivasan, N., Perumalsamy, K., Sridhar, K., Rajendran, G. and Kumar, A. A., (2024). Comprehensive Study on Bias In Large Language Models, *International Refereed Journal of Engineering and Science*, 13(2), pp. 77–82 .

Utama, P. A., Moosavi, N. S. and Gurevych, I., (2020). Towards Debiasing NLU Models from Unknown Biases, in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 7597–7610

Worldometer, (2024). Worldometer GDP per capita dataset. https://www.worldometers.info/gdp/gdp-per-capita/, accessed: 23.09.2024.

Zhao, J., Zhou, Y., Li, Z., Wang, W. and Chang, K.-W., (2018). Learning Gender-Neutral Word Embeddings, in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4847–4853. http://arxiv.org/abs/1809.01496.

Zhu, S., Wang, W. and Liu, Y., (2024). Quite Good, but Not Enough: Nationality Bias in Large Language Models – A Case Study of ChatGPT, *arXiv*. http://arxiv.org/abs/2405.06996.