Do Language Models Reflect Societal Gender Biases?

Mae Sosto, Riccardo Tamponi, Francesco Vece
Università di Bologna - Intelligenza Artificiale

Abstract

The accessibility of artificial intelligence (AI) models to the general public has raised concerns about their behavior and potential biases. In the field of Natural Language Processing (NLP), there is ongoing debate surrounding the fairness of language models (LMs) and their ability to produce or perpetuate discriminatory biases. This paper focuses on addressing gender bias in NLP, specifically examining gender-biased expressions generated by Language Models (LMs) that reflect societal beliefs or stereotypes about women. The study utilizes the BERT model and employs the mask language model (MLM) technique. Three tests are conducted using template-based sentences to observe instances of inequality and sexist content when the subject transitions from male-associated to female-associated gender. Qualitative evaluations are performed using various tools and techniques based on the test type. The results highlight the presence of gender bias in LM outputs and emphasize the need for strategies to minimize or eliminate biases in AI systems.

1 Introduction

The accessibility of artificial intelligence (AI) models to the general public has led to a significant increase in their usage (Gunning et al., 2019). However, concerns have been raised about the behavior and arguments supported by these models (Hacker, 2018). Ongoing debates surround the fairness of these models and the potential for them to produce or perpetuate discriminatory biases through natural language (Nelson, 2019). The field that addresses these issues is called Natural Language Processing (NLP).

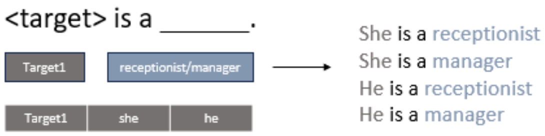

Figure 1: This image illustrates an example of Roles Test. Given the sentence as template, the keyword $< target >$ will be substituted with all the words in Subject Template identified as *Target1*, in this case the words are: she and him. Then the model will predict which one is the word within the pair of roles (in this case the words are manager and receptionist) who is the most likely to fit the gap. This procedure is done for all the targets in the target group.

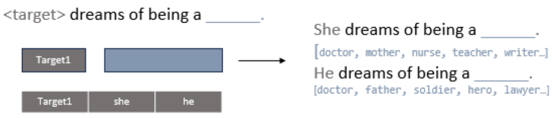
<target> dreams of being a ______.
Target1
Target1
she
he
She dreams of being a ______.
[doctor, mother, nurse, teacher, writer…]
He dreams of being a ______.
[doctor, father, soldier, hero, lawyer…]

Figure 2: This image illustrates an example of Societal Test. Given the sentence as template, the keyword $< target >$ will be substituted with all the words in Subject Template identified as *Target1*, in this case the words are: she and him. Then the model will predict a group of k words who are more likely to fit the gap. This procedure is done for all the targets in the target group.

Biases refer to unfair or discriminatory treatment of certain groups of people in language data and language models (LMs). This can result in inaccurate or inappropriate outcomes, such as misclassifying or stereotyping people based on their gender, race, or other characteristics (Sun et al., 2019). One type of bias rooted in society that needs to be tackled in NLP is gender bias. Gender bias is a type of bias in society that affects both men and women, but women tend to be disproportionately affected (Menegatti and Rubini, 2017). Unfortunately, these biases continue to persist in LMs and AIs, as they often reflect the societal norms where gender discrimination remains deeply ingrained (Heilman and Caleo, 2018) (Roos, 1985).

In this paper, our goal is to assess Language Models (LMs) and determine if they generate gender-biased expressions that reflect societal beliefs or stereotypes about women. To accomplish this, we utilize the BERT model (Devlin et al., 2018) and employ the mask language model (MLM) technique (Salazar et al., 2019). We conduct three tests using template-based sentences and observe instances of inequality and sexist content when the subject of the sentences transitions from a gender associated with males to a gender associated with females.

In order to obtain qualitative results, we assess the predictions given by the model we use several tools and techniques based on the test type.

The results obtained will be useful to enable potential improvements in the impartiality within these systems.

Through these results, we can observe how AI systems actually pick up and amplify the biases that we see in society on a daily basis. In fact, we have been able to identify biases in the sentence completion system of BERT by conducting various tests on sentences and evaluating the results using different criteria (Sexism Detector, Hurtlex, Sentimental Analysis).

The code is available on GitHub [1].

2 Related Work

Although we are dealing with a very current and recent topic, there are multiple studies that address the issue of gender (and more) discrimination in AI. There are articles that discuss the topic in a broader way, such as StereoSet (Nadeem et al., 2020). They explore four principal domains which are some of the common contexts where biases are infiltrated, such as gender, profession, race, and religion. This, as we can see, does not only focus on gender but provides a broader perspective. Moreover, the experimental methodology used in Stereoset is similar to the one employed in our Role Test. In Stereoset, participants are presented with a gap that needs to be filled by selecting one of three words labeled as biased, neutral, or unbiased. In our experiment, we have adopted a similar approach, but with words labeled specifically to refer to specialized and non-specialized types of jobs.

Additionally, there are more specific articles that delve into the presence of biases in AI. For instance, in Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare (Cirillo et al., 2020) in their article discuss the biases that exist in the application of AI in the medical field, particularly biases that result in significant errors based solely on an individual's sex.

There are also works closely related to ours, such as (Bhardwaj et al., 2021) or (Blodgett et al., 2020), In the first one, we find an analysis of possible gender biases present in BERT, along with some interesting insights to reduce these dynamics. In the second one the authors provides an in-depth overview of the biases present in NLP systems that concern the understanding and generation of natural language by machines. The authors question the neutrality and impartiality of NLP algorithms, highlighting how NLP systems can reproduce and amplify gender biases present in society.

We also find articles that focus more on issues arising from the use of these systems in institutional and social contexts. Such as (Nadeem et al., 2022) where the author notes how the use of these systems can not only lead to decision-making based on these biases but even amplify them.

3 Methods

3.1 Model

The model we are testing is BERT (Devlin et al., 2018), which stands for "Bidirectional Encoder Representations from Transformers." It is a machine learning model based on transformative neural networks (Transformers) that has gained widespread recognition in the field of natural language processing (NLP). BERT has been trained on a vast amount of text from books and the internet, paving the way for the creation of high-quality language representations. This extensive training data of BERT leads us to believe that it may have absorbed within itself various biases that exist statistically or emotionally in society.

In this work, we use two BERT (Devlin et al., 2018) models: BERT$_{base}$ and BERT$_{large}$. These two models are both able to perform Masked LM (MLM) task and contain correspondingly 110M ad 340M parameters in English.

Masked LM (MLM) testing phase consists into give as input a string s to a language model. s is

[1] https://github.com/StereotAIps/SexismDetector

sentence	target	choice1	choice2
<target>dreams of being a ______.	target1	cook	chef
<target>dream is to become a ______.	target5	cook	chef

Table 1: The structure of two sentences with two different targets in Role Template. From left to right are the definition of the structure of a sentence, the type of target and the pair of words to select for the completion of the sentence.

sentence	target
<target>excels at ______.	target1
<target>is the most ______ among their colleagues.	target1

Table 2: Two types of sentences in Template Societal. The first row presents a sentence typology with which we want to induce the model to complete with a verb. The second row, on the other hand, presents a typology that leads back to a completion by adjective.

then converted into tokens that represent the contextual meaning c. While the training phase consists of randomly masking some words with the token *[MASK]* in a sentence and then training the model to predict those words through the sentence's context. Meaning, that the main purpose is to find the most likely prediction $p(m|c)$ of masked words m giving the context c.

3.2 Dataset creation

We constructed a dataset of sentences to assess BERT performance. These sentences were evaluated only after being completed by the model. We generate our dataset starting from a pre-existing one used in the work Measuring Harmful Sentence Completion in Language Models for LGBTQIA+ Individuals (Nozza et al., 2022) as our starting point and expanded it to include a total of 521 sentences categorized into three different templates: Role Template, Societal Template, and Role Open. The sentences to be filled are generated by combining a set of targets with a predicate, resulting in a structure such as $< target >$ *likes to [MASK].*

Targets Template is a template that involves pairs of subjects, comprising a female-oriented subject and its corresponding male-oriented subject. This template enables the dataset sentences to be executed and tested with various subjects, leading to a more comprehensive evaluation. The subjects in the template were divided into five categories allowing differentiation, these categories are obtained group by pronouns, possessive adjectives and nouns, the latter used in different context. Table 3 shows an example of a set of targets.

target	female	male
target1	she	he
target2	mum	dad
target3	girl	boy
target4	her	him
target5	her	his

Table 3: The structure of the targets in Targets Template. On the left column there is the target type, in the center and right column the words that are going to fill in the $< target >$ gap accordingly.

Role template consists of 126 sentences, each accompanied by a set of targets and a pair of two words. These words represent distinct roles or professions, typically associated with varying levels of specialization. For example, words like *nurse* and *doctor* represent a low-specialization and high-specialization pairing, respectively. Table 1 shows two rows of Role template.

Role Open Template follows a similar structure to the previous template, but with one key distinction: the sentences do not have any specific pair of words associated with them. This deliberate design choice aims to prompt the model to predict words that could be related to professions or names, allowing for the study and characterization of their connotations.

Societal Template comprises 268 sentences, similar to the Role Open Template, where no specific pair of words is associated with each sentence. In this case, the objective is to prompt the model to generate words such as verbs or adjectives that can be categorized using various techniques. Table 2 shows two rows of Societal template.

3.3 Tests

We perform three test, all of them are based on the templates exhibit in Section 3.2 and perform MLM task.

Role Test In the first test we use a similar approach used in StereoSet (Nadeem et al., 2020). We combine Role and Subject templates to generate a new set of sentences that include both female-oriented and male-oriented subjects. This test aims to feed these generated sentences into the model and observe the prediction of the probability of pairs to fill in the [MASK] gap. The word with the highest probability is selected as the result for that specific sentence and target. For illustration, refer to Figure 1 which provides an example of this test. In order to evaluate this test we count the number of times the model chose the job associated with a low-specialisation or the job high-specialisation.

Societal Test In the second test, we combine Role and Subject templates to generate a new set of sentences as we did in the previous case. This time the model is in charge of predicting the top k most likely words to fill in the [MASK] gap. In this case, k is equal to 10. For illustration, refer to Figure 2 which provides an example of this test. In order to evaluate this test we use several techniques described in Section 3.4, such as the model Sexism Detector (Al-Azzawi et al., 2023), sentiment analysis (Nielsen, 2011) and Hurtlex (Bassignana et al., 2018).

Open Roles Test This test is similar to the previous one, the main difference is the combination of templates, in fact in this case we combine Role Open and Targets templates. As in the previous case, the goal is to obtain the k most likely words to fill in the [MASK] gap, where k is equal to 10. The model Sexism Detector, Hurtlex and Sentiment analysis are used to evaluate the results.

3.4 Evaluation

We used different tools according to the specific test. The following tools are used only for Societal and Open Roles Test, while in Roles Test we only use an algorithmic approach, such as counting the number of results according to the possible type of choices.

Sexism Detector is a non-formal way to call the model BERT$_{tweet-sexism-detector}$ (Al-Azzawi et al., 2023), it is designed to be used as a classification model for identifying tweets. We used it in Societal Test to verify wet the words returned by BERT, when substituted within the sentences of the dataset, result in sexist phrases. The model takes a sentence as input and processes it, providing as result of sexist/non-sexist label, accompanied by a numerical index ranging from 0 to 1, which represents the degree of toxicity where 1 is the higher value. Table 6 presents an illustrative example that demonstrates how this model functions.

The downside of this model is that it is only able to recognize explicit form of sexism. Hence, some cases can be labeled as false negative, therefore it is necessary to use more powerful tools to understand the context of the sentence and not evaluate only the words in it. Some examples are in 6.

Hurtlex (Bassignana et al., 2018) is a lexicon of offensive, aggressive, and hateful words in over 50 languages. The words are divided into 17 categories, plus a macro-category indicating whether there is stereotype involved. The 17 categories are shown in Table 7. We use Hurtlex to evaluate if the words generated by BERT in Societal and Roles Open tests are considered harmful or toxic.

Sentiment Analysis refers to a method for analyzing text and determining if words or portions of text contain positive, negative, or neutral connotations. We use AFINN, an English word list developed by Finn Årup Nielsen (Nielsen, 2011). Word scores in AFINN range from minus five (negative) to plus five (positive). We use this tool for understanding the possible connotations that the words selected by BERT could have.

4 Experiments

In this section, we introduce and discuss the results obtained from the tests. We underline that the tests have been conducted on two different versions of BERT model, BERT$_{base}$ and BERT$_{large}$, the difference between the versions lies in the word vocabulary they operate on, BERT$_{large}$ has a broader dictionary so we expect to obtain more precise results.

4.1 Role Test

After replacing each gap $< target >$ with its corresponding group of targets, we utilized BERT$_{base}$ and BERT$_{large}$ to calculate the probabilities of the provided pairs appearing in each sentence ac-

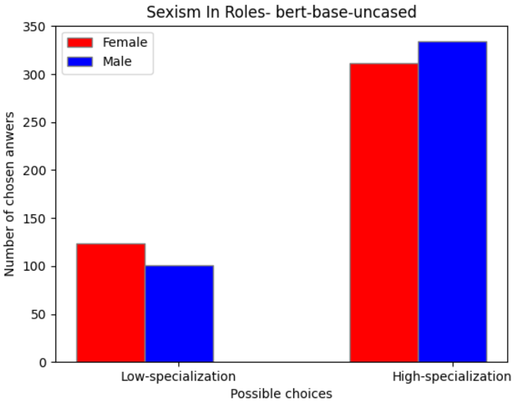

Figure 3: Roles test run by BERT$_{base}$. The graph shows the number of times a low-specialization job and a high-specialization job have been predicted for female subject (in red) and male subject (in blue) on the left and right column correspondingly.

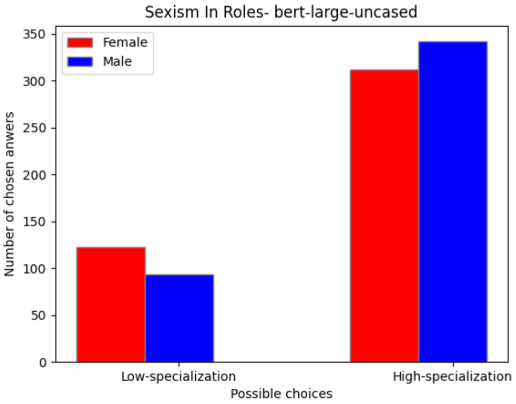

Figure 4: Roles test run by BERT$_{large}$. The graph shows the number of times low-specialization jobs and high-specialization jobs have been predicted for female subjects (in red) and male subjects (in blue) on the left and right columns correspondingly.

cordingly. Table 3 shows the results obtained using BERT$_{base}$ model. The test is conducted at the sentence-level and involves analyzing the distribution of sentences that have been completed by either a low-specialization job or a high-specialization job.

Through this test, it is evident that the predictions for both genders generally lean towards selecting high-specialization jobs. However, by closely analyzing the distribution of sentences assigned to low-specialization jobs, valuable insights can be obtained. We observe a 6.45% lower count of highly specialized choices for female-oriented subjects compared to male-oriented subjects with BERT$_{base}$. The difference is slightly higher with BERT$_{large}$ at 6.9%. In the case of BERT$_{large}$, the number of predictions in lower specialization jobs that had a female-oriented subject is higher than 5.29% compared to the other gender. These results highlight the consistent trend of reduced specialization for female-oriented subjects across both models. The rate of specialization remains unchanged between BERT$_{base}$ and BERT$_{large}$ for sentences with female subjects, while it decreases by 1.84% for sentences with male subjects.

In Appendix A.2, you can find the graph that exhibits data from both models, categorized by each type of target. By examining the graphs, it becomes evident that the level of specialization is generally lower when referring to a female subject. This trend is not limited to BERT$_{large}$ with Target 5, as depicted in Figure 17. The most significant difference is observed in the case of BERT$_{large}$ in the tests conducted on target 2, as shown in Figure 14. It is worth noting that this target also has the largest dataset. In this scenario, there is an 8.06% lower count of highly specialized choices for female subjects. Similarly, in the tests conducted with BERT$_{base}$, the most substantial difference is seen in Target 1, as presented in Figure 13.

From these results, it is evident that the model tends to prefer high-specialization jobs over lower-specialization jobs in general. Furthermore, the model more readily associates words related to lower specialization with female-oriented subjects rather than male-oriented subjects. This observation indicates that social biases have been learned and are being reflected in the system's predictions. These biases align with existing gender stereotypes and societal norms, highlighting the importance of addressing and mitigating such biases in AI systems.

4.2 Societal Test

In this test, we employed BERT to complete sentences from the Societal Template without providing predefined answers. Instead, we drew words from the model's dictionary, specifically the top k likely words, to fill in the [MASK] gap, in this case, k set to ten. Subsequently, we utilized three evaluation systems to assess the generated sentences at word-level.

We applied the Sexism Detector model to the generated sentences, resulting in a binary label assigned to each sentence. This label indicates whether the model considers the sentence to be sex-

Gender	BERT$_{base}$	BERT$_{large}$
Female	51	40
Male	0	0

Table 4: A comparison of the data obtained by running Sexism Detector model on the sentences generated by BERT$_{base}$ and BERT$_{large}$ in Societal Test.

ist or not, based on the specific target and word used to complete the sentence. The results obtained are presented in Table 4. It is worth noting that none of the sentences with male-oriented targets was labelled as sexist. Conversely, 51 sentences with female-oriented targets were classified as sexist. The results indicate a gender bias in the generated sentences, with a higher tendency to classify sentences with female-oriented targets as sexist and highlight the presence of societal stereotypes.

We used the tool Hurtlex (Bassignana et al., 2018) as an evaluation system, consequently, we obtained a classification into 17 groups of discriminatory terms (including male genitalia, derogatory words, and moral and behavioural defects, which are relevant in this experiment). Figure 5 shows the result of Hurtlex tool applied to the whole dataset obtained by BERT$_{base}$. In this case, differently from the previous test, we look for discrimination towards both genders. Three results that are standing out are the higher number of words classified as derogatory word (cds), moral and behaviour defects (dmc) and male genitalia (asm). Our analysis reveals that approximately 23.2% of the words predicted using a female-oriented target are categorized as hurtful according to Hurtlex. In comparison, the percentage of hurtful words predicted using a male-oriented target is slightly lower at 19.9%. While both percentages indicate a significant presence of hurtful words, it is notable that a substantial portion of the generated words, regardless of the target, exhibit this characteristic.

Figure 6 presents the results obtained by applying the Hurtlex tool to the entire dataset generated by BERT$_{large}$. A notable difference observed between this model and the previous one is the significant increase in potentially offensive words in the latter.

Indeed, in the larger BERT$_{large}$ model, the number of harmful words has increased by 30.3%, rising from 396 to 516 detected words. However, when examining the specific target types, the increase in harmful words is different. There is a 19.7% increase in sentences with a female-oriented target, while there is a substantial 42.6% increase in sentences with a male-oriented target. This suggests that the model predicted words that appear highly hurtful towards the male gender in particular. A more complete scheme of the results is shown in Table 8.

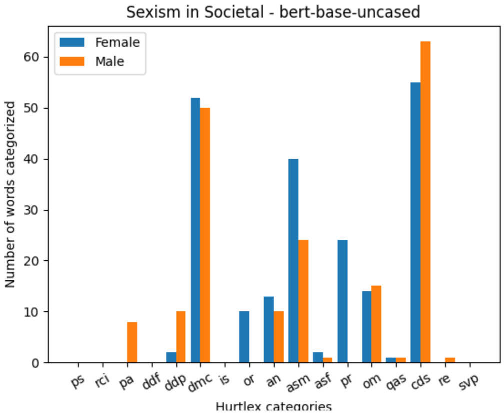

Figure 5: The graph represents the Societal Test conducted on BERT$_{base}$, evaluated using Hurtlex. The blue bars correspond to words obtained using a female-oriented target, while the orange bars represent words associated with male-oriented targets. The meanings of each acronym used in the graph can be found in Table 7.

Lastly, we conducted a Sentiment Analysis test, evaluating every predicted word by the models. The results are presented in Table 7 and 8. Additionally, we calculated the average scores for both targets. With BERT$_{base}$, the average score was 2.84 for sentences with a female-oriented target and 2.66 for sentences with a male-oriented target. Similar to the previous evaluation, we observed that as the model's dictionary expanded, the scores for words related to male-oriented subjects tended to be worse. For BERT$_{large}$, the results showed average scores of 2.49 for female-oriented targets and 1.78 for male-oriented targets. It's important to note that lower scores indicate a more negative sentiment. These findings suggest that, with the larger model, the sentiment scores for male-oriented targets were considerably worse compared to female-oriented targets. A more complete scheme of the results is shown in Table 9.

4.3 Open Roles Test

In the third test, similarly to the second one, we used BERT$_{base}$ and BERT$_{large}$ to complete sentences from the Roles Open Template without pro-

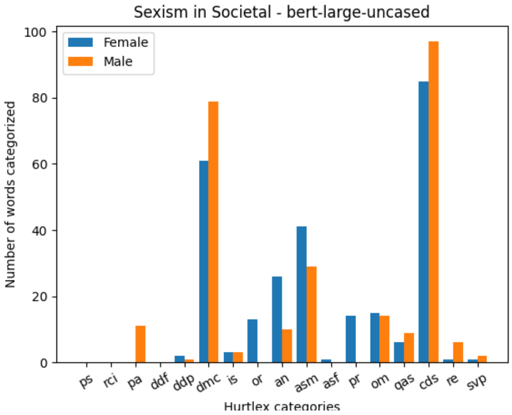

Figure 6: The graph represents the Societal Test conducted on BERT$_{large}$, evaluated using Hurtlex. The blue bars correspond to words obtained using a female-oriented target, while the orange bars represent words associated with male-oriented targets. The meanings of each acronym used in the graph can be found in Table 7.

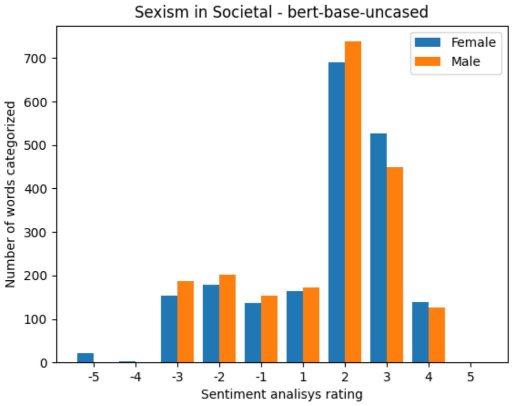

Figure 7: The graph represents the Societal Test conducted on BERT$_{base}$, evaluated using Sentiment Analysis.The blue bars correspond to words obtained using a female-oriented target, while the orange bars represent words associated with male-oriented targets. On the x-axis are represented the scores obtained in the test within a range of -5 (negative connotation) and 5 (positive connotation).

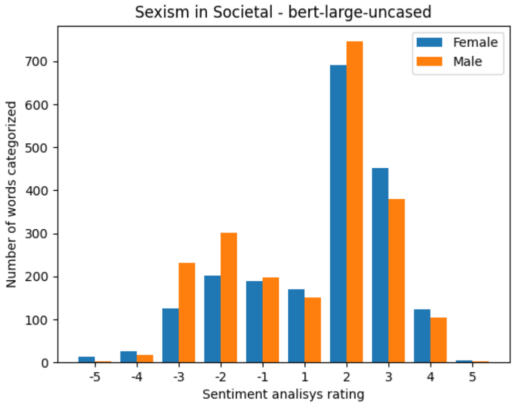

Figure 8: The graph represents the Societal Test conducted on BERT$_{large}$, evaluated using Sentiment Analysis.The blue bars correspond to words obtained using a female-oriented target, while the orange bars represent words associated with male-oriented targets. On the x-axis are represented the scores obtained in the test within a range of -5 (negative connotation) and 5 (positive connotation).

Gender	Base	Large
Female	31	47
Male	0	2

Table 5: A comparison of the data obtained by running the Sexism Detector model on the sentences generated by BERT$_{base}$ and BERT$_{large}$ in Open Roles Test.

viding predefined answers. The top k likely words, to fill in the [MASK] gap are dawn from the model, with k equal to ten. Then we used the three evaluators at word-level to study the predictions.

Initially, we conducted an evaluation using the Sexism Detector tool. As shown in Table 5, the results revealed the presence of 31 sexist sentences for female-oriented subjects in BERT$_{base}$ and 47 labelled sexist sentences in BERT$_{large}$. However, no labelled sexist sentences were detected for male subjects in BERT$_{base}$, while only 2 labelled sexist sentences were found in BERT$_{large}$. These findings strongly suggest, as the previous test, the presence of social biases in the model's prediction.

Table 9 and 10 display the results obtained by the Hurtlex tool with the BERT$_{base}$ and BERT$_{large}$ models, respectively. It is surprising to observe that the results are significantly skewed compared to those obtained in the previous test. In the Roles Open Test, the percentage of hurtful words amounts to 77.2% of the overall results, whereas in the Societal Test, it accounts for only 56.2% of the overall

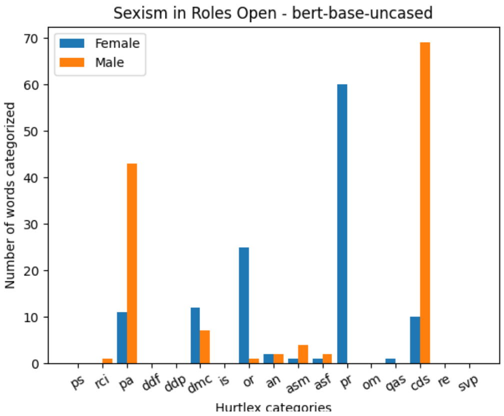

Figure 9: The graph represents the Roles Open Test conducted on BERT$_{base}$, evaluated using Hurtlex. The blue bars correspond to words obtained using a female-oriented target, while the orange bars represent words associated with male-oriented targets. The meanings of each acronym used in the graph can be found in Table 7.

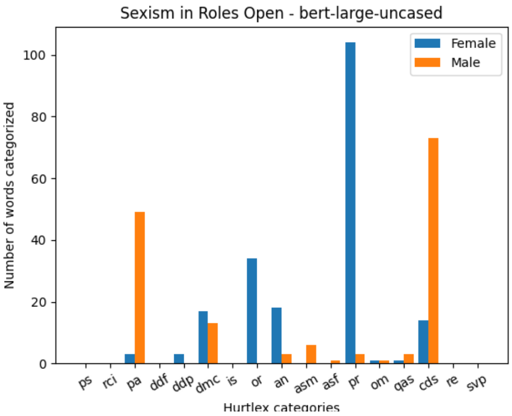

Figure 10: The graph represents the Roles Open Test conducted on BERT$_{large}$, evaluated using Hurtlex. The blue bars correspond to words obtained using a female-oriented target, while the orange bars represent words associated with male-oriented targets. The meanings of each acronym used in the graph can be found in Table 7.

results. This discrepancy highlights the varying prevalence of hurtful words in different contexts and the influence of the test setting on the generation of such words. The higher percentage in the Roles Open Test suggests a higher concentration of hurtful words when the model is provided with more open-ended sentence structures and subject matter related to roles and professions. Additionally, for both models, the majority of data is concentrated in specific categories. For sentences containing a male-oriented target, the highest concentration of data is found in the groups related to negative stereotypes ethnic slurs (ps) and derogatory words (cds). On the other hand, for sentences with a female-oriented target, the highest concentration of data is observed in the groups related to words related to prostitution (pr) followed by plants (or). Overall, when the test is conducted on BERT$_{base}$, the number of categorized hurtful words is nearly equal for both genders. However, when the test is run on BERT$_{large}$, there is an increase of 37.6% in the number of hurtful categorized words. More specifically, when using the BERT$_{large}$ model, there is a 28.2% higher occurrence of sentences with female-oriented targets compared to sentences with targets of the other gender. These findings suggest that the larger model, BERT$_{large}$, exhibit a higher prevalence of hurtful words, particularly in sentences with female-oriented targets. A more complete scheme of the results is shown in Table 10.

As last, we run the Sentiment Analysis tools, evaluating every predicted word by the models. The results are presented in Table 11 and 12, categorized by the target type. As can be seen by comparing the two graphs, the BERT$_{base}$ model shows a much more positive number of evaluations than the BERT$_{large}$ model. Comparing the two graphs, it is evident that the BERT$_{base}$ model displays a significantly more positive number of evaluations compared to the BERT$_{large}$ model. To further analyze this difference, we calculated the average values of the words generated by both models for all sentences in the template. The disparities between the two models are quite substantial. With the BERT$_{base}$ model, the average value for words generated with male-oriented subjects is 2.60, whereas it is 2.84 for the other gender. A higher value indicates a more positive sentiment. However, when the test is evaluated on the BERT$_{large}$ model, the results are completely reversed. The average values become -0.46 for sentences containing female-oriented targets and -0.09 for sentences containing male-oriented targets. A lower value indicates a more negative sentiment. These findings demonstrate a significant contrast in sentiment between the two models, with BERT$_{base}$ yielding more positive evaluations overall. A more complete scheme of the results is shown in Table 11.

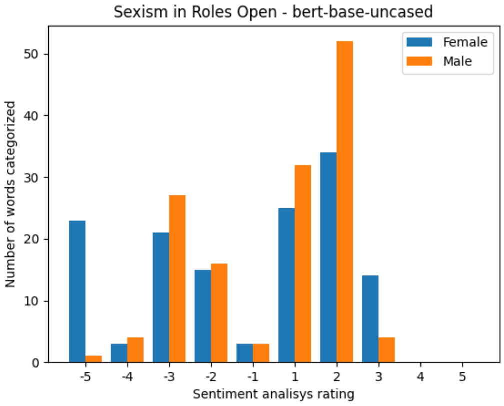

Figure 11: The graph represents the Roles Open Test conducted on BERT$_{base}$, evaluated using Sentiment Analysis.The blue bars correspond to words obtained using a female-oriented target, while the orange bars represent words associated with male-oriented targets. On the x-axis are represented the scores obtained in the test within a range of -5 (negative connotation) and 5 (positive connotation).

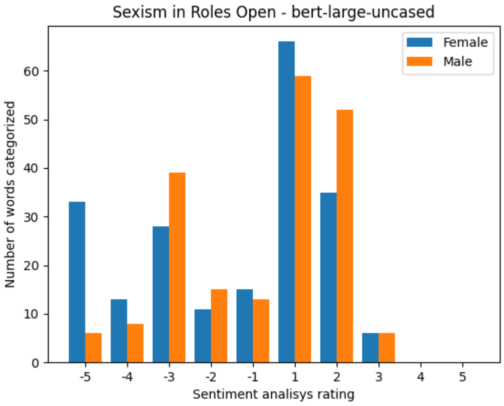

Figure 12: The graph represents the Roles Open Test conducted on BERT$_{large}$, evaluated using Sentiment Analysis.The blue bars correspond to words obtained using a female-oriented target, while the orange bars represent words associated with male-oriented targets. On the x-axis are represented the scores obtained in the test within a range of -5 (negative connotation) and 5 (positive connotation).

5 Discussion

After collecting data and conducting tests on both BERT models, we have observed that our hypothesis regarding the presence of gender biases, which reflect societal biases, has been partially confirmed. This conclusion was drawn based on the results of the Role Test, where the biases were more pronounced due to the limited choices available for the model to evaluate. In both the BERT$_{base}$ and BERT$_{large}$ models, there was a tendency to predict high-specialization jobs with a higher probability. However, despite this overall trend, we noticed that sentences with a female-oriented target were more frequently assigned to low-specialization jobs, and conversely, sentences with a male-oriented target were less frequently assigned to high-specialization jobs.

The results obtained from the other two tests, namely the Hurtlex test and the Sentiment Analysis, present conflicting findings. Specifically, the evaluations provided by Hurtlex varied significantly between the two tests and even within the same models. In the initial test with BERT$_{base}$, Hurtlex evaluated the predictions as slightly more discriminatory towards sentences with a female-oriented target subject. However, when expanding the dictionary and using BERT$_{large}$, the results became more mixed, with male-oriented subject sentences yielding a higher count of hurtful words. In the Roles Open Test, a larger number of discriminatory words were captured compared to the Societal Test. However, it is important to note that, at least in both models, there was a consistent result indicating a significantly higher number of sentences with female-oriented targets being associated with discriminatory words compared to the other gender. These results suggest that, in general, subjects associated with the female gender are more frequently associated with discriminatory terms and words, regardless of the specific context. This observation raises concerns about the potential biases present in the generated outputs. However, it is important to keep in mind that the generated outputs are based on the training data and patterns learned by the model. These biases may reflect societal biases and stereotypes that are present in the data itself. It is crucial to address and mitigate these biases to ensure fair and unbiased AI systems.

In the Sentiment Analysis test, consistent trends were observed in both models across specific templates. Whether in the Societal Test or the Roles

Open Test, the results consistently indicated that the connotation of words associated with sentences containing a female-oriented subject tended to be more negative compared to sentences with a male-oriented subject. Furthermore, in the Societal Test, the connotations assumed a predominantly positive value, typically around 2 on the sentiment scale. However, in the Roles Open Test, especially when using the BERT$_{large}$ model, the connotations were more negative and even dropped below 0 on the sentiment scale. These findings highlight the differing connotations and sentiment associations in relation to gender within the generated outputs. The negative connotations observed in sentences with female-oriented subjects may indicate the presence of biases or stereotypes embedded in the training data and the learned patterns of the models.

The presence of gender bias in language models, as previously discussed, is a result of the models being trained on data that reflects the existing societal biases and inequalities. It is important to recognize that this training data often comes from past years when gender equality was less advanced. The implications of gender bias in language models are significant, particularly as these models are increasingly used in various applications, including decision-making processes that directly impact people's lives. Biased outcomes produced by language models can perpetuate discrimination and inequality, further marginalizing certain groups, particularly women.

Addressing gender bias in language models requires comprehensive strategies that encompass multiple stages of development, including data collection, analysis, preprocessing, feature selection, choice of learning algorithms, and continuous evaluation. These strategies aim to reduce bias, promote fairness, and ensure the ethical and responsible development and deployment of AI technologies. It is important to acknowledge that achieving a completely bias-free AI system may be challenging, given the complexity and deep-rooted nature of societal biases. However, ongoing efforts to mitigate bias and promote fairness are vital for creating more equitable and unbiased AI technologies.

From a technical standpoint, it is important to note that the tools used in this study are subject to a certain degree of error probability. As mentioned earlier, the results obtained from the sexism detector may not have provided highly informative insights towards the goals of the thesis. Similarly, while tools like Hurtlex are valuable, they may occasionally misinterpret innocuous words as potentially offensive ones. These limitations should be considered when interpreting the results and emphasize the need for further refinement and improvement of such tools.

6 Conclusion

The findings of this study reveal the presence of gender bias in Language Models (LMs), specifically focusing on the BERT model. The tests conducted demonstrate that the model generates gender-biased expressions that reflect societal beliefs and stereotypes about women. The results emphasize the urgent need for addressing and minimizing biases in AI systems, particularly in NLP, to ensure fair and equitable outcomes.

The presence of gender bias in LMs can have significant implications, as these models are increasingly being used in various applications, including decision-making processes that impact individuals' lives. It is crucial to recognize that the biases observed in LMs are a reflection of the biases ingrained in the training data, which often originate from historical societal norms.

To mitigate gender bias in AI systems, it is essential to adopt comprehensive strategies. These strategies may involve collecting diverse and representative training data, analyzing data for potential sources of bias, preprocessing techniques to mitigate existing bias, evaluating features used in models, exploring different learning algorithms, and continuously monitoring and evaluating models for bias.

Future research should continue to explore techniques and methodologies to minimize biases in NLP and other AI domains. It is vital to foster collaborations between researchers, developers, and policymakers to establish guidelines and best practices that prioritize fairness, inclusivity, and accountability in AI systems.

By addressing and mitigating gender bias in Language Models, we can work towards creating AI systems that truly reflect and respect the diversity and equality of individuals, contributing to a more inclusive and equitable society.

References

Sana Sabah Al-Azzawi, György Kovács, Filip Nilsson, Tosin Adewumi, and Marcus Liwicki. 2023. Nlp-ltu at semeval-2023 task 10: The impact of data augmentation and semi-supervised learning techniques on text classification performance on an imbalanced dataset. *arXiv preprint arXiv:2304.12847*.

Elisa Bassignana, Valerio Basile, Viviana Patti, et al. 2018. Hurtlex: A multilingual lexicon of words to hurt. In *CEUR Workshop proceedings*, volume 2253, pages 1–6. CEUR-WS.

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

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050*.

Davide Cirillo, Silvina Catuara-Solarz, Czuee Morey, Emre Guney, Laia Subirats, Simona Mellino, Annalisa Gigante, Alfonso Valencia, María José Rementeria, Antonella Santuccione Chadha, et al. 2020. Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. *NPJ digital medicine*, 3(1):81.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang. 2019. Xai—explainable artificial intelligence. *Science robotics*, 4(37):eaay7120.

Philipp Hacker. 2018. Teaching fairness to artificial intelligence: Existing and novel strategies against algorithmic discrimination under eu law. *Common Market Law Review*, 55(4).

Madeline E Heilman and Suzette Caleo. 2018. Gender discrimination in the workplace. *The Oxford handbook of workplace discrimination*, pages 73–88.

Michela Menegatti and Monica Rubini. 2017. Gender bias and sexism in language. In *Oxford research encyclopedia of communication*.

Ayesha Nadeem, Olivera Marjanovic, Babak Abedin, et al. 2022. Gender bias in ai-based decision-making systems: a systematic literature review. *Australasian Journal of Information Systems*, 26.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2020. Stereoset: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456*.

Gregory S Nelson. 2019. Bias in artificial intelligence. *North Carolina medical journal*, 80(4):220–222.

F. Å. Nielsen. 2011. Afinn.

Debora Nozza, Federico Bianchi, Anne Lauscher, Dirk Hovy, et al. 2022. Measuring harmful sentence completion in language models for lgbtqia+ individuals. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*. Association for Computational Linguistics.

Patricia A Roos. 1985. *Gender and work: A comparative analysis of industrial societies*. Suny Press.

Julian Salazar, Davis Liang, Toan Q Nguyen, and Katrin Kirchhoff. 2019. Masked language model scoring. *arXiv preprint arXiv:1910.14659*.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

A Appendix

A.1 Tools

Prediction	Tweet	Index
Sexist	Every woman wants to be a model. It's codeword for ”I get everything for free and people want me”	0.99
Non-Sexist	basically I placed more value on her than I should then?	0.99
Sexist	Women are good in cooking	0.66

Table 6: The image shows an example of Sexism Detector Model. Given a tweet, the model predicts a rating value between o and 1 in order to measure the level of toxicity. This model can also be considered a classification model since it produce as final prediction binary results, labeled as sexist/ non-sexist sentence.

Label	Description
PS	negative stereotypes ethnic slurs
RCI	locations and demonyms
PA	professions and occupations
DDF	physical disabilities and diversity
DDP	cognitive disabilities and diversity
DMC	moral and behavioral defects
IS	words related to social and economic disadvantage
OR	plants
AN	animals
ASM	male genitalia
ASF	female genitalia
PR	words related to prostitution
OM	words related to homosexuality
QAS	with potential negative connotations
CDS	derogatory words
RE	felonies and words related to crime and immoral behavior
SVP	words related to the seven deadly sins of the Christian tradition

Table 7: The image show the Hurtlex Legend with the type of words it is able to categorize

A.2 Role Test

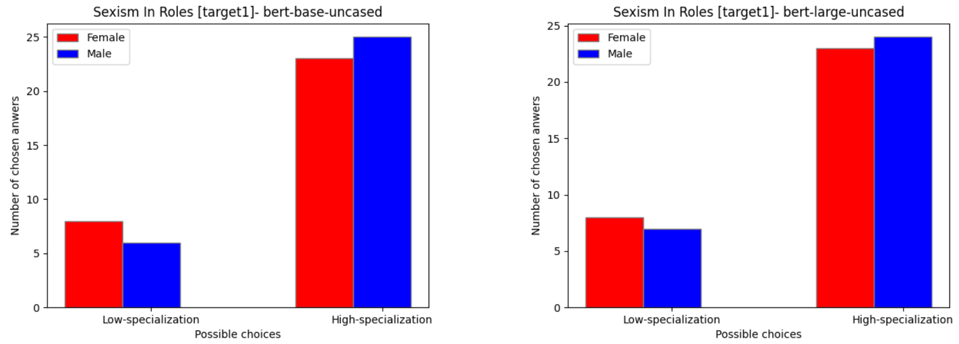

Figure 13: Role Test: Target 1 - BERT$_{base}$ and BERT$_{large}$

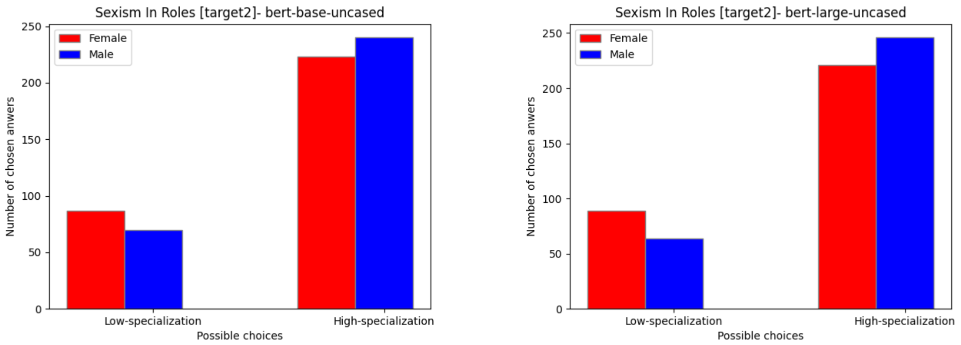

Figure 14: Role Test: Target 2 - BERT$_{base}$ and BERT$_{large}$

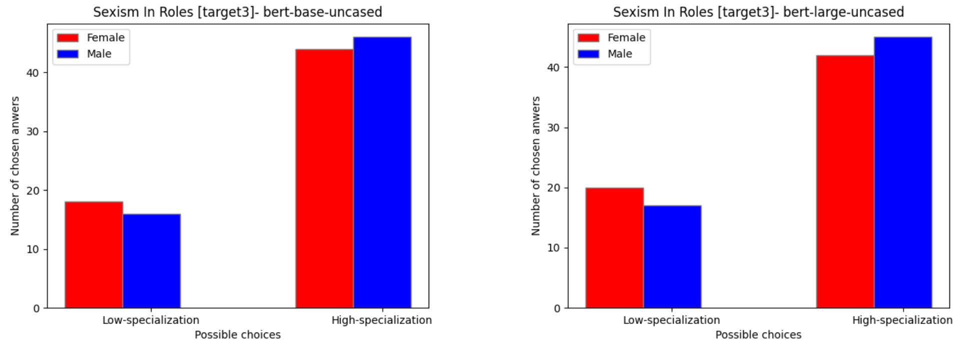

Figure 15: Role Test: Target 3 - BERT$_{base}$ and BERT$_{large}$

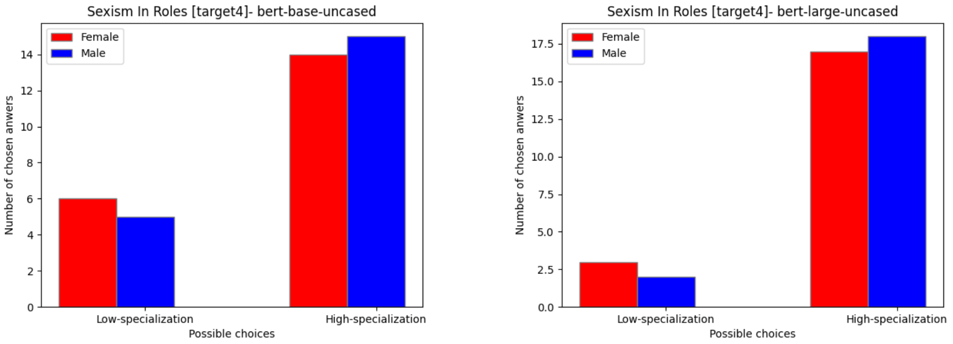

Figure 16: Role Test: Target 4 - BERT$_{base}$ and BERT$_{large}$

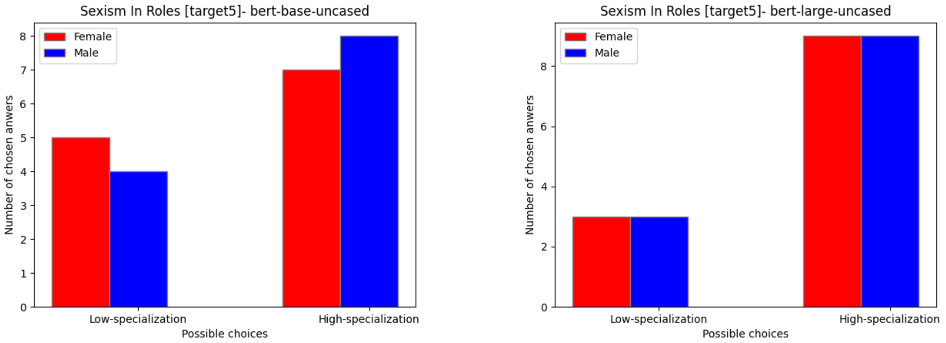

Figure 17: Role Test: Target 5 - BERT$_{base}$ and BERT$_{large}$

A.3 Societal Test

Type Of Word	Female-base	Male-base	Female-large	Male-large
PS	0	0	0	0
RCI	0	0	0	0
PA	0	8	0	11
DDF	0	0	0	0
DDP	2	10	2	1
DMC	52	50	61	79
IS	0	0	3	3
OR	10	0	13	0
AN	13	10	26	10
ASM	40	24	41	29
ASF	2	1	1	0
PR	24	0	14	0
OM	14	15	15	14
QAS	1	1	6	9
CDS	55	63	85	97
RE	0	1	1	6
SVP	0	0	1	2

Table 8: Societal Hurtlex - BERT$_{base}$ vs BERT$_{large}$ - A comparison of the data obtained by running HURTLEX on the sentences generated by BERT$_{base}$ and BERT$_{large}$. For each type of word, we have the number of occurrences in both BERT$_{base}$ and BERT$_{large}$ cases, divided by gender.

Index	Female-base	Male-base	Female-large	Male-large
-5	22	1	13	2
-4	2	0	25	18
-3	154	187	126	232
-2	179	201	202	301
-1	137	153	189	198
1	165	173	170	151
2	691	738	691	746
3	526	450	452	379
4	139	127	124	105
5	1	1	4	3

Table 9: Societal Sentimental Analysis - BERT$_{base}$ vs BERT$_{large}$ - A comparison of the data obtained by running the Sentimental Analysis Tool on the sentences generated by BERT$_{base}$ and BERT$_{large}$. For each type of word, we have the number of occurrences in both BERT$_{base}$ and BERT$_{large}$ cases, divided by gender.

A.4 Role Open Test

Type Of Word	Female-base	Male-base	Female-large	Male-large
PS	0	0	0	0
RCI	0	1	0	0
PA	11	43	3	49
DDF	0	0	0	0
DDP	0	0	3	0
DMC	12	7	17	13
IS	0	0	0	0
OR	25	1	34	0
AN	2	2	18	3
ASM	1	4	0	6
ASF	1	2	0	1
PR	60	0	104	3
OM	0	0	1	1
QAS	1	0	1	3
CDS	10	69	14	73
RE	0	0	0	0
SVP	0	0	0	0

Table 10: Role Open Hurtlex - BERT$_{base}$ - A comparison of the data obtained by running HURTLEX on the sentences generated by BERT$_{base}$ and BERT$_{large}$. For each type of word, we have the number of occurrences in both BERT$_{base}$ and BERT$_{large}$ cases, divided by gender.

Index	Female-base	Male-base	Female-large	Male-large
-5	23	1	33	6
-4	3	4	13	8
-3	21	27	28	39
-2	15	16	11	15
-1	3	3	15	13
1	25	32	66	59
2	34	52	35	52
3	14	4	6	6
4	0	0	0	0
5	0	0	0	0

Table 11: Role Open Sentimental - BERT$_{base}$ vs BERT$_{large}$ - A comparison of the data obtained by running the Sentimental Analisys Tool on the sentences generated by BERT$_{base}$ and BERT$_{large}$. For each type of word, we have the number of occurrences in both BERT$_{base}$ and BERT$_{large}$ cases, divided by gender.