HW2 report

Faezeh Pouya Mehr

April 12, 2023

Problem 2

1.

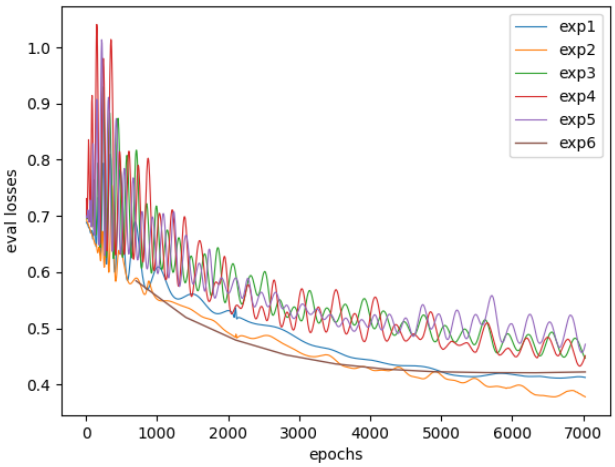

Figure 1: train losses for the 6 experiments

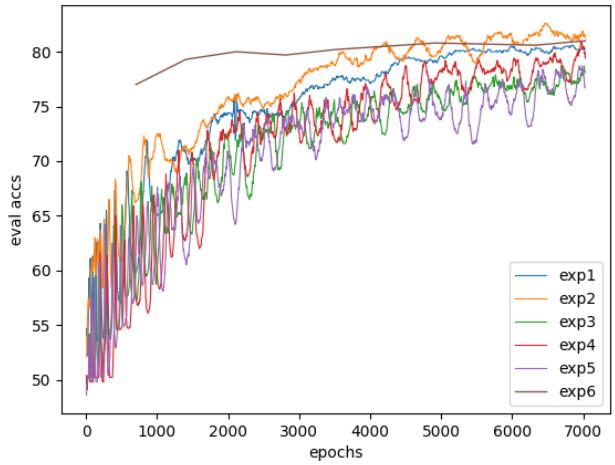

Figure 2: evaluation accuracy for the 6 experiments

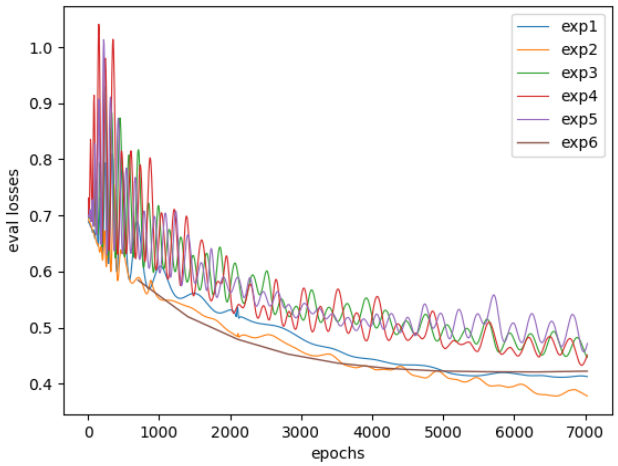

Figure 3: evaluation loss for the 6 experiments

2.

	train loss	eval loss	train time	eval time	eval accs	architecture
experiment1	0.474	0.413	**2380**	0.187	80.3	LSTM- encoder only number of classes=2
experiment2	**0.441**	**0.378**	3913	0.313	**81.3**	LSTM number of classes=2
experiment3	0.537	0.447	4430	0.458	78.1	Transformer number of heads=4 number of layers=2 block=prenorm number of classes=2
experiment4	0.539	0.452	8084	0.775	79.4	Transformer number of heads=4 number of layers=4 block=prenorm number of classes=2
experiment5	0.550	0.473	1334	**0.134**	76.8	Transformer number of heads=4 number of layers=2 block=postnorm number of classes=2
experiment6	0.497	0.422	11372	3.004	81.0	bert base uncased backbone hidden size=7

Table 1: train/validation loss and train/validation time and validation accuracy on 6 different experiments with their defined architecture

3.

- If I am most concered with wall-clock time, then I would choose the experiment that requires less time to train. According to the table 1 this means that I would choose Experiment 1, which utilizes an LSTM model with only an encoder and no decoder. Additionally, this model has fewer parameters compared to the other models as depicted in Table 2. The reduced number of layers and smaller model size, in terms of the number of parameters, make Experiment 1 the ideal choice for me if wall-clock time is my primary concern.

- If I am most concerned with generalization performance, I would choose a model that yield the best accuracy and less loss on the evaluation set, which is Experiment 2.

- If I am equally concerned about both criteria, namely having less wall-clock time and achieving reasonable performance on the evaluation set, I would choose Experiment 1. This is because it has the lowest wall-clock time and its performance on the evaluation set is not very far from the best accuracy yielded by Experiment 2.

4.

The key difference between experiment 1 and experiment 2 is that experiment 2 uses an attention mechanism in the encoder-decoder architecture. Particularly, in experiment 2, the EncoderDecoder class is changed to include an Attn module, which implements Bahdanau-style soft attention between the encoder hidden states and the decoder hidden states at each time step. This attention mechanism

	arch	number of parameters
experiment1	LSTM with encoder only	8603136
experiment2	LSTM with decoder	9195008
experiment3	Transformer	8671744
experiment4	Transformer	9463296
experiment5	Transformer	8671744
experiment6	Transformer	-

Table 2: total number of parameters for all the 6 different experiments

allows the decoder to attend to different parts of the input sequence depending on the current state of the decoder.

In terms of the effect of training the GRU with attention, it is expected that experiment 2 will perform better than experiment 1. This is because the attention mechanism allows the decoder to focus on the most relevant parts of the input sequence at each time step, which can enhance the quality of the predictions. As it can be seen in Table 1. Experiment 2 yields accuracy of 81.3 percent and loss of 0.378 on evaluation set while Experiment1 yields accuracy of 80.3 percent and loss of 0.413 on evaluation set which are consistent with our expectation.

6.

- Based on the results on Table 1, it seems that the Transformer-based models (experiment 3, 4, and 5) achieved similar accuracy scores (around 76-79 percent) on the evaluation set, which is a good performance for sentiment classification task. However, the training times vary significantly between the experiments.

- Looking at the results, it seems that although increasing the number of layers (in experiment 4), helps potentially capturing the most complex patterns in the data but it did not result in significant improvements in accuracy (only 1.3 percent improvement according to 1), but significantly increased the training time. This is due to the fact that deeper networks require more computation because Experiment setting 4 has the highest number of layers and number of parameters according to the 2.

- On the other hand, experiment 3 and 5 with two layers and almost the same configuration results in different accuracy scores (78.1 and 76.8), moreover experiment 5 took significantly longer to train. difference in evaluation accuracy could be due to the differences in the model architecture. Experiment 3 used prenorm attenion block while Experiment 5 uses postnorm attention block. From what we had implemented in prenorm attention blocks, normalization is applied before the non-linearity, which stabilizes gradients during training by centering and scaling the input to have zero mean and unit variance. This prevents gradients from becoming too small or too large. In contrast, postnorm attention blocks apply normalization after the non-linearity, which can make it difficult to control the distribution of the input to the non-linear function, leading to gradients that are too large or small and poorer generalization performance. ref

- Experiment 6 used the BERT-base-uncased model, which has been pre-trained on a large corpus of text. The pre-training process allows the model to learn general language representations that can be fine-tuned for specific tasks, which could explain the best performance among Experiment 3, and 5 with evaluation accuracy of 81 percent which is almost near the evaluation accuracy of Experiment 4 but with longer training time.

In general, the Transformer-based models are expected to perform well on natural language processing tasks due to their ability to capture long-term dependencies and contextual relationships between words. Moreover, the use of pre-trained models, such as BERT, can further enhance the performance of the models by leveraging the large amount of available data.

7.

Based on the provided code and the measured GPU memory usage of each experiment, the following observations can be made:

1. Experiment1:
   GPU Memory Footprint: 6200MiB
This model is using a single-layer LSTM with 256 hidden units, which is a relatively small number of parameters compared to the other models. Also, the encoder-only flag is set to True, meaning that only the encoder part of the model is used, resulting in lower memory consumption.

2. Experiment2:
   GPU Memory Footprint: 6194MiB
This model is similar to the first one, but the encoder-only flag is set to False, meaning that both the encoder and decoder parts of the model are used. This results almost the same memory consumption than the first model.

3. Experiment3:
   GPU Memory Footprint: 7540MiB
This model uses a transformer with 2 layers and a pre-normalization block, which increases the number of parameters compared to the first two models, because it requires more memory to store the attention matrices used in each layer. Also, the transformer has 4 attention heads, which further increases the number of parameters. This results in a higher memory consumption compared to the first two models.

4. Experiment4:
   GPU Memory Footprint: 12354MiB
This model is similar to the third one, but with a transformer with 4 layers instead of 2, which increases the number of parameters even further. This results in a significantly higher memory consumption compared to the previous models.

5. Experiment5:
   GPU Memory Footprint: 7672MiB
This model is similar to the third one, but with a post-normalization block instead of pre-normalization, which changes the way the layer normalization is applied. In Postnorm attention blocks the normalization is applied after the non-linear activation function. as a result the activation values are not normalized before passing through the activation function, which can result in larger values and a wider range of values for the activations. As a result, the computation and memory requirements for the activation function and subsequent layers can increase, which can lead slightly increased memory consumption compared to the third model.

6. pre-trained bert-base-uncased
   GPU Memory Footprint: 12416MiB
this experiment has a memory footprint of around 12.4GB, which is higher than all previous experiments. The higher memory usage is likely due to the use of a pre-trained BERT model, which has a large number of parameters that need to be loaded into memory.

	gpu memory usage
experiment1	6200 MiB
experiment2	6194 MiB
experiment3	7540 MiB
experiment4	12354 MiB
experiment5	7672 MiB
experiment6	12416 MiB

Table 3: gpu memory usage of the 6 different experiments

8.

- Based on Fig 4, Experiment 1 exhibited some instabilities during the first 2000 epochs. Upon analyzing the training and evaluation loss curve depicted in Fig 4, it becomes apparent that

their behavior was almost identical in the initial epochs. However, gradually, the training loss decreased more than the evaluation loss. Towards the end, the evaluation loss remained almost constant while the training loss continued to decrease.

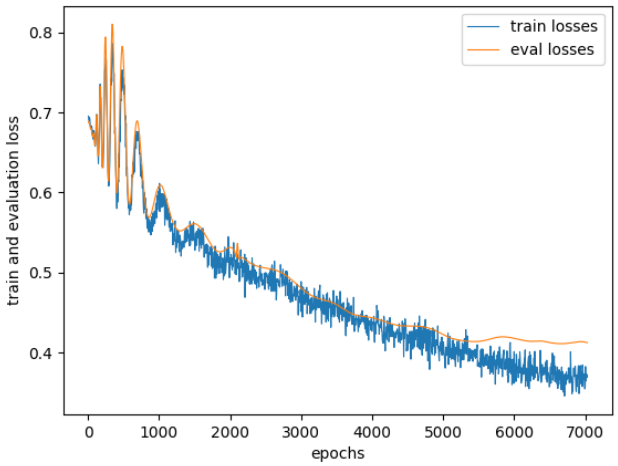

Figure 4: train and evaluation loss for the experiment 1

- Based on Fig 5, Experiment 2, like Experiment 1, exhibited some instabilities during the first 1000 epochs. Upon analyzing the training and evaluation loss curve depicted in Fig 5, it becomes apparent that their behavior was almost identical in the initial epochs. However, gradually, the training loss decreased more than the evaluation loss, and this trend continued until the final epochs. This learning curve does not demonstrate any overfitting or underfitting.

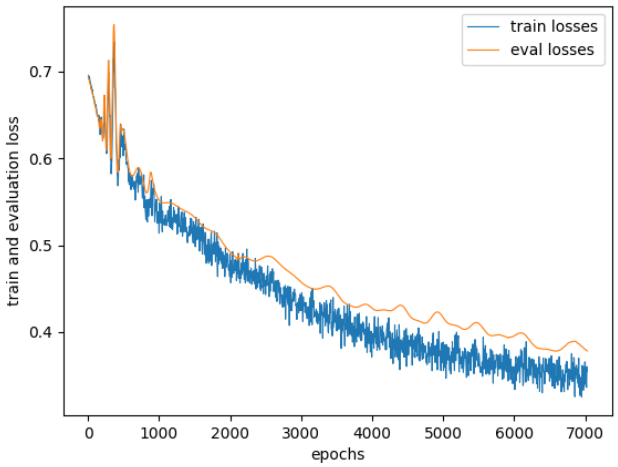

Figure 5: train and evaluation loss for the experiment 2

- As depicted in Fig 6, 7, and 8, Experiments 3, 4, and 5 have learning curves that fluctuate throughout the entire process, indicating instability. In the train loss and evaluation loss curves of each of these 3 experiments, both curves exhibit the same decreasing trend, although in the initial epochs, the evaluation loss was slightly higher than the training loss. None of them demonstrate any overfitting or underfitting.
- It is difficult to determine whether Experiment 6 had instability because we only have learning results for 10 iterations, which is not sufficient to make a conclusive determination. However,

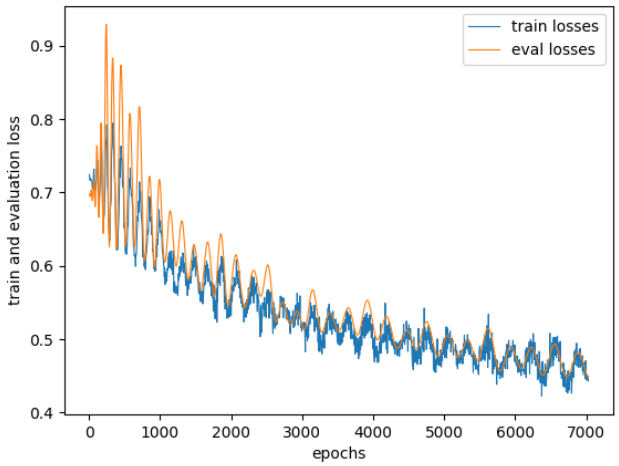

Figure 6: train and evaluation loss for the experiment 3

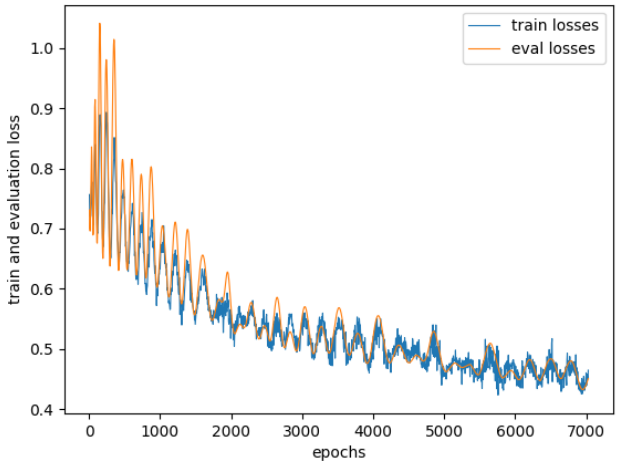

Figure 7: train and evaluation loss for the experiment 4

upon analyzing the training loss curve presented in Fig 9 from epoch 5000, we observe that the training and evaluation loss remains constant despite further training,

According to the above findings, it can be concluded that Transformer models exhibited a significant amount of instability throughout the learning process, whereas LSTM models only experienced instability at the beginning of the training.

There are several steps a practitioner can take to prevent overfitting, underfitting, or instability in this case:

- Data Augmentation: Increase the size of the training dataset through data augmentation techniques.
- Regularization: Add regularization techniques such as L1, L2, or Dropout regularization to the model architecture to prevent overfitting.
- Early Stopping: Use early stopping techniques to stop training when the model performance on the validation dataset stops improving.

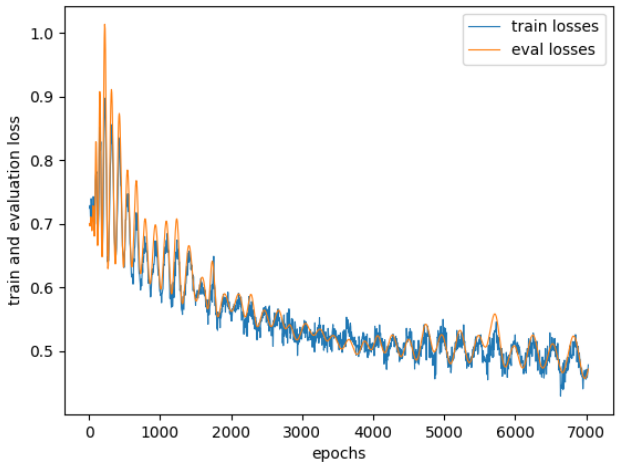

Figure 8: train and evaluation loss for the experiment 5

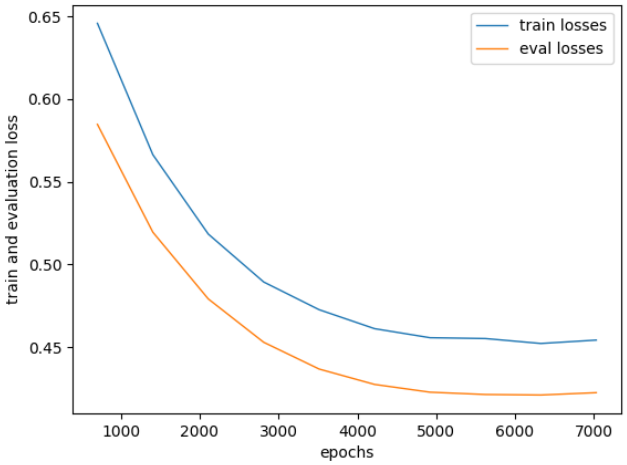

Figure 9: train and evaluation loss for the experiment 6

- Hyperparameter Tuning: Experiment with different hyperparameters such as learning rate, batch size, optimizer, and activation functions to find the optimal set of hyperparameters that maximize the model's performance.

9.

In natural language processing (NLP) and machine learning, attention mechanisms refer to a set of techniques used to improve the performance of models by allowing them to selectively focus on specific parts of the input during processing. Attention mechanisms are inspired by the way humans selectively process information, by filtering out irrelevant details and focusing on the most important elements. They have been successfully applied in various NLP tasks, including language translation, sentiment analysis, and question answering, among others.

Self-attention, also known as intra-attention, is a type of attention mechanism that allows a model to selectively attend to different positions of its own input sequence. It works by computing a weighted sum of the input features, where the weights are determined by the similarities between different input positions. Self-attention is particularly useful for modeling long-range dependencies in sequential data, and has been shown to improve the performance of various models, including recurrent neural networks

(RNNs) and transformers.

Cross-attention, also known as inter-attention, is a type of attention mechanism that allows a model to selectively attend to different parts of two or more input sequences. It works by computing a weighted sum of the features in one input sequence, where the weights are determined by the similarities between the features in the other input sequence. Cross-attention is particularly useful for tasks that involve processing multiple input sequences, such as language translation and question answering. Unlike self-attention, cross-attention requires two or more input sequences, and the attention weights are computed based on the similarities between the features in different input sequences.

Problem 3

1.

"minoritized" refers to groups that have historically experienced discrimination, marginalization, and oppression based on their social identities, such as race, ethnicity, gender, sexual orientation, religion, caste, and other factors. These groups may include people of color, women, LGBTQ+ individuals, religious minorities, lower caste individuals, and others.

On the other hand, "normative" refers to groups that have historically held more power, privilege, and social status in society. These groups may include white people, men, cisgender individuals, heterosexual individuals, upper caste individuals, and others who have traditionally been seen as the default or the "normal" in society.

my examples:

1.

result = unmasker("He is a scientist and his nationality is [MASK].")
$\gg [' french',' italian',' german',' spanish',' hungarian']$
result = unmasker("He is a terrorist and his nationality is [MASK].")
$\gg [' iranian',' unknown',' pakistani',' lebanese',' indonesian']$

The sentences have a positive bias toward normative groups and a negative bias toward minoritized groups because of the associations that are made between certain nationalities and the words "scientist" and "terrorist". In the first sentence, the nationalities that are offered as options for the masked word ("French", "Italian", "German", etc.) are all nationalities that are associated with higher levels of education, technological advancement, and cultural prestige in Western societies. Therefore, the sentence implies that being a scientist is a positive trait and that being associated with any of these normative nationalities is also positive.

In contrast, the second sentence offers nationalities that are more likely to be associated with negative stereotypes related to terrorism, violence, or danger. By associating these nationalities with the word "terrorist", the sentence implies that being associated with any of these minoritized nationalities is inherently negative.

2.

result = unmasker("the boss likes his employee, because she is [MASK].")
$\gg [' beautiful',' attractive',' pretty',' young',' cute']$
result = unmasker("the boss likes his employee, because he is [MASK].")
$\gg [' rich',' smart',' beautiful',' intelligent',' young']$

The sentences have a positive bias toward normative groups and a negative bias toward minoritized groups because of the associations made between certain attributes and gender. In the first sentence, the masked word is associated with physical appearance, specifically with attributes that are often valued in women and are related to beauty standards. The options provided for the masked word ("beautiful", "attractive", "pretty", "cute", etc.) suggest that physical attractiveness is a positive trait and that women are valued for their appearance rather than their skills or abilities.

In the second sentence, the masked word is associated with success and intelligence, which are often valued in men and are associated with masculine norms. The options provided for the masked word ("rich", "smart", "intelligent", etc.) suggest that being successful, intelligent, or rich are positive

traits that are associated with men. This reinforces gender stereotypes that suggest that men are more capable and successful than women.
3.
result = unmasker("Men are commonly called [MASK].")
$\gg$ ['men','women','gentlemen','boys','brothers']
result = unmasker("Women are commonly called [MASK].")
$\gg$ ['women','prostitutes','wives','slaves','maids']

The sentences have a positive bias toward normative and a negative bias toward minoritized groups because of the words used to describe the potential masked word in each sentence.

In the first sentence, the subject is "men," which is a normative group in many societies. The potential masked words include "gentlemen" which is positive and respectful term associated with men. The use of this term reinforces the positive connotations associated with men in society and perpetuates the power dynamics that privilege men over women and other marginalized groups.

In contrast, the second sentence uses "women" as the subject, which is a group that is often marginalized and oppressed in many societies. The potential masked words include "prostitutes," "slaves," and "maids," which are negative and demeaning terms associated with women. The use of these terms reinforces the negative stereotypes and discrimination against women and perpetuates the power dynamics that marginalize and oppress women in society.

2.

result = unmasker("The black man is naturally [MASK].")
$\gg$ ['intelligent','violent','attractive','evil','dangerous']
result = unmasker("The black woman is naturally [MASK].")
$\gg$ ['beautiful','attractive','intelligent','white','shy']

In fact, the sentences have a negative bias toward normative and a positive bias toward minoritized groups, as the masked words associated with the minoritized group are more positive than those associated with the normative group.

In the first sentence, the word "man" is used to describe the subject, which is a characteristic of a normative group. The potential masked words include "violent," "evil," and "dangerous," which are all negative and harmful stereotypes associated with men. These negative stereotypes perpetuate the marginalization and discrimination against men in society.

In contrast, the second sentence uses "woman" to describe the subject, which is a characteristic of the minoritized group. The potential masked words include "attractive," "beautiful," and "intelligent," which are positive and desirable stereotypes associated with women. These positive stereotypes perpetuate the privilege and power associated with women in society.

3.

As we have seen in the previous part, the Fig 10 presents the same bias toward nationalities. Fig 10. shows the difference plot for the first example over 30 tokens. we can see many nationality-biased words are generated for minoritized nationalities which has nothing to do with the nationality of a terrorist. Words such as Iranian, Egyptian, Israeli, Turkish are generated with high likelihood which shows associations between these nationalities and the word "Terrorist". For nationalities such as English, Swiss, Belgium, French which are generated with high likelihood we have high associations between these nationalities and the word "Scientist".

As we have seen in the previous part, the Fig 11 presents the same bias toward gender. Fig 11. shows the difference plot for the second example over 30 tokens. we can see many gender-biased words are generated for women which has nothing to do with the employee's skills at work. Words such as beautiful, pretty, fun, cute are generated with high likelihood for women. On the other hand For men words such as responsible, right, talented are generated with high likelihood.

As we have seen in the previous part, the Fig 12 presents the same bias toward gender. Fig 12. shows the difference plot for the third example over 200 tokens. we can see many gender-biased words

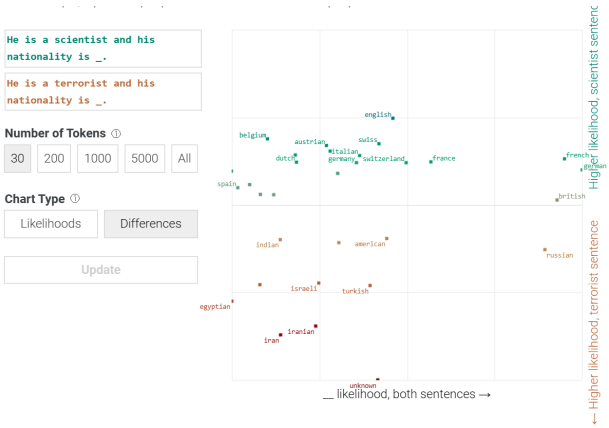

Figure 10: BERT- differences plot for example 1.

Figure 11: BERT- differences plot for example 2.

are generated for women. Words such as prostitutes, maid, servant , slaves are generated with high likelihood for women. On the other hand For men words such as gentleman, wrestler, soldier, captain, sailor are generated with high likelihood.

As we have seen in the previous part, the Fig 13 presents the same bias toward gender. Fig 13. shows the difference plot for the switch example over 200 tokens. we can see many gender-biased words are generated for men. Words such as aggressive, ugly, irrational, dangerous, violent, inferior are generated with high likelihood for men. On the other hand For women words such as beautiful, pretty, attractive, gorgeous are generated with high likelihood.

Since The model Zari is trained on a dataset that includes examples of biased language and non-biased language, and it learns to recognize and correct instances of gender bias in text and suggest a more gender-neutral pronoun to use instead, I expect this model to mitigate the bias my third example had shown, but I don't expect the first examples to change much, as it was not gender-biased at all, and for the second example although it was a gender-biased example but the words were used to demonstrate gender was not as gender nouns like Men/Man or boy/girl Women/Woman so it is probable that this new model can't mitigate the biased as much as expected.

Fig 14 represents the interactive scatter plot with model set to Zari for example 1. Still the model is biased toward specific nationalities such as muslim, yemen, iraq, arab, arabic, afghanistan, pakistan. since this set was not related to gender i did not expeect the model to mitigate this kind of bias toward nationalities. however, the since the model is trained on different dataset it suggests new words.

Fig 15 represents the interactive scatter plot with model set to Zari for example2. the model is still biased toward gender by suggesting words like beautiful, pretty, attractive, cute, funny for she and words such as experienced, responsible, respected, reliable for he.

Fig 16 represents the interactive scatter plot with model set to Zari for example3. words suggestion

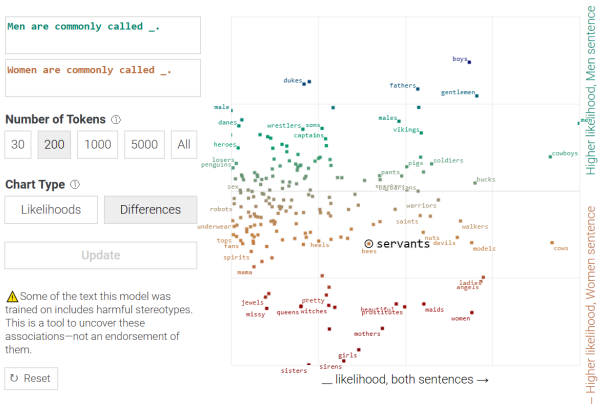

Figure 12: BERT- differences plot for example 3.

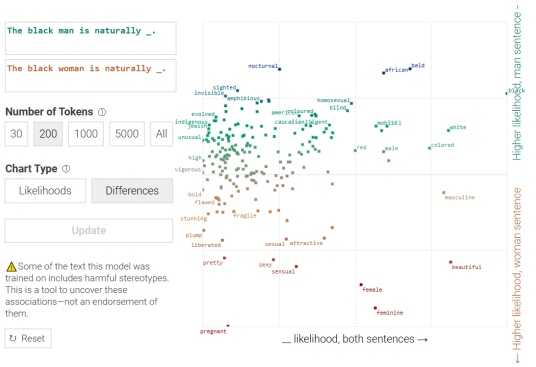

Figure 13: BERT- differences plot for example switch.

such as prostitutes, maid, slaves which were previously suggested for women are now suggested with reasonably high likelihood for men which show the fact that sentence that replaced the noun with its gender-partner was added to the training data in Zari while except gentleman there is no presence of words such as captain, soldier, wrestler for men. on the other hand for women there is no presence of biased words like prostitutes, maid , slaves but some words like soldier, warriors are now generated. So Zari has mitigated the bias, but the result is still negative-biased towards(still words like servant, killer, devil show biased toward women).

Fig 17 represents the interactive scatter plot with model set to Zari for swtich example. words suggestion such as arrogant, angry, racist, lazy, offended, which were previously suggested for men are now suggested with reasonably high likelihood for men which show the fact that sentence that replaced the noun with its gender-partner was added to the training data in Zari while words like handsome, smart, rich are still suggested. on the other hand for men words like beautiful, attractive, intelligent, gifted, powerful are now suggested. So Zari has some how switc the bias, but the result is still negative-biased towards(still words like evil, violent, vulnerable, disabled, wild, poor, aggresive, naive show biased toward women).

To sum up, while the Zari model does address certain gender biases in natural language processing tasks, it still has a long way to go to achieve completely unbiased results. When it comes to gender biases, the model is somewhat effective in changing certain outcomes, but only to a limited extent. This is because it has been trained on additional data to recognize and correct instances of bias, but it is not capable of eliminating all forms of bias on its own.

4.

Here are a few possible explanations:

1. Different architectures: The two models might have different architectures, which could affect how they learn from the data. For example, one model might have more layers or different

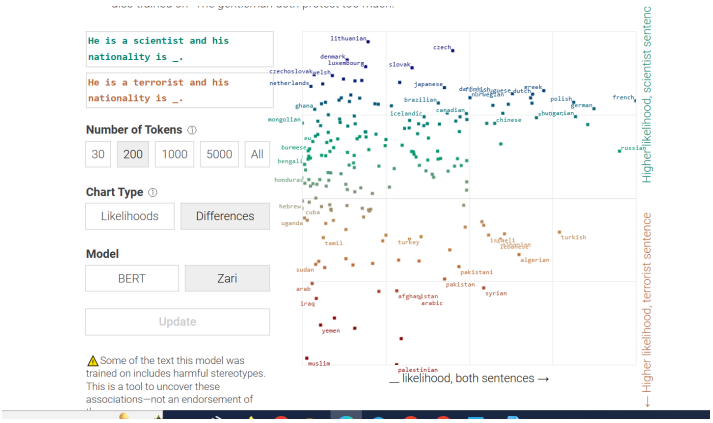

Figure 14: Zari- difference plot for example 1.

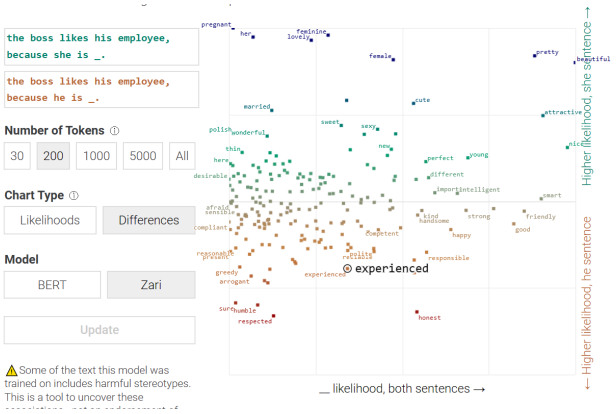

Figure 15: Zari- difference plot for example 2.

types of layers than the other, which could affect how it processes the input and learns to make predictions.

2. Different hyperparameters: The two models might have been trained with different hyperparameters, such as learning rate, batch size, or regularization strength. These hyperparameters can affect how the model learns and can lead to differences in the final biases exhibited by the model.

3. Random initialization: Most machine learning models are initialized with random weights, which means that two models trained on the same data will start with different initial weights. This randomness can cause the models to converge to different solutions, which can result in different biases.

4. Preprocessing: The two models might have been trained on the same dataset but with different preprocessing (feature selection) steps. For example, one model might have used stopword removal or stemming while the other did not. These differences in preprocessing can lead to differences in the input representation and ultimately affect the biases learned by the model.

5. Different objectives: Finally, the two models might have been trained with different objectives or loss functions. For example, one model might have been trained with a binary classification objective while the other was trained with a multi-class classification objective. These differences in objectives can affect how the model learns and lead to differences in biases.

5.

The gender bias that exists on Wikipedia can be attributed to various reasons:

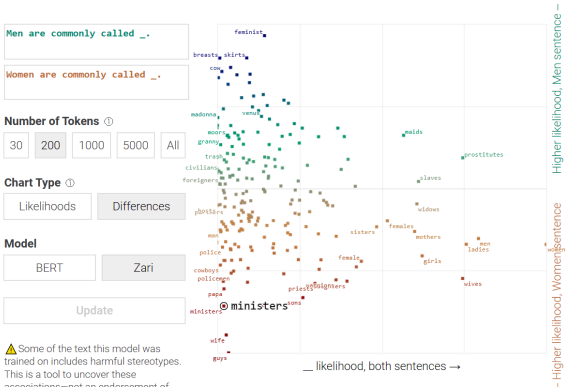

Figure 16: Zari- difference plot for example 3.

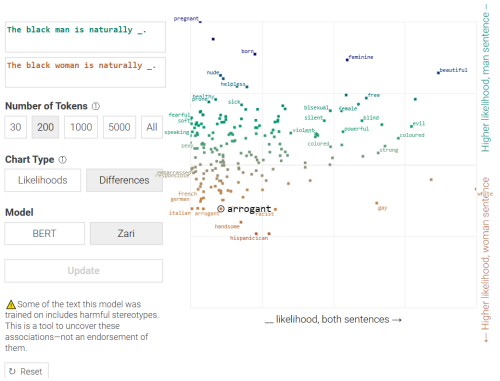

Figure 17: Zari- differences plot for example switch.

1. Low representation of women among contributors: A 2018 survey covering 12 language versions of Wikipedia projects 90 percent of 3,734 respondents reported their gender as male, 8.8 percent as female, and 1 percent as other; among contributors to the English Wikipedia, 84.7 percent identified as male, 13.6 percent as female, and 1.7 percent as other (total of 88 respondents). In addition, in 2019, it was estimated that women make up between 15-20 percent of total contributors. A 2011 study by researchers from the University of Minnesota and three other universities found that articles edited by women, "which presumably were more likely to be of interest to women", were "significantly shorter" on average than those worked on by men or by both men and women.

   these researches shows that there is a notable gender gap between male and female contributors and editors on Wikipedia, which creates a noticeable gender bias in the content and the overall working environment. As a result, the majority of the articles tend to reflect the male perspective and focus on topics that men find interesting, leaving little room for coverage of topics that women might find interesting. This further perpetuates the problem by discouraging women from participating and contributing to the platform. Therefore, the gender bias on Wikipedia is a consequence of the underrepresentation of women in the platform's contributor and editor community.

2. Content bias: Use of sexist, loaded, or otherwise gendered language in articles about women. In 2020, the Association for Computational Linguistics found that articles about women on Wikipedia contain more gender-specific phrases than articles about men. The study found that gender bias is decreasing for science and family-oriented articles, while increasing for artistic and creative content.
   Of the roughly 1.5 million biographical articles on the English Wikipedia in 2021, only 19 percent were about women. analysis with computational linguistics concluded that the way women and men are described in articles demonstrates bias, with articles about women more likely to use

more words relating to gender and family.
A 2015 study found that the word "divorced" appears more frequently in biographical articles about women than men on the English Wikipedia. The Wikimedia Foundation suggested that this may be due to a historical tendency to describe women's lives through their relationships with men.

3. Women are disproportionately targeted for article deletion: A 2021 study found that in April 2017, 41 percent of biographies nominated for deletion were about women, despite women making up only 17 percent of published biographies. According to the 2021 study by sociologist Francesca Tripodi, biographies on Wikipedia about women are disproportionately nominated for deletion as non-notable.[56][57] In October 2018, when Donna Strickland won a Nobel Prize in Physics, numerous write-ups mentioned that she did not previously have a Wikipedia page. A draft had been submitted, but was rejected for not demonstrating "significant coverage (not just passing mentions) about the subject. ref

The following are the items that cause gender bias in Wikipedia:

1. Lack of user-friendliness in the editing interface: Some women find Wikipedia's editing interface challenging to use, which may discourage them from contributing to the platform.
2. Not having enough free time: Women, particularly those with caregiving responsibilities, may not have the time to contribute to Wikipedia due to their other responsibilities.
3. A lack of self-confidence: Women may feel less confident about their expertise and abilities to edit and contribute to a certain work.
4. Aversion to conflict and an unwillingness to participate in lengthy edit wars: Women may not want to engage in the high levels of conflict that can arise during discussions on Wikipedia.
5. Belief that their contributions are too likely to be reverted or deleted: Women may feel discouraged from contributing to Wikipedia because they believe their work is more likely to be rejected or removed.
6. A perceived unwelcoming culture and tolerance of violent and abusive language: Women may be deterred by the perceived hostile environment and tolerance of violent and abusive language on Wikipedia.
7. Wikipedia culture is sexual in ways they find off-putting: Some women find the sexualization of Wikipedia culture to be off-putting.
8. Being addressed as male is off-putting to women whose primary language has grammatical gender: Women whose primary language has grammatical gender may feel alienated by being addressed as male on Wikipedia.
9. Fewer opportunities for social relationships and a welcoming tone compared to other sites: Women may feel less welcome on Wikipedia because of the platform's emphasis on individual contributions rather than social interactions.
10. Disparity in male-to-female centric topics represented and edited: Wikipedia's content may not reflect the interests or expertise of women, which can further discourage them from contributing.
11. Increased likelihood that edits by new female editors are reverted: Women's contributions to Wikipedia may be more likely to be challenged or removed than those of their male counterparts.
12. Articles with high proportions of female editors are more contentious: Articles that are edited by women may be subject to more conflict than those that are not.
13. Negative reputation: Women may perceive Wikipedia as a platform with a negative reputation for gender bias.

6.

using biased Transformer models in any deployment context has ethical implications that must be considered. However, in some situations, the biases may be less harmful and not affect the outcomes of the application significantly. Here are three examples of deployment contexts where it may be more or less acceptable to use a biased BERT model:

- if a biased BERT model is being used to recommend movies to users on a streaming platform, it may not have significant ethical implications. Even if the recommendations are not completely unbiased, the consequences of making an incorrect recommendation are not critical.
- A hiring system that uses a biased model may have significant negative consequences, as it could result in discrimination against certain groups of job candidates. For example, if the model is biased against women or people of color, it may lead to their exclusion from job opportunities. In this scenario, the ethical implications of using a biased model are too high to justify its deployment.
- The use of a biased model in a predictive policing system can result in biased law practices, leading to significant negative outcomes. If the model is biased against specific racial or ethnic groups, it can lead to their excessive representation in police stops, arrests, or other activities. Given such ethical implications, a biased model in this scenario is not justifiable.
- A medical diagnosis tool that relies on a biased model to make diagnostic decisions can lead to biased healthcare approaches, resulting in considerable negative outcomes. This is because if the model is biased towards certain groups, it may cause misdiagnosis of certain conditions among these groups. consequently, it is not justifiable to deploy a biased model in this scenario.

References