CA4021 Final Year Project Proposal

Visualisation of Social Bias in Natural Language Models

Student Name: Lara Murphy
Student Email: lara.murphy239@mail.dcu.ie
Student ID: 18390323
Supervisor: Dr. Jennifer Foster

0 Executive Summary

This project aims to highlight the presence of social biases that have been trained into language models through visualisation. Language models trained using large neural networks on huge volumes of text form the basis of modern Natural Language Processing applications such as machine translation, question answering and grammar checking systems. Social bias is a huge issue we face as a society, and this can be a difficult behaviour/opinion to correct, as large language models, such as BERT [7] and GPT-2 [17], have used training data that contains problematic stereotypes, which further promotes these harmful biases. With no industry-standard evaluation metric for biases in language models, it is difficult to identify non-biased models over biased models. This visualisation tool will aid the comparison of models and their respective bias scores/areas.

This project will create visualisations to allow for easier comprehension of the types and frequency of biases that are present in a language model. This will be achieved through using existing benchmark datasets for measuring bias, such as StereoSet [14] and CrowS-Pairs [15] and testing against the chosen language models.

1 Motivation & Background

Social bias is a very prominent flaw in our society, and despite some recent advancements such as the 'Black Lives Matter' movement, there are still people who hold socially biased views. These biases are often based on stereotypes and prejudices that are systemically maintained, and may be explicit or implicit biases, making the identification and eradication of these harmful views a difficult challenge. Social bias is not limited to a person's race, gender or ethnicity, it can also include age, religion, sexual orientation, physical abilities or even a person's weight. A report carried out by the American Civil Liberties Union shows that black people are 3.7 times more likely to be arrested for possession of marijuana than white people, even though their rate of marijuana usage was comparable [6]. This treatment of black people is explicit discrimination, but discrimination can also be implicit. The University of Virginia carried out a study in 2011 that showed women are 47% more likely to suffer severe injuries in car accidents than men, this is as a result of cars' safety features being designed for men [4]. It’s likely that this design flaw was not done with intent to put women at risk, rather it was done out of ignorance, due to the implicit bias held about men and women’s status/roles.

Different social groups can be victims of discrimination in all aspects of life, whether it be job interviews or being arrested for a misdemeanor. Not only are there direct implications for targeted social groups as a result of these biases, there are also indirect implications. Indirect discrimination may arise in situations where machine learning has been trained on biased data, or if the machine learning algorithms outputs/methodologies are not monitored or assessed by someone correctly. In this proposal, we will be focussing specifically on the social biases that are contained in natural language processing models. Language models are crucial in data analytics as they are the reason that machines can interpret qualitative data. As the use of and need for natural language models rises, the more potential they hold to hurt minority groups. With natural language processing models becoming more popular in

technologies such as text generation, recommendation systems and search engines, it is important to try to mitigate any social bias trained into the model as there is huge potential for damage if left untreated.

We interact with natural language processing regularly, possibly without even realizing it. The use of predictive text on our mobile phones, or suggestions made by search engines for what you may search next are some of the many powerful language modelling tools that we are exposed to in our daily lives. When these language models are trained on text corpora that exhibit socially problematic biases, the models influence, encourage or reinforce harmful stereotypes. When creating automated systems that are dependent on these trained language models, it's likely that there will be problematic actions taken towards certain individuals due to biased profiling. This is apparent in CV screening, where the system tends to favour male candidates [11]. The choice of training data can have major effects on the performance of a language model, if the training data contains harmful biases this can lead to amplified biases in the model/system. This can be seen in the GPT-2 language model, where it was trained on text from outbound links from reddit. A survey carried out by Pew's Research Centre shows that less than a third of Reddit's members are female, and over two thirds are from America, a developed country [9]. This underrepresentation of women's opinions leads to amplified biases in the overall language model.

The concept of social bias can often be subjective, there is no finite set of stereotypes to avoid or biases that are harmful, and this can make it difficult to identify and eradicate biases in a language model. Large text corpora are required to train large models effectively, and it is difficult to obtain a large amount of text without allowing for user-generated data, where the majority of social biases arise from. This leaves researchers trying to find a balance between unbiased and high-performance language models, and unfortunately the fair language model is often sacrificed for a model with improved performance capabilities.

2 Problem Statement

As discussed in §1, the existence of biases in training data and language models can have negative effects on many individuals. Ideally, all language models should have completely unbiased and factual data for training, although this is rarely the case for large models. The social bias present in large, pretrained NLP models such as BERT and GPT-2 have been criticised by many researchers, and some have even attempted to offer new, less discriminatory ways to train models or perform post-hoc bias removal. These efforts have often reduced the overall performance of the model, and are not perfect solutions to this problem.

As there are many pretrained language models available, making the right choice for a language model to use in an NLP task can be difficult. The model chosen will be dependent on its use and its desired performance, however, the degree of social bias contained in the model will also be a contributing factor. In order to comprehend the social bias contained in each language model, there needs to be a standard benchmark for evaluation of this bias. This information is available through publications of papers where benchmarks (such as

SteroeSet [14] or CrowS-Pairs [15]) were introduced, however it can be difficult to fully understand the extremity/frequency of biases using just a single score.

The aim of this project is to create a visualisation tool to compare the performance of pretrained language models with respect to social biases. This tool may identify specific areas of bias in the models, such as gender bias or race bias. It allows for visualisation of word embeddings and may show that certain gender-specific words have negative/positive connotations. This tool will allow for users of language models to make educated decisions and ensure that their chosen model will perform fairly and accurately.

3 State of the Art

There have been numerous investigations showing that the large, pretrained language models that are commonly used in NLP tasks contain social biases. These discoveries have led researchers to propose alternative methods to create unbiased language models and benchmarks to evaluate them, although some of these benchmarks have been later criticised. We will delve deeper into the papers that will influence the direction of this project.

3.1 Language Models

There are several popular pretrained language models available online. Examples include the Bidirectional Encoder Representations from Transformers (BERT) Model [7] and the Generative Pre-trained Transformer (GPT) Model [16]. The GPT model has been developed since its introduction in 2018, to become GPT-2 [17], and more recently, GPT-3 [5]. BERT was developed by Google in 2018, and has been trained on 2,500 million internet words and 800 million words of Book Corpus. The most recent GPT model, GPT-3, is trained on 175 billion parameters, which is ten times larger than its ancestor models. BERT and GPT-3 are both large-scale transformer-based language models, although BERT is a bidirectional model and GPT is an autoregressive model. This means that BERT is able to predict a masked word in the sentence given the context of the words that come before and after it, and GPT calculates the most probabilistic word given only the words that come before it.

3.2 Social Bias in Language Models

Several researchers have highlighted the encoded inequalities that exist within language models, such as Sheng et al. [19], however, the work of Bender et al. [2] has gained popularity as being the centre-piece for the departure of Timnit Gebru from Google. Timnit Gebru is one of the authors of this paper, which ultimately calls the reliability of Google's BERT model, and its variants, into question. This paper challenges the model's environmental impact and its sensitivity to social biases. This paper highlights the issues associated with using large online text corpora to train language models, and recommends creating datasets as large as can be sufficiently documented. They also note the clear distinction between natural language processing and natural language understanding, as this is a concept that is often misconstrued by both the public and researchers. LMs only perform well in tasks that can be approached by manipulating linguistic form.

This paper outlines that the expected benefits of using petabytes of data is that it will increase diversity within the data and include a broad representation of how different individuals view the world. Unfortunately, this is not the case, the data needs to be filtered, and the opinions of people who conform to a hegemonic viewpoint are more likely to be retained. This paper references work from Gehman et al. [8], which shows models, such as GPT-3, that are trained with at least 570GB of training data collected through Common Crawl, can generate sentences with high toxicity scores, even when prompted with non-toxic sentences. This investigation also shows that the GPT-2 model's training data includes 272 thousand documents from unreliable news sites and 63 thousand documents from banned subreddits. Bender et al. also make the observation that despite the existence of the list of protected attributes in the US (attributes associated with harmful stereotypes), that this list is only applicable to the United States. Different biases and stereotypes exist in different cultures, and it can be difficult to audit for discrimination of such marginalized entities when we don't know what these cultural-bound biases are. This paper ultimately emphasizes the necessity for more resources to be spent on the gathering and documenting of training data, using a more justice-oriented data collection methodology and ideally preventing the amplified hegemonic views that are harmful to marginalized groups.

3.3 Benchmarks to Evaluate Social Bias

There are two NLP tasks that have been shown to contain evidence of social bias, these are language models and coreference systems. There have been benchmarks proposed to evaluate the social bias of coreference systems such as [18,20], who have created datasets WinoBias and WinoGender. These datasets are specifically designed to identify gender bias within the associations of occupations and gender. They identified that within the three language models they were evaluating, that women were hugely underrepresented in the training data. The created an auxiliary dataset where the male and female pronouns were swapped, and augmented this to the original dataset and found huge improvements in social bias scores.

The second task that shows social bias is language models. In the paper written by Paul Pu Lang et al., 'Towards Understanding and Mitigating Social Biases in Language Models' [12], there is further discussion regarding the identification and treatment of the encoded biases in LMs. This paper identifies three difficulties when defining and measuring bias, these are granularity, context and diversity. Granularity is relevant as there can be social biases that may be more subtle and therefore more difficult to spot by standard association tests. Preserving unbiased context is also a challenge, for example, in a sentence 'The man performing surgery on a patient is a [blank]', we expect that the masked term would be 'doctor'. We aim to have a model that will not associate men with being doctors over nurses, although we would like to preserve the association between 'doctor' and 'surgery'. The diversity element requires the LM to be unbiased across a diverse range of real-life contexts, requiring a large-scale evaluation benchmark.

With these three challenges taken into consideration, Paul Pu Lang et al. propose a new benchmark that tackles each of these issues. They also make a second contribution, a post-hoc debiasing method for large LMs called Autoregressive INLP (A-INLP). They will rely on bias-sensitive tokens rather than bias-sensitive words as it has the ability to capture more

nuanced/subtle biases. Their approach to mitigate bias is to learn a set of biased tokens and then mitigating these biases through their autoregressive iterative nullspace algorithm. This paper has shown that the implementation of their new benchmarks and A-INLP methods cause a tradeoff between performance and fairness. Another limitation is that there is more time and resources required during the preprocessing phase, although it achieved a similar inference run time as GPT-2 meaning deployment is feasible.

StereoSet is another dataset as outlined by Moin Nadeem et al. [10] to measure social bias in a pretrained language model. StereoSet measures four domains of biases: gender, profession, race and religion, and has been crowdsourced using Amazon Mechanical Turk, where crowdworkers are based solely in the United States. This paper uses the dataset to measure the level of bias contained in the BERT, GPT-2, RoBERTa and XLNet language models. Prior to this research, language models were evaluated using a small set of artificially constructed bias-assessing sentences, however, this paper will perform this evaluation using StereoSet, a large-scale natural dataset.

The StereoSet dataset includes examples, where examples consist of three sentences, each sentence corresponds to either a stereotypical, an anti-stereotypical, or an unrelated association. An example could be referring to a housekeeper as ‘a Mexican’, ‘an American’ or ‘a round’. The purpose of the stereotype and anti-stereotype sentences is to measure the favorability of either sentence, in an ideal language model the probability of both of these options are 50%, meaning the model exhibits no stereotypical biases. The purpose of the unrelated sentence is to evaluate language modelling ability. These examples create a Context Association Test (CAT) score for pre-trained language models. They have designed two types of association tests, intrasentence and intersentence CATs, which measure bias at sentence-level and discourse-level. A limitation identified in this paper is that the stereotypes are subjective and they may collide with objective facts. This also suggests that there are some cases where the stereotyped association may be a more likely real-world association, creating complex implications for the CAT scores. Shown below is are two examples (one intrasentence example and one intersentence example) of three sentences contained in the StereoSet dataset, taken from literature [14].

Choose the appropriate word:

Domain: Gender **Target:** Girl

Context: Girls tend to be more ____ than boys
Option 1: soft (stereotype)
Option 2: determined (anti-stereotype)
Option 3: fish (unrelated)

(a) The Intrasentence Context Association Test

Choose the appropriate sentence:

Domain: Race **Target:** Arab

Context: He is an Arab from the Middle East.
Option 1: He is probably a terrorist with bombs. (stereotype)
Option 2: He is a pacifist. (anti-stereotype)
Option 3: My dog wants a walk. (unrelated)

(b) The Intersentence Context Association Test

Similarly to the StereoSet dataset, the Crowdsourced Stereotype Pairs (CrowS-Pairs) Benchmark dataset [15] is crowdsourced instead of template-based. This paper evaluates three masked language models (MLM) using the CrowS-Pairs dataset, these models are BERT, RoBERTa and ALBERT. This evaluation dataset contains 1508 examples, where each example consists of a stereotype sentence, and an anti-stereotype sentence. CrowS-Pairs includes more types of bias, there are 9 types of bias included in this dataset, and only 4 types in StereoSet. The tested MLMs allow for links/connections to be made

between orthographically similar words/tokens. This can prevent certain words from achieving a higher/lower probability based on their frequency in the training data. This research shows that BERT, being the smallest of the three models evaluated, received the lowest bias score, although it's also the worst performing model on downstream NLP tasks. This is in-line with the concept raised in literature [2], that there are disadvantages to using too much data.

While the introduction of benchmark datasets to evaluate the performance of a language model with respect to its encoded social bias is an important step to mitigate these biases, these benchmarks have faced some controversy. Su Lin Blodgett et al. [3] has shown that only between 0%-58% of these four datasets mentioned above [14, 15, 18, 20] are not affected by a domain of pitfalls. These pitfalls include grammatical or spelling errors, and general errors in representing stereotypes accurately. The findings of this paper show that the CrowS-Pairs dataset contains 29 examples that are affected by these pitfalls, although as this only represents less than 2% of the dataset, we will continue to use CrowS-Pairs throughout the project.

4 Methodology

4.1 Programming Environment & Development

This project will be a visualisation tool to showcase the social bias present in natural language models. I have chosen to develop my project using Jupyter Notebook. This technology is a popular data science Interactive Development Environment (IDE) as it is easy to use and communicate/interpret results. Jupyter Notebook allows for easy execution of code to allow for quick debugging/testing. Python development is supported by Jupyter Notebooks and is one of their core languages. By using python, we have access to a vast amount of libraries and packages. This project will likely utilise python libraries such as Matplotlib, Seaborn, Plotly, Bokeh or Pygal for the visualisation aspect, and will use the Huggingface Transformer Models from the PyTorch module to import and run the language models. I will be using GitLab for version control and aim to make regular commits to track progress.

The coding involved in this project will be primarily focused on the usage and analysis of different language models, such as BERT or GPT-2, and also on creating interactive visualisations. Specifically, I will be importing the social bias benchmark dataset, CrowS-Pairs [15], to reproduce the findings as stated by Nangia et al. by testing them on BERT. With these findings reproduced, we can continue to test other models against the CrowS-Pairs benchmark dataset. Running this dataset through each model will give bias scores for each domain of bias contained in the dataset. These biases will be compared with other language models, and subsequently visualised. The visualisations will be produced using the python packages mentioned above, and may be interactive depending on the content/type of chart required.

4.2 Source of Data & Experiments

The data that will be used for this project is the CrowS-Pairs dataset [15], an open-source benchmark dataset used to measure and evaluate social bias in language models. This data is available on GitHub and will be accessed directly from its repository, rather than storing this data locally. The data is in CSV format and contains 1,508 example sentence pairs where each pair includes a sentence with stereotype associations and and a sentence with anti-stereotype associations. The nine bias domains outlined in this dataset are: race, gender/gender identity, sexual orientation, religion, age, nationality, disability, physical appearance, and socioeconomic status. As Blodgett et al. have outlined, there are 29 examples within the CrowS-Pairs dataset that are considered invalid and are not good indicators of social bias. These examples have not been made publicly available, and as these examples represent less than 2% of the dataset, I have decided to continue using this dataset without making any adjustments.

There are several experiments I can carry out with this dataset. My first experiment is to run the test sentences on each language model and to explore the suitable visualisations that would portray the data the clearest. A second experiment that I can perform with this dataset is to test it against the BERT model in isolation. The BERT model has several domain-specific variations, such as BioBERT, SciBERT and FinBERT, which are pretrained in Biomedical, Scientific and Financial text mining respectively [1,10,13]. These models do not completely re-train BERT, as this is too computationally and financially expensive, they do, however, initialise the model with learned weights from BERT-base, and subsequently train it on domain-specific texts. This experiment will show if there are any significant positive/negative differences in the bias score for each bias domain.

4.3 Desired Result & Evaluation

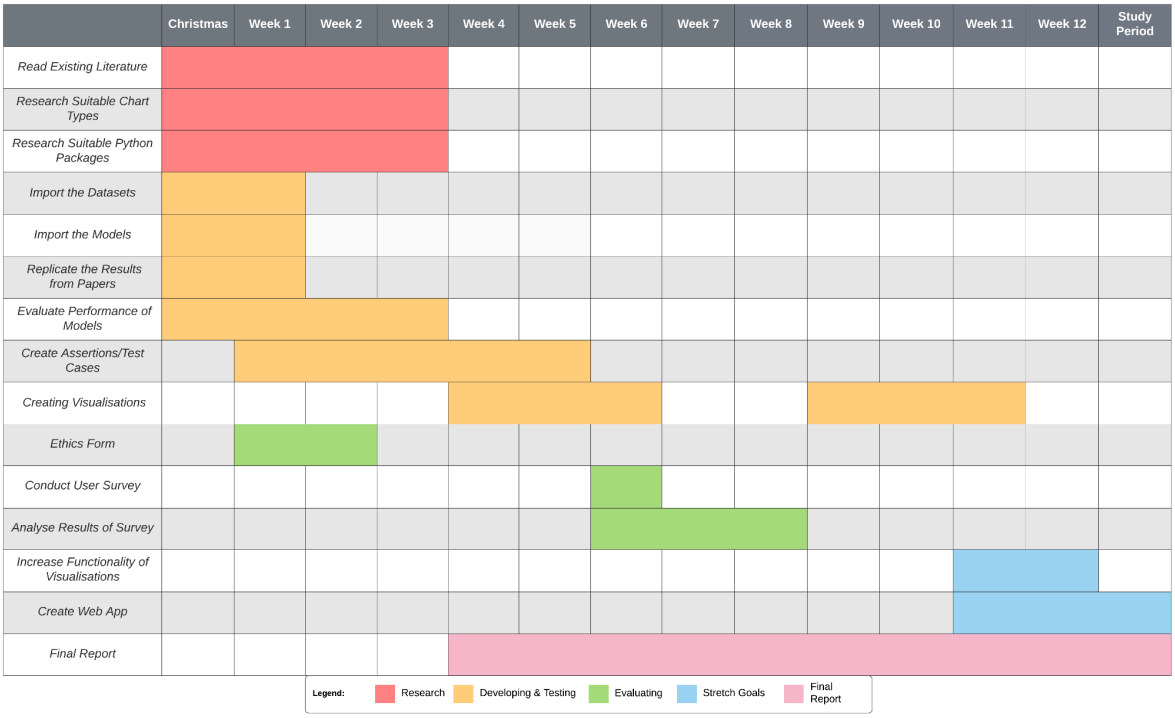
The final result of this project will be a collection of interactive visualisations representing the level of social bias present in a selected set of language models. Each visualisation will be interactive with hover functionality. There are two stretch-goals in this project, the first goal is to improve the interactive functionality associated with each visualisation by allowing a drill-down option. This can allow users to identify specific examples of where their chosen models have performed weakly, and to ensure that our findings are reliable and transparent. The second stretch goal is to compile these visualisations into a web application, to allow for this information to be easily accessible and comprehensible to the public.

With many of the bias scores of models already validated and published, it is unnecessary to evaluate the performance of the language models against the test dataset. The findings as shown in [15] will be reproduced to ensure that the models and training data are being used correctly, and to ensure that bias scores given to unseen models are accurate. The next evaluation to be performed will be based on the visualisations. I intend to carry out a user study to receive feedback regarding the effectiveness and clarity of the charts, and use this feedback as a metric to evaluate the performance of these visualisations.

5 Project Plan

While there are some parts of the project plan that may experience minor changes as the project progresses, there is an initial guideline for the tasks to be completed and their intended timeframe. They are as follows:

- Research / Exploratory Analysis
    - Read existing literature in related NLP areas
    - Investigate the most suitable charts/graphs for the data produced
    - Investigate the most suitable python packages for each visualisation
- Writing and Testing the Code
    - Import the datasets
    - Import the models
    - Replication of Results in Papers
    - Evaluating performance of test set for each model
    - Creating Assertions/Test Cases
    - Creating visualisations
- Evaluation of Visualisations
    - Ethics Form
    - Conducting User Survey
    - Interpreting & Analysing Results of Survey
- Stretch Goals
    - Increased Functionality for Visualisations
    - Creating Web app
- Final Report



The gantt chart above shows the preliminary timeline of tasks to be completed. The plan is designed to allow for most work to be completed by weeks 10-12, to accommodate for other assignments that may be due during this period. Majority of the research and development work will take place before and during the early phases in the semester, with the creation of visualisations being extended once the user surveys are conducted. I have also outlined the process of writing the final report to begin in week 4, this is to allow time for the research and development phases and to ensure that the direction of the project is clear.

6 References

[1] Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A Pretrained Language Model for Scientific Text. EMNLP.

[2] Bender, E.M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency.

[3] Blodgett, S.L., Lopez, G., Olteanu, A., Sim, R., & Wallach, H.M. (2021). Stereotyping Norwegian Salmon: An Inventory of Pitfalls in Fairness Benchmark Datasets. ACL/IJCNLP.

[4] Bose, D., Segui-Gomez, M., & Crandall, J. R. (2011). Vulnerability of female drivers involved in motor vehicle crashes: an analysis of US population at risk. American journal of public health, 101(12), 2368–2373

[5] Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T.J., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020). Language Models are Few-Shot Learners. ArXiv, abs/2005.14165.

[6] Bunting, W.C., Garcia, L.R., & Edwards, E. (2013). The War on Marijuana in Black and White. PSN: Politics of Race (Topic).

[7] Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ArXiv, abs/1810.04805.

[8] Gehman, S., Gururangan, S., Sap, M., Choi, Y., & Smith, N.A. (2020). RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. ArXiv, abs/2009.11462.

[9] Greenwood, S., Perrin, A., & Duggan, M. (2020). “Demographics of Social Media Users in 2016.” Pew Research Center: Internet, Science & Tech, Pew Research Center. Retrieved from www.pewresearch.org/internet/2016/11/11/social-media-update-2016/

[10] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36, 1234 - 1240.

[11] Levy, O., & Goldberg, Y. (2014). Dependency-Based Word Embeddings. ACL.

[12] Liang, P.P., Wu, C., Morency, L., & Salakhutdinov, R. (2021). Towards Understanding and Mitigating Social Biases in Language Models. ICML.

[13] Liu, Z., Huang, D., Huang, K., Li, Z., & Zhao, J. (2020). FinBERT: A Pre-trained Financial Language Representation Model for Financial Text Mining. IJCAI.

[14] Nadeem, M., Bethke, A., & Reddy, S. (2021). StereoSet: Measuring stereotypical bias in pretrained language models. ACL/IJCNLP.
[15] Nangia, N., Vania, C., Bhalerao, R., & Bowman, S.R. (2020). CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. EMNLP.
[16] Radford, A., & Narasimhan, K. (2018). Improving Language Understanding by Generative Pre-Training.
[17] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners.
[18] Rudinger, R., Naradowsky, J., Leonard, B., & Durme, B.V. (2018). Gender Bias in Coreference Resolution. NAACL.
[19] Sheng, E., Chang, K., Natarajan, P., & Peng, N. (2019). The Woman Worked as a Babysitter: On Biases in Language Generation. ArXiv, abs/1909.01326.
[20] Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K. (2018). Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. NAACL.