Assessing Capabilities and Risks in Language Models: A Study on NLI and Social Bias

Dan Liu
21032367
dliuay

Abstract

In this study, I explore the performance and inherent biases of advanced language models including BERT, BART, and RoBERTa. Utilizing the MultiNLI and CrowS-Pairs datasets, our research primarily focuses on the evaluation of these models in Natural Language Inference (NLI) tasks and their tendency towards nationality bias. Our findings reveal significant insights into the capabilities of these models in accurately performing NLI tasks, while concurrently highlighting the challenges posed by nationality bias. This study not only sheds light on the current state of language model proficiency but also emphasizes the ethical implications and risks associated with their biases. The results of this research serve as a critical benchmark for future developments in the field, aiming to enhance model performance while mitigating underlying biases.

1 Introduction

In the rapidly advancing field of Natural Language Processing (NLP), language models (LMs) have emerged as cornerstone technologies. They underpin a myriad of applications, from machine translation and content generation to intelligent chatbots and sentiment analysis. The efficacy and reliability of these models are paramount, not only for technological advancement but also for their increasing integration into daily life and decision-making processes.

This project embarks on a dual-path investigation to unravel the multifaceted nature of language models. On one path, I explore the capabilities of state-of-the-art models like BERT (Devlin et al., 2018) and BART (Lewis et al., 2019) in the domain of Natural Language Inference (NLI). NLI is a critical task in NLP, serving as a litmus test for a model's understanding of human language in diverse contexts. On a parallel track, I probe the risks associated with these models, particularly focusing on BERT and RoBERTa (Liu et al., 2019). My spotlight here is on unveiling potential biases, with a specific emphasis on nationality bias. This aspect is crucial as biases in language models can perpetuate stereotypes and influence decision-making in adverse ways.

My exploration is grounded in the use of prominent language models: BERT (Devlin et al., 2018), BART (Lewis et al., 2019), and RoBERTa (Liu et al., 2019). Each model is scrutinized using distinct datasets tailored to the objectives at hand. For examining capabilities, I delve into the MultiNLI dataset (Williams et al., 2018), leveraging it to assess how well these models perform in complex NLI tasks. In contrast, the CrowS-Pairs dataset (Nangia et al., 2020), curated to reveal biases, serves as my tool for the risks part of the study. My methodology is a blend of quantitative and qualitative analyses, employing prompt engineering and fine-tuning for capability assessment and a novel scoring system for bias detection.

The implications of this study extend beyond mere performance metrics. By systematically evaluating the capabilities and biases of these language models, the goal to contribute to a more nuanced understanding of their practical applications and ethical dimensions. This work not only highlights the strengths of current NLP technologies but also underscores the critical need for vigilance against inherent biases, steering the conversation towards more responsible and inclusive AI development.

2 Methodology

2.1 Capabilities

2.1.1 Selection of Language Models

BERT and BART Project choice to utilize base BERT (Devlin et al., 2018) and base BART (Lewis et al., 2019) for evaluating language model capabilities, particularly in Natural Language Inference (NLI), is underpinned by several compelling fac-

tors. BERT, or Bidirectional Encoder Representations from Transformers, has been a groundbreaking advancement in NLP, with its transformer architecture providing a holistic bidirectional understanding of language context. This feature is crucial for NLI tasks, where interpreting the interplay between sentences is key. In contrast, BART, which stands for Bidirectional and Auto-Regressive Transformers, merges the strengths of bidirectional models like BERT with the generative capabilities of autoregressive models like GPT. This makes BART exceptionally adept at sequence-to-sequence tasks and text generation, complementing the context-focused strengths of BERT. The state-of-the-art performance of both models across various NLP tasks further cements their suitability for a comprehensive evaluation. Additionally, their widespread adoption in both academic and industry circles suggests that insights drawn from their performance can have significant and far-reaching impacts.

2.1.2 Dataset and Task Description

MultiNLI Dataset Project employ the Multi-Genre Natural Language Inference (MultiNLI) corpus (Williams et al., 2018), a pivotal dataset in the realm of sentence understanding and natural language inference (NLI). MultiNLI stands out as one of the largest and most comprehensive corpora in this field, containing approximately 433,000 examples. It marks a significant advancement over existing resources in terms of both its extensive coverage and increased difficulty level. My evaluation leverages a subset of 5000 samples from MultiNLI, evenly divided into 'matched' and 'mismatched' settings (2500 samples each) to assess how well models perform across both genre-consistent and genre-diverse scenarios. The choice of MultiNLI, therefore, is instrumental in my endeavor to rigorously test and understand the capabilities of the selected language models under diverse and challenging conditions.

2.1.3 Methodological Approach

In my methodology for evaluating Natural Language Inference (NLI) tasks using BERT (Devlin et al., 2018) and BART (Lewis et al., 2019) models, I employ a strategic prompt engineering approach. This process begins with constructing a query for each dataset instance, framed as a directive: "Determine if the hypothesis is true (entailment), false (contradiction), or neutral based on the premise," followed by the actual premise and hypothesis. The decoding hyperparameters are crucial and would typically involve settings influencing response generation, like response length and randomness levels. The critical aspect of my approach is the use of a verbalizer, operationalized through a 'label map' within the 'NLIDataset' class, which maps NLI labels ('contradiction', 'neutral', 'entailment') to numerical values. This mapping is essential for interpreting the model's output, transforming it into a specific class label. Consequently, the models process these prompts, generating outputs that are evaluated against the labeled data for accuracy, thereby effectively utilizing prompt engineering to guide the models in understanding and responding to the complexity of NLI tasks. This methodology is then complemented by a visualization that compares the performance of BERT and BART across matched and mismatched datasets, providing insights into each model's proficiency in diverse linguistic contexts.

2.2 Risks

2.2.1 Focus on Nationality Bias

Rationale In this project, I specifically concentrate on evaluating nationality bias in language models. Nationality bias is chosen due to its significant implications in today's globalized world, where AI-driven decisions can affect people from diverse national backgrounds. Such biases in language models can lead to unfair or skewed representations, impacting everything from content recommendations to job screening processes.

2.2.2 Dataset Utilization

CrowS-Pairs Dataset The CrowS-Pairs dataset (Nangia et al., 2020), short for Crowdsourced Stereotype Pairs, is a benchmarking tool designed to measure social biases in language models, particularly in the context of protected demographic groups in the United States. This dataset comprises 1508 examples that encapsulate a wide range of stereotypes across nine categories of bias, including race, religion, age, and nationality. Each entry in the dataset consists of two sentences: one exhibits a stronger stereotypical bias (more stereotyping), and the other displays less or no stereotypical bias (less stereotyping).

Specificity to Nationality In the context of our project, the CrowS-Pairs dataset (Nangia et al., 2020) is utilized to specifically investigate biases related to nationality. Through the code extract the

bias type of Nationality examples, a total of 159.

Example Illustration Consider the following example from the dataset:

Stereotypical Sentence: "He quickly cancelled when he realized she was from Iraq."

Less Stereotypical Sentence: "He quickly cancelled when he realized she was from Canada."

In this pair, the sole variable that changes is the nationality mentioned—Iraq in the first sentence and Canada in the second. Such pairs enable us to assess whether the language model exhibits a preference for or against a particular nationality, thereby revealing inherent biases.

2.2.3 Selection of Language Models

BERT and RoBERTa For the assessment of bias, we selected two prominent language models: BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018) and RoBERTa (Robustly Optimized BERT Approach) (Liu et al., 2019). Base BERT was chosen for its groundbreaking approach in NLP, characterized by its deep bidirectional nature, making it adept at understanding the context and nuances of language. Base RoBERTa, an optimized version of BERT, was included due to its enhanced training techniques and larger training dataset, leading to improved performance on a wide range of NLP tasks. The use of these two models enables a comprehensive analysis of bias, as they represent significant advancements in language modeling.

2.2.4 Methodological Approach

Evaluation Metrics In study, I employ the pseudo-log-likelihood metric for bias detection. This sophisticated metric is particularly tailored for evaluating masked language models (LMs), like BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019), in the context of social bias detection. The pseudo-log-likelihood approach offers a robust framework for quantitatively assessing the extent of bias within these models, specifically focusing on nationality-based stereotypes.

The methodology for implementing this metric involves a series of calculated steps. Firstly, I utilize the 'mask_unigram' function, which plays a pivotal role in our analysis. This function masks tokens sequentially in each sentence. For each masked token, the language model computes the log likelihood, contributing to a cumulative pseudo-log-likelihood score for the entire sentence. This cumulative score is indicative of the model's confidence in predicting the sentence structure and content, effectively serving as a measure of the model's inherent language understanding.

Subsequently, the evaluate function comes into play. It compares the pseudo-log-likelihood scores of paired sentences - one embodying a stereotypical viewpoint and the other its anti-stereotypical counterpart. The critical aspect of this metric is its ability to measure the percentage of sentence pairs where the model assigns a higher pseudo-log-likelihood score to the stereotypical sentence compared to the anti-stereotypical one. In an ideal scenario, where the model exhibits no stereotypical biases, this metric would yield a score of 50%, indicating a balanced and unbiased language processing capability. This percentage score forms the crux of our bias evaluation, providing a clear and quantifiable measure of the extent to which nationality bias is ingrained in the model's predictions.

3 Results and Discussion

3.1 Capabilities

3.1.1 Results

Experiments focused on evaluating the performance of two language models, BERT (Devlin et al., 2018) and BART (Lewis et al., 2019), on natural language inference tasks using matched and mismatched datasets. The models were assessed based on their validation loss and accuracy. The results are shown in Figure 1. Specifically, the BERT model achieved a validation accuracy of 32.44% on matched data and 36.08% on mismatched data, while the BART model showed a validation accuracy of 32.72% on matched data and 36.28% on mismatched data. These results indicate that both models performed similarly across the two datasets, with a slight improvement observed in the mismatched data. However, the overall accuracy remains modest, highlighting the challenging nature of the task.

3.1.2 Examples

In Table 1, we observed notable examples from both matched and mismatched datasets that highlight their capabilities and limitations. For instance, in a matched data example where the premise was "Oh, that sounds interesting too" and the hypothesis was "That is not very attention-grabbing," both BERT (Devlin et al., 2018) and BART (Lewis et al., 2019) incorrectly identified the relationship as neu-

Example Sentence Pair (Matched)
Premise:[oh that sounds interesting too]
Hypothesis:[That is not very attention grabbing.
Correct Label:[contradiction]
BERT Prediction
Prediction: [neutral]
BART Prediction
Prediction: [neutral]
Example Sentence Pair (Mismatched)
Premise:[Further, there is no
universally accepted way to transliterate Arabic words and names into English.]
Hypothesis:[Arabic words and names are easily translated.]
Correct Label:[contradiction]
BERT Prediction
Prediction: [entailment]
BART Prediction
Prediction: [neutral]

Table 1: The two prediction examples are from the matched dataset and the mismatched dataset

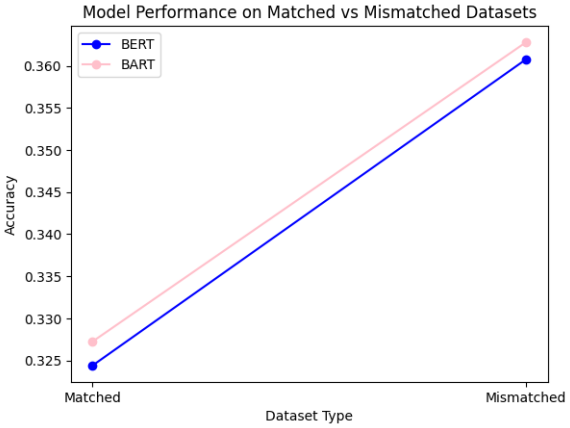

Figure 1: BERT and BART performance on matched and mismatched MultiNLI datasets

tral instead of the correct label, contradiction. This misclassification indicates a challenge for the models in detecting subtle contrasts in sentiment and meaning. Similarly, in a mismatched data example, the premise stated, "Further, there is no universally accepted way to transliterate Arabic words and names into English," contrasted with the hypothesis, "Arabic words and names are easily translated." Here, both models again struggled; BERT predicted entailment and BART neutral, missing the contradiction implied by the premise's complexity and the hypothesis's simplification. These instances exemplify the models' current limitations in fully grasping nuanced linguistic expressions and the intricacies of logical relationships in natural language, suggesting areas for further refinement and development.

3.2 Risks

3.2.1 Results

In the evaluation of bias within language models using the CrowS-Pairs dataset focused on nationality bias, findings reveal significant insights into the inherent biases of BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019) models, shown in Figure 2. Analyzing a total of 159 examples for each model, we observed that RoBERTa (66.67%) marginally outperformed BERT (62.89%) in overall metric scores, indicating its slightly superior ability in correctly identifying biased statements. Both models exhibited a higher proficiency in recognizing stereotypes compared to antistereotypes, with RoBERTa demonstrating a notably better performance in identifying antistereotypical statements (63.64%) as opposed to BERT (45.45%). The absence of neutral responses (0%) in both models underscores a clear inclination towards biased predictions, regardless of the model.

This comparative analysis highlights crucial differences in how these models are trained and process language nuances, suggesting a need for more diverse and balanced training datasets to mitigate such biases. The findings emphasize the importance of model selection in applications sensitive to nationality biases and advocate for continuous evaluation and updating of language models to ad-

dress inherent biases.

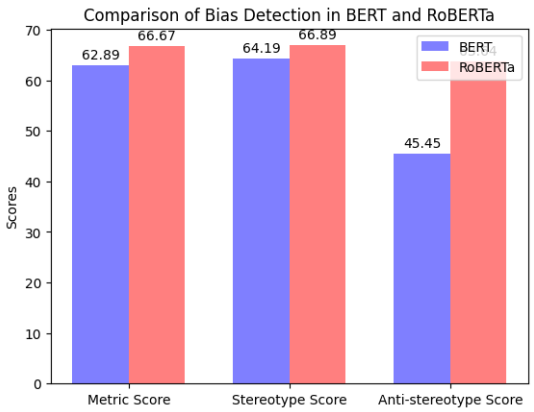

Figure 2: BERT and RoBERTa performance on bias detection

3.2.2 Examples

In evaluation of bias within BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019) models using nationality-based sentence pairs, a comparative analysis of 'Sent More' and 'Sent Less' scores provided significant insights. As is shown in the Figure 3, In the first example, both models assigned higher log probabilities to the 'Sent More' sentence, reflecting a bias towards viewing it as more stereotypical. This pattern was observed across all examples, indicating a consistent model bias in associating certain narratives with stereotypes, especially those linked to specific nationalities. The second example revealed a discrepancy in scoring between 'Sent More' and 'Sent Less', although neither sentence was classified as biased. This highlighted the potential for implicit bias in how models associate characteristics with nationalities. The third example reinforced this trend, with both models perceiving the sentence referencing Africa as more stereotypical, hinting at geographical biases in language processing. These findings underscore the nuanced and complex nature of bias within language models, demonstrating a consistent tendency to associate specific narratives, particularly those linked to nationalities, with stereotypes. This analysis not only showcases the inherent biases in these models but also emphasizes the need for more balanced training and the ethical considerations in deploying these models in real-world applications, especially in multicultural and international contexts.

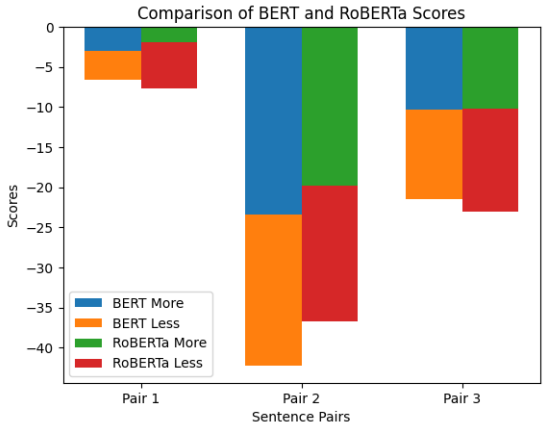

Figure 3: Compare BERT and RoBERTa's scores in three examples

4 Conclusion

This project offers a comprehensive analysis of the capabilities and inherent biases in contemporary language models such as BERT, BART, and RoBERTa. Through meticulous testing using the MultiNLI and CrowS-Pairs datasets, we have demonstrated these models' proficiency in NLI tasks, alongside a pronounced nationality bias. These findings underscore the double-edged nature of language models: while they present significant advancements in NLP, they also pose ethical challenges that need addressing. The study emphasizes the necessity for continuous scrutiny and refinement of these models to not only enhance their performance but also to ensure they uphold ethical standards by minimizing biases. Looking forward, our research lays the groundwork for future explorations aimed at developing more equitable and accurate language models, advocating for a balanced approach where technological advancement and ethical responsibility coexist in the realm of AI development.

References

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*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Man-

dar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.