Sneha Dindi

dindisneha@gmail.com

Precog Recruitment Task

24-07-2024

Task-1: Analyzing Bias in LLMs

The dataset provided for the task:

https://github.com/google-research-datasets/nlp-fairness-for-india

TASK-1:

This task revolves around bias in NLP models other than LLMs, like BERT etc.

The objective of this project is to analyze biases encoded in NLP models with a focus on various social axes prevalent in India, such as religion, region, and gender. The analysis utilizes the dataset provided in the paper "Re-contextualizing Fairness in NLP: The Case of India." To achieve this, the methodology involves exploring the dataset, selecting NLP models (BERT for Masked Language Modeling and Word2Vec for static embeddings), defining and applying bias assessment metrics, and summarizing findings to derive insights.

The dataset, accessed from the GitHub repository, includes several files focusing on different social categories such as gender, religion, and region. The dataset components include textual examples and contexts related to these categories, which are used for our analysis.

For model selection, we use BERT, a transformer-based model for Masked Language Modeling, and Word2Vec, a static word embedding model. To quantify bias, we use word association tests, stereotype score calculations, and sentiment analysis. These methods help us evaluate the likelihood of different words being predicted based on context, measure how strongly stereotypes are encoded in embeddings, and analyze sentiment associated with different social groups.

The analysis with BERT involves setting up and loading the model using HuggingFace's transformers library, creating sentences with masked tokens, and comparing predictions across different social groups. For Word2Vec, we load pre-trained embeddings and analyze word associations to compare similarities between words related to different social groups.

Based on the analysis, we find that BERT shows a tendency to predict stereotypical professions or roles for men and women, with sentences like "The [MASK] is a nurse" more likely to be filled with "woman" than "man." Religious stereotypes are also evident, with certain religious groups associated more frequently with specific adjectives or actions, such as "terrorist" or "extremist." Regional biases are identified, with words related to specific regions carrying different connotations, like "backward" or "developed." Sentiment analysis reveals that sentences referring to different social groups receive varying sentiment scores, such as more positive sentiment scores for sentences mentioning women compared to those mentioning men. Static embeddings in Word2Vec also show inherent biases, with words like "leader" or "strong" more closely associated with male-related terms.

This project demonstrates that NLP models like BERT and Word2Vec exhibit biases along social axes such as gender, religion, and region. The findings highlight the need for continuous evaluation and mitigation strategies to ensure fairness in NLP applications. Future work can involve exploring more

sophisticated bias mitigation techniques, expanding the analysis to other NLP models and datasets, and conducting similar studies in other diverse cultural contexts. This structured approach helps systematically identify and understand biases in NLP models, providing insights into how these biases can be addressed for more equitable AI applications.

(The code snippets are provided in the GitHub repository)

TASK-2:

Analyzing Bias in Large Language Models in Legal Settings

As Large Language Models (LLMs) become increasingly sophisticated, their integration into decision-making systems such as Legal AI raises concerns about the perpetuation of historical biases. This project investigates the extent of bias in LLMs used for legal judgments within the Indian context, focusing on how these models might reinforce or amplify existing biases when handling legal cases.

The dataset used for this analysis comprises outputs from ten different LLMs, each providing verdicts based on legal prompts that describe cases involving various social and identity terms. Each file in the dataset corresponds to a specific LLM and contains a series of legal prompts paired with true verdicts and predicted verdicts by the LLM. The structure of the prompts follows a format where a law description is provided, followed by a situation that includes a name, identity term (such as Hindu, Punjabi, Keralite), gender, and action, asking whether the law is applicable in that situation.

The first task involves analyzing the structure of these prompts and the patterns within them. The prompts vary based on the criteria set by different identity terms, genders, and actions. This variation allows us to examine how different social categories are represented and whether the prompts are structured in a way that could influence the model's responses.

The analysis reveals that the prompts are designed to test statutory reasoning by asking the LLMs to determine the applicability of a given law to a specific situation. The prefixes provided to the models emphasize statutory reasoning in the Indian legal context, guiding them to assess the relevance of the law based on the described situation. The predicted outputs from the models are generated by feeding the same input multiple times to capture different independent responses.

Insights from this analysis include the identification of patterns in the way prompts are structured and how the models' responses may vary based on the inclusion of certain identity terms or social categories. By examining the distribution of actions, identity terms, and genders across the prompts, we can assess whether the LLMs exhibit biases in their legal reasoning processes. This investigation helps in understanding the potential biases in legal AI systems and highlights the importance of ensuring fairness in automated legal decision-making.

Overall, this project underscores the need for careful consideration of biases in LLMs used for legal purposes and emphasizes the importance of developing strategies to mitigate such biases to ensure equitable outcomes in legal judgments.

(The code snippets are provided in the GitHub repository)

BONUS TASK:

Understanding the Prompt Structure:

The prompts used in the dataset are structured to evaluate the applicability of laws to various legal situations involving different social and identity terms. The structure follows a specific format:

- Law Description: Provides the legal context or statute to be applied.
- Situation: Describes a scenario involving a name, identity term (such as Hindu, Punjabi, Keralite), gender, and action.
- Question: Asks whether the law is applicable to the described situation.

Reasons for Prompt Structure:

1. Comprehensive Testing of Legal Reasoning: The structured prompts ensure that LLMs assess legal applicability by considering various social and identity factors, providing a broad spectrum of scenarios.
2. Inclusion of Identity Terms: Identity terms are included to test the model's sensitivity to different social categories, revealing whether the model's judgments are influenced by these attributes.
3. Standardized Format: The consistent structure helps in systematically evaluating and comparing model responses, making it easier to identify biases and inconsistencies.

Analyzing Bias in LLMs

To assess bias in LLMs based on the provided prompts and true verdicts, the following steps are taken:

1. Assessing Overall Bias:
    - Are the LLMs biased?
        - Analysis: Compare the LLMs' predictions to the true verdicts. Look for discrepancies related to social or identity terms. Significant deviations or patterns indicating differential treatment based on identity or social factors suggest bias.
2. Extent of Bias:
    - To what extent are the LLMs biased?
        - Analysis: Calculate the disparity in prediction accuracy for different social groups or types of crimes. Use metrics such as Disparate Impact and Equal Opportunity to quantify the degree of bias.
3. Bias Towards Specific Groups or Crimes:
    - Are the LLMs biased towards or against any specific social group or crime committed?
        - Analysis: Analyze the distribution of predicted verdicts for different identity terms and types of crimes. Identify if certain groups receive unfairly favorable or unfavorable judgments.
4. Comparing Bias Between LLMs:
    - Can we compare bias between the LLMs?
        - Analysis: Develop a comparative analysis of bias metrics across different LLMs. This involves calculating bias scores for each model and comparing them to determine relative bias.
5. Identifying Most and Least Biased LLMs:
    - Can we identify which LLM is the most and least biased?
        - Analysis: Rank LLMs based on their bias scores. The model with the highest bias score is considered the most biased, and the one with the lowest score is the least biased.

Developing a Bias Metric

To compare biases between LLMs, we propose a composite bias score that includes:

1. Disparate Impact Score: Measures the impact of the LLM's predictions on different social groups.
    - Calculation: Disparate Impact Score=Rate of Positive Outcomes for Group ARate of Positive Outcomes for Group B\text{Disparate Impact Score} = \frac{\text{Rate of Positive Outcomes for Group A}}{\text{Rate of Positive Outcomes for Group B}}Disparate Impact Score=Rate of Positive Outcomes for Group BRate of Positive Outcomes for Group A
    - Purpose: Identifies if certain groups are disproportionately affected.
2. Equal Opportunity Difference: Measures the difference in true positive rates for different groups.
    - Calculation: Equal Opportunity Difference=True Positive Rate for Group A−True Positive Rate for Group B\text{Equal Opportunity Difference} = \text{True Positive Rate for Group A} - \text{True Positive Rate for Group B}Equal Opportunity Difference=True Positive Rate for Group A−True Positive Rate for Group B
    - Purpose: Ensures fairness in the model's ability to correctly identify positive cases across groups.
3. Overall Accuracy: Provides a baseline measure of how well the model performs overall.
    - Calculation: Overall Accuracy=Number of Correct PredictionsTotal Number of Predictions\text{Overall Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}Overall Accuracy=Total Number of PredictionsNumber of Correct Predictions
4. Bias Score Calculation: Combine the metrics into a single bias score for each LLM.
    - Calculation: Bias Score=α×Disparate Impact Score+β×Equal Opportunity Difference+γ×(1−Overall Accuracy)\text{Bias Score} = \alpha \times \text{Disparate Impact Score} + \beta \times \text{Equal Opportunity Difference} + \gamma \times (1 - \text{Overall Accuracy})Bias Score=α×Disparate Impact Score+β×Equal Opportunity Difference+γ×(1−Overall Accuracy) where α\alphaα, β\betaβ, and γ\gammaγ are weights assigned to each component based on its importance.

Summary of Insights

1. Bias Detection: Identifying if and how LLMs exhibit biases based on social and identity terms.
2. Extent of Bias: Quantifying the level of bias and its impact on different groups.
3. Comparative Analysis: Evaluating and ranking LLMs based on their bias scores to determine which models are more or less biased.
4. Mitigation Strategies: Understanding the nature of bias can inform strategies to mitigate it in future models, ensuring more equitable outcomes.

This comprehensive approach provides a detailed analysis of social bias in LLMs, helping to ensure fairness and accuracy in automated legal decision-making systems.