










 Anngladys / AI-Tools-Assignment



 Code  Issues  Pull requests  Actions  Projects  Wiki  Security 

[AI-Tools-Assignment](#) / [part3_ethics_optimization](#) / [ethical_considerations.md](#) 

Anngladys feat: Added initial practical tasks and ethics structure

6df3479 · 1 hour ago



36 lines (27 loc) · 4.82 KB

Preview

Code

Blame



Raw



Ethical Considerations in AI Models

Potential Biases

MNIST Handwritten Digits Model

While less prone to social biases like those found in language models, the MNIST dataset and models trained on it can exhibit **data collection bias**.

- **Bias Source:** If the dataset predominantly features handwriting from a specific demographic (e.g., adults, people from a certain region, right-handed individuals), the model might perform suboptimally on handwriting samples from underrepresented groups (e.g., children, left-handed individuals, different cultural writing styles). This could lead to unequal performance and potentially exclude users whose handwriting doesn't fit the 'norm' the model was trained on.

Amazon Product Reviews Model (Sentiment and NER)

This model, particularly the sentiment analysis, is highly susceptible to various biases.

- **Algorithmic Bias in NER:** The Named Entity Recognition (NER) model, even if using a pre-trained spaCy model, is trained on general text. It might struggle to identify niche product names, brand variations, or specific jargon common in product reviews from particular industries or communities. This could lead to **under-recognition** of entities relevant to certain products or user groups.
- **Sentiment Analysis Bias (Rule-Based):**
 - **Vocabulary Bias:** Our simple rule-based system relies on predefined positive/negative word lists. It may fail to capture sentiment expressed

through sarcasm, irony, slang, regional dialects, or nuanced expressions. For instance, "sick" can be negative in health contexts but positive as slang ("that's sick!").

- **Product Category Bias:** A word might have different sentiment implications depending on the product. "Hot" is positive for a new gadget but negative for a laptop that overheats. The rule-based approach won't differentiate this contextually.
- **Demographic Bias:** If certain user demographics (e.g., younger users, specific cultural groups, non-native English speakers) use language patterns, abbreviations, or emotional expressions differently, the model could misclassify their sentiment, leading to an inaccurate representation of their opinions.

Mitigation Strategies

Mitigating Bias with Tools (or approaches)

- **TensorFlow Fairness Indicators (for MNIST - Conceptual Application):**
 - While direct application for raw image data like MNIST is complex, if metadata about the writers (e.g., age group, gender, region) were available and associated with MNIST samples, TensorFlow Fairness Indicators could be used.
 - **How:** One would define sensitive groups based on this metadata. Fairness Indicators would then measure and visualize performance metrics (e.g., accuracy, false positive rates) across these groups. If disparities are found, it would signal a need for more diverse data collection or re-weighting of samples during training to ensure equitable performance. *For this specific assignment, its direct application for MNIST is limited without additional metadata.*
- **spaCy's Rule-Based Systems and Extensions (for Amazon Reviews):**
 - **Mitigating NER Bias:**
 - **Custom Rules/Matchers:** If the statistical NER model consistently misses specific product names or brand patterns (e.g., "XYZ-Pro Max"), we can augment it with spaCy's rule-based `Matcher` to explicitly recognize these. This allows for precise extraction of known entities regardless of the statistical model's performance on them.
 - **Fine-tuning:** For a more robust solution, fine-tuning a spaCy NER model on a diverse, domain-specific dataset of product reviews (annotated for entities) would significantly improve its performance and reduce bias towards general text patterns.
 - **Mitigating Sentiment Bias:**

- **Expanded & Contextual Lexicons:** Enhance the `positive_words` and `negative_words` lists to be more comprehensive and domain-specific. Incorporate multi-word expressions (e.g., "great value", "not working well").
- **Negation Handling:** Implement simple rules to reverse sentiment for negated words (e.g., "not good" should be negative). SpaCy's dependency parser could help identify negation tokens.
- **Part-of-Speech (POS) and Dependency Parsing:** Leverage spaCy's capabilities to understand the grammatical context. For instance, "hot" modifying a positive noun like "deal" vs. "hot" modifying a negative noun like "surface temperature." This moves beyond simple keyword matching to more sophisticated contextual analysis.
- **User Feedback & Iteration:** Continuously collect user feedback on sentiment classifications and use it to refine the rule-based system or train more advanced models.
- **Ensemble Approaches:** Combine the rule-based system with a pre-trained general sentiment model (like one from Hugging Face Transformers) and analyze where they differ, learning from their disagreements.