Natural Language Processing (NLP) and Ethics in AI

Summary Slides

Important Concepts

- NLP enables computers to understand, interpret, and generate human language
- Tokenization in NLP is the process of breaking down text into smaller units, such as words, subwords, or characters, which allows models to analyze and process language more effectively
- Tokenization allows unstructured data (text) to be structured and easy for a computer program to compute

Important Concepts

- Stop words are common words in a language (such as "the," "is," "in," and "and" in English) that are often filtered out in NLP tasks because they carry minimal semantic meaning

- Stop Words are sometimes removed in NLP because they do not carry significant meaning in general. But for some tasks such as Named Entity Recognition (NER) and others, stop words must be kept

- NER is an NLP task that identifies and classifies key information, like names of people, organizations, locations, and dates, within text

Important Concepts

- **Word embedding debiasing (soft debiasing)** refers to methods or techniques used to reduce bias in word embeddings, which are vector representations of words
- Word embedding debiasing reduces bias by adjusting word representations to weaken associations with sensitive attributes
- **Reduces bias** implies that these techniques aim to mitigate or eliminate biased associations, not directly improving the model prediction accuracy
- **Adjusting word representations** indicates that the word embeddings are modified
- **Weakening associations** means that the debiasing process specifically targets and reduces the influence of sensitive attributes (e.g., gender, race, etc.) in the word embeddings

Important Concepts

- **Soft** debiasing weakens biased associations, while **hard** debiasing removes them entirely
- Data augmentation creates more balanced training data by introducing diverse examples to reduce bias in models

Important Concepts

- **Adversarial training** is a technique in machine learning where a model is trained to become more robust by learning from adversarial examples—inputs that are deliberately modified to fool or challenge the model

- Example:

  - **Original Training Sentence**: "The scientist made a groundbreaking discovery."
  - **Adversarial Training Example**: "The homemaker made a groundbreaking discovery."

Important Concepts

- In general, **Fairness constraints** can be incorporated during training an AI model to promote balanced predictions for diverse demographic groups

- Example:

  – A task is to predict loan eligibility in a credit scoring system

  – Add a fairness constraint to ensure the model's predictions are equally accurate across demographic groups (e.g., men and women, or different racial groups

  – The fairness constraint prevents the model from being biased, especially if the data lacks diversity and has more eligible men than women in the training examples