TITLE: *Bias Detection and Mitigation in Text Classification.*

TEAM MEMBERS:

1. JANANI GANJI
2. SUSHMA NAGUBANDI
3. RAKESH MADISETTY

PROJECT TOPIC: *Bias Detection and Mitigation in Text Classification*

BRIEF DESCRIPTION: The project aims to investigate techniques to detect and mitigate biases in text classification models to promote fairness and equity.

Statement of Project Objectives

Objectives:

- Investigate existing approaches for bias detection and mitigation in text classification.
- Implement and evaluate different techniques for bias detection and mitigation.
- Apply the chosen techniques to a real-world text classification task to assess their effectiveness.
- Analyze the impact of bias detection and mitigation techniques on classification performance and fairness metrics.

Statement of Value

Importance of the Project:

- Addressing biases in text classification models is crucial for promoting fairness and equity in AI applications.
- Ensuring models are unbiased improves decision-making processes and reduces the risk of perpetuating harmful stereotypes.
- By mitigating biases, text classification systems can provide more accurate and equitable results for diverse user populations.

Review of the State of the Art

Bias Detection Techniques:

- Statistical Analysis: Methods such as statistical parity, demographic parity, and equal opportunity assess disparities across demographic groups in model predictions.
- Adversarial Attacks: Craft examples to expose biases by perturbing data, revealing differential treatment based on sensitive attributes.
- Fairness Metrics: Include disparate impact, equal opportunity difference, and demographic parity to quantify biases in mode predictions.

Bias Mitigation Strategies:

- Data Rebalancing: Techniques like oversampling or generating synthetic data address class imbalances and mitigate biases in training data.
- Adversarial Training: Involves training models with adversarial examples to enhance robustness against biases and adversarial attacks.
- Fairness-aware Learning: Algorithms explicitly optimize for fairness metrics during training to ensure equitable outcomes.

Interpretability and Explainability:

- Interpretability Methods: Aim to provide insights into model decisions and identify sources of bias.
- Techniques like LIME and SHAP offer local explanations for predictions, aiding in bias detection and mitigation.

Fairness Metrics:

- Disparate Impact: Measures disparities in predictions between different demographic groups.
- Equal Opportunity Difference: Quantifies differences in true positive rates between groups, focusing on fairness in positive predictions.
- Demographic Parity: Ensures consistent distribution of predictions across different demographic groups, promoting equal representation and treatment.

Citations of Key Papers and Studies:

- Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora necessarily contain human biases. Science, 356(6334), 183-186.
- Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. Advances in Neural Information Processing Systems, 29, 4349-4357.
- Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013). Learning fair representations. In International Conference on Machine Learning (pp. 325-333). PMLR.
- Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 259-268).

Brief Discussion of Limitations and Challenges:

- **Data Imbalance:** Imbalanced datasets can lead to biased models, as minority classes may be underrepresented, resulting in skewed predictions and unfair treatment.
- **Algorithmic Bias:** ethical implications associated with biased models, including perpetuating stereotBiases in training data or model architectures can perpetuate unfair treatment across different demographic groups, reinforcing existing societal biases and discrimination.
- **Ethical Considerations:** There are ypes, reinforcing inequalities, and potentially causing harm to disadvantaged groups.
- **Generalization:** Bias detection and mitigation techniques may not generalize well across diverse datasets and application domains, highlighting the need for robust and scalable approaches.

Overview of the Approach:

- **Data Preprocessing:**
  - ✓ Clean and prepare dataset.
  - ✓ Tokenize text and split dataset.

- **Baseline Model Training:**
  - ✓ Train initial models without bias mitigation.

- **Bias Detection:**
  - ✓ Identify biases in dataset and predictions.

- **Bias Mitigation:**
  - ✓ Apply techniques like adversarial training, data augmentation, fairness-aware learning.

- **Model Evaluation:**

  Assess performance using classification and fairness metrics.

Datasets, Models, Tools, and Techniques.

Datasets:

Selection of a Relevant Text Classification Dataset with Known Biases:

- The dataset should represent real-world text classification tasks and contain biases that need to be addressed.
- Example datasets include:

  IMDB Movie Reviews: Contains movie reviews labeled as positive or negative, with potential biases related to sentiment analysis.

  Twitter Sentiment Analysis: Consists of tweets labeled with sentiment, with potential biases related to language usage and cultural contexts.

  Toxic Comments Classification: Contains comments labeled as toxic or non-toxic, with potential biases related to demographic or social factors.

Models:

Implementing Baseline Classification Models and Integrating Bias Detection and Mitigation Techniques:

- Baseline models such as logistic regression, decision trees, or deep neural networks will be implemented for text classification tasks.
- Bias detection and mitigation techniques will be integrated into these models to address biases present in the dataset.
- Example techniques include:

  Adversarial Training: Modifying the training process to make the model robust against adversarial examples that highlight biases.

  Fairness-aware Learning: Incorporating fairness constraints into the model optimization process to promote fair outcomes.

  Data Rebalancing: Adjusting the class distribution in the training dataset to mitigate biases towards certain classes.

Tools:

- Python, TensorFlow/PyTorch, scikit-learn, Fairness Indicators and Aequitas.
  - ✓ Python will serve as the primary programming language for implementing models and conducting experiments.
  - ✓ TensorFlow or PyTorch will be used to build and train deep learning models for text classification.
  - ✓ scikit-learn will be utilized for implementing baseline machine learning models and conducting data preprocessing and evaluation.
  - ✓ Fairness Indicators and Aequitas are tools specifically designed for evaluating fairness in machine learning models and will be used to assess the fairness of the trained models.

Techniques:

Adversarial Training, Fairness-aware Learning and Data Augmentation.

- Adversarial Training involves training models with adversarial examples to improve robustness against biases.
- Fairness-aware Learning techniques explicitly optimize for fairness metrics during model training to mitigate biases and ensure equitable outcomes.
- Data Augmentation methods will be used to generate synthetic data to address class imbalances and mitigate biases present in the training data.

Deliverables

List of Deliverables:

- Implemented bias detection and mitigation techniques in text classification models.
- Evaluation results comparing model performance with and without bias mitigation techniques.
- Analysis of the effectiveness of bias mitigation approaches and their impact on classification performance and fairness.
- Documentation of the experimental setup, methodology, and findings.

Evaluation Methodology

- Evaluation Metrics:
  - ✓ Classification Metrics: Accuracy, precision, recall, F1 score.
  - ✓ Fairness Metrics: Equal opportunity difference, disparate impact, demographic parity.
- Methodology:
  - ✓ Cross-validation or train-test split for model evaluation.
  - ✓ Statistical tests to compare the performance of models with and without bias mitigation techniques.