Detecting Sexism in Text

Integrating Explainable AI and Contextual Intelligence

Pranshu Jain

NIT Calicut

April 30, 2025

Introduction

- Addressing bias in textual data is crucial for fairness in AI.
- Explainable AI (XAI) techniques help interpret model decisions.
- This research applies ANN, SNN, and BERT with explainability methods.

Bias in AI (Bolukbasi et al., 2016)

- Found that word embeddings (e.g., Word2Vec) inherit societal biases.
- Demonstrated gender bias in analogies (e.g., man is to computer programmer as woman is to homemaker).
- Proposed debiasing techniques to reduce bias in word embeddings.
- Highlighted the ethical concerns of biased AI models.

Counterfactual Explanations for Sexism Detection (Yang et al., 2023)

- Used counterfactuals to generate minimally edited text for model explanation.
- Showed how altering sexist words impacts model predictions.
- Improved fairness by identifying features that trigger sexist classifications.
- **Drawbacks:** Generating counterfactuals is difficult for longer texts.

DUTIR at SemEval-2023 Task 10: Semi-supervised Learning for Sexism Detection in English

- To enhance sexism detection using semi-supervised learning techniques. (ACL Anthology)
- Employed Unsupervised Data Augmentation (UDA) with the RoBERTa model, incorporating Easy Data Augmentation (EDA) for consistency training. (ACL Anthology)
- Drawbacks: The semi-supervised approach, while effective, may still require substantial labeled data for optimal performance and may not generalize well to all forms of sexist content.

Sarcasm Detection Using Feature-Variant Learning Models

- Detect sarcasm on social media using machine learning.
- Applied SVM, Decision Trees, Logistic Regression, Random Forest, KNN, and Neural Networks.
- Drawbacks: Limited to sarcasm detection; needs adaptation for sexism detection.

Artificial Neural Network (ANN)

- **Explanation:** ANN is a feedforward neural network that mimics human brain functionality.
- **Advantages:**
  - Efficient for structured data.
  - Can capture complex patterns.
  - Easy to implement and train.
- **Why It Is Used:** ANN helps in detecting sexism by learning patterns in text through multiple hidden layers.

Feature	**Description**
Architecture	Fully connected layers
Activation	ReLU, Softmax, Focal Loss
Training Method	Backpropagation
Performance	Moderate (73%)

Spiking Neural Network (SNN)

- **Explanation:** SNN mimics biological neurons and processes information through spikes.
- **Advantages:**
  - Energy-efficient computations.
  - Closer to human cognition.
  - Captures temporal dependencies in text.
- **Why It Is Used:** SNN helps analyze sexism detection with more biologically inspired interpretability.

Feature	Description
Neuron Model	LIF (Leaky Integrate and Fire)
Information Processing	Spikes instead of activations
Learning Rule	STDP (Spike-Timing-Dependent Plasticity)
Performance (68%)	Limited but interpretable

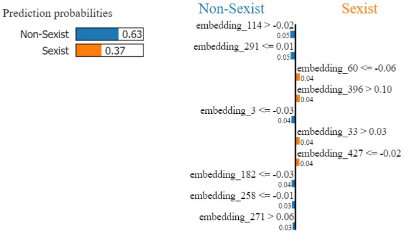
Bidirectional Encoder Representations from Transformers (BERT)

- **Explanation:** BERT is a deep transformer-based model trained on vast textual data.
- **Advantages:**
  - Superior contextual understanding.
  - Pretrained on large datasets.
  - High accuracy in NLP tasks.
- **Why It Is Used:** BERT effectively captures the nuanced meanings in sexist text for robust classification.

Feature	Description
Multitask Learning Architecture	Transformer-based
Tokenization	WordPiece Tokenizer
Redefine Loss Function with MC Dropout	Task Weighted CE
Performance	High accuracy (88%)

Explainability in ANN

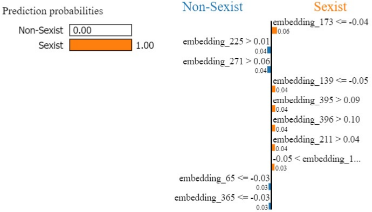
- **Why Explainability in ANN?**
  - ANN functions as a black box, making its decisions hard to interpret.
  - Explainability helps identify crucial words in sexism classification.



Feature	Value
embedding_114	-0.00
embedding_291	0.00
embedding_60	-0.07
embedding_396	0.10
embedding_3	-0.05
embedding_33	0.03
embedding_427	-0.03
embedding_182	-0.09
embedding_258	-0.06
embedding_271	0.07

Explainability in SNN

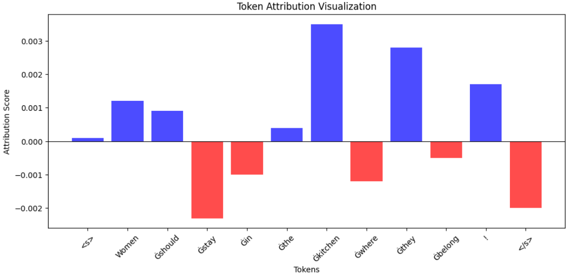
- **Why Explainability in SNN?**
  - SNN processes information using spikes, making feature attribution complex.
  - Explainability helps in analyzing spike-based decision patterns.



Feature	Value
embedding_173	-0.06
embedding_225	0.04
embedding_271	0.08
embedding_139	-0.06
embedding_395	0.10
embedding_396	0.10
embedding_211	0.07
embedding_110	-0.04
embedding_65	-0.04
embedding_365	-0.05

Explainability in BERT

- **Why Explainability in BERT?**
  - BERT's attention mechanisms make its decision-making process opaque.
  - Explainability helps in understanding how words influence classification.



Conclusion

- **Key Takeaways:**
  - Explainable AI enhances trust and transparency in sexism detection.
  - ANN and SNN provide interpretability using LIME and SHAP.
  - BERT's complex decision-making is explained using Integrated Gradients.
- **Challenges:**
  - Ensuring explainability without compromising model performance.
  - Handling ambiguous and context-dependent sexist language.
- **Future Directions:**
  - Improving explainability techniques for deep learning models.
  - Extending the dataset for better generalization.
  - Exploring multimodal approaches for sexism detection in text and speech.

[1] Xavier Ferrer, Tom van Nuenen, Jose M. Such, Mark Coté, and Natalia Criado.
Bias and discrimination in ai: a cross-disciplinary perspective.
In *arXiv preprint arXiv:2008.07309*, 2020.

[2] Author(s) Name.
Sarcasm detection using feature-variant learning models.
In *Conference/Journal Name*, Year.

[3] Indira Sen, Mattia Samory, Claudia Wagner, and Isabelle Augenstein.
Counterfactually augmented data and unintended bias: The case of sexism and hate speech detection.
In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2022.

[4] Bingjie Yu, Zewen Bai, Haoran Ji, Shiyi Li, Hao Zhang, and Hongfei Lin.
Dutir at semeval-2023 task 10: Semi-supervised learning for sexism detection in english.
In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, 2023.