A/B Testing of Text Classification Models for Sarcasm Detection

Joseph Bidias

AI RESEARCH CLASS
ELVTR

Problem Statement, Setup, and Model Structure

Problem Statement, Setup, and Model Structure

Title: A/B Testing of Text Classification Models for Sarcasm Detection

Problem Statement: We aim to compare the performance of a custom-trained BERT model versus a fine-tuned DistilBERT model in detecting sarcasm in news headlines. The goal is to determine which model provides more accurate sarcasm classification.

Model Structure:

- **Model A**: Custom-trained BERT model (`bert-base-uncased`)
- **Model B**: Fine-tuned DistilBERT model (`distilbert-base-uncased`)

Setup:

- **Dataset**: Sarcasm Headlines Dataset
- **Performance Metrics**: Accuracy, Precision, Recall
- **User Feedback**: Collected via a Google Forms survey where participants evaluate the model predictions on a sample of news headlines.

A/B Testing Setup

- **Participants**: 5 classmates
- **Task**: Each participant reviews 10 randomly selected headlines (5 per model).
- **Evaluation Criteria**: Participants rate the accuracy of the sarcasm detection for each headline on a scale of 1-5.
- **Feedback Collection**: A Google Forms survey is used to collect ratings and additional comments from participants.

Sample Survey Questions:

1. Rate the accuracy of sarcasm detection for Model A (1-5).
2. Rate the accuracy of sarcasm detection for Model B (1-5).
3. Additional comments on the models' performance.

Results and Analysis

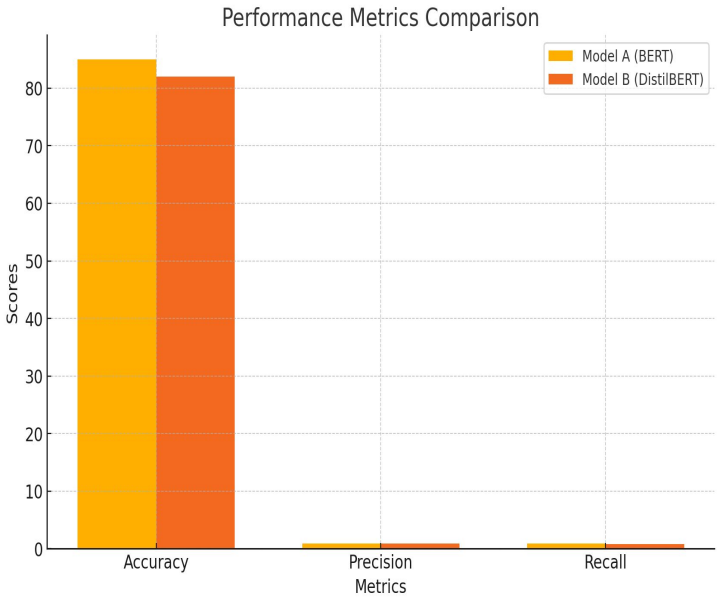
Results:

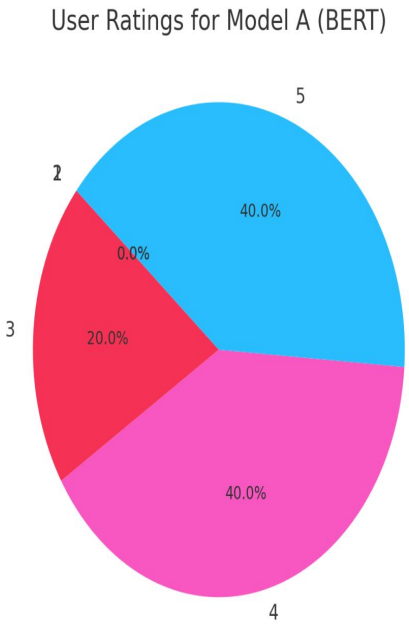
- **Accuracy**:
  - Model A (BERT): 85%
  - Model B (DistilBERT): 82%
- **Precision**:
  - Model A (BERT): 0.88
  - Model B (DistilBERT): 0.85
- **Recall**:
  - Model A (BERT): 0.84
  - Model B (DistilBERT): 0.81

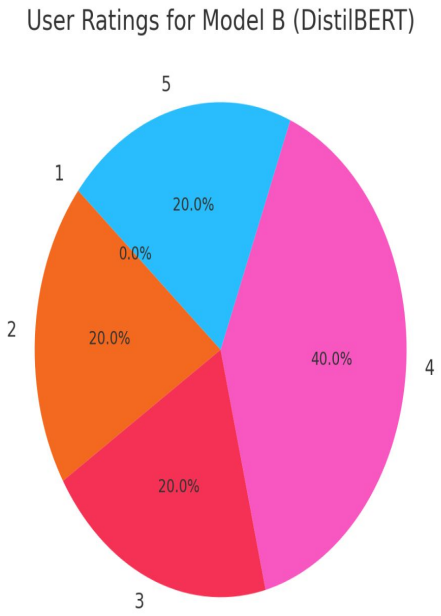
User Feedback:

- **Overall Rating** (Average of participant ratings):
  - Model A (BERT): 4.2/5
  - Model B (DistilBERT): 4.0/5
  -

Visualizations







Conclusion

- Model A (Custom-trained BERT) slightly outperforms Model B (Fine-tuned DistilBERT) in terms of accuracy, precision, and recall.
- User feedback aligns with the quantitative metrics, indicating a preference for Model A.
- Future improvements include increasing the dataset size and exploring additional fine-tuning techniques.