# Title: Balancing Data and Model Optimization in NLP Sentiment Analysis

---

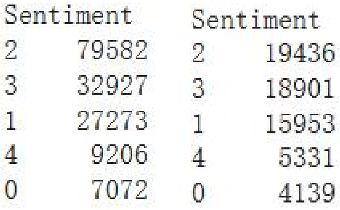
## 1. Introduction

- **Objective**: To evaluate the performance of NLP models (BERT and GPT-2) on a sentiment analysis task under different data preprocessing and weighting strategies.
- **Dataset**: Sentiment dataset with 5 classes, where class distribution is highly imbalanced:
  - Largest class: 79,582 samples (Neutral)
  - Smallest class: 7,072 samples (Negative)

---

## 2. Methodology

### Data Cleaning:



```
Sentiment        Sentiment
2    79582       2    19436
3    32927       3    18901
1    27273       1    15953
4     9206       4     5331
0     7072       0     4139
```

- Removed special characters, stop words.
- Dropped rows with less than two words which are neutral sentiments.
- Goal: Balance class distribution and reduce noise.

### Model Optimization:

- Baseline models: BERT and GPT-2 (on cleaned and uncleaned datasets).
- Implemented class-weighted loss for BERT to address class imbalance.
- Evaluation Metrics:
  - **Accuracy**: Overall correctness.
  - **F1-Score**: Balance between precision and recall.
  - Weighted and Macro-averaged metrics for better class comparison.

---

## 3. Results and Observations

### Performance Comparison Table

| Model | Accuracy | Macro F1 | Weighted F1 | Observation |
|---|---|---|---|---|
| BERT (Cleaned) | 63.93% | 60% | 64% | Cleaning reduced overall accuracy. |
| GPT-2 (Cleaned) | 61.14% | 56% | 61% | Performed slightly worse than BERT. |
| BERT (Uncleaned) | 69.80% | 63% | 70% | Uncleaned data retained useful sentiment info. |
| GPT-2 (Uncleaned) | 68.75% | 59% | 69% | Close to BERT but consistently lower. |
| BERT (Weighted) | 70.04% | 63% | 70% | Class-weighting helped balance performance. |

## 4. Analysis

**Data Cleaning**:

- While it aimed to remove noise, essential sentiment indicators (like short texts) were inadvertently removed.
- Uncleaned data preserved these features, leading to better performance.

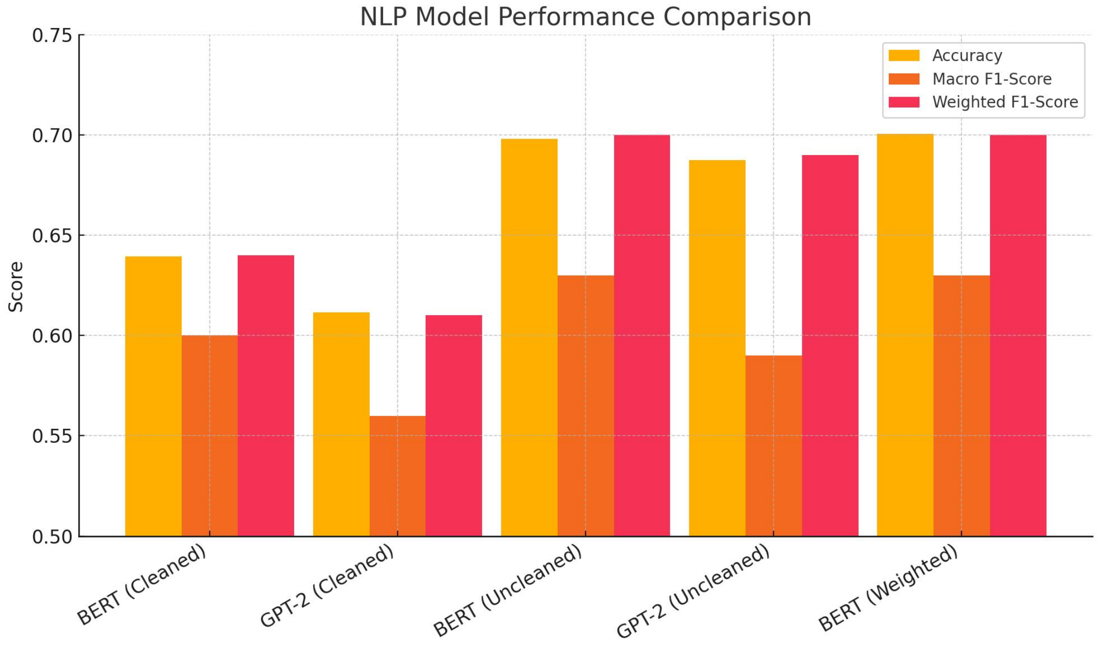
**Class-Weighting**:

- Improved a little bit BERT's ability to balance class performance, particularly for minority classes.
- No significant improvement in overall F1-Score

**Model Characteristics**:

- **BERT**: Outperformed GPT-2 due to better handling of context and smaller classes.
- **GPT-2**: Lagged behind, likely due to its preference for longer context and generative training objectives.

## 5. Conclusion



- **Best Model**: BERT with class-weighting on uncleaned data performed the best with **70.04% accuracy** and balanced metrics.
- **Insights**:

    1. **Data Cleaning**: Must be carefully designed to retain sentiment-relevant features.
    2. **Class Imbalance**: Addressed effectively through class-weighted loss functions.
    3. **Model Choice**: BERT's contextual understanding outperformed GPT-2 in this task.