Data-Efficient Text Classification for Low-Resource Indian Languages: A Case Study on Gujarati

---

1. Introduction

Natural Language Processing (NLP) systems often rely on large annotated datasets, which are unavailable for many Indian regional languages. Gujarati, spoken by millions, remains a low-resource language in NLP research. This project explores data-efficient learning strategies for Gujarati text classification by comparing traditional machine learning approaches with modern multilingual transformer-based models.

---

2. Dataset Description

The dataset consists of Gujarati movie reviews annotated with three sentiment classes represented as numeric labels. Due to limited availability of publicly labeled Gujarati datasets, the project operates under low-resource constraints, reflecting real-world challenges in multilingual NLP.

---

3. Methodology

A classical TF-IDF with Logistic Regression model was implemented as a baseline. For advanced modeling, a multilingual transformer architecture was selected to leverage cross-lingual transfer learning. The models were trained and evaluated using standard train-test splits, with performance measured using accuracy and F1-score.

---

4. Experiments and Results

The baseline model achieved an accuracy of approximately 70%. Transformer-based models demonstrated improved performance, achieving accuracy of approximately 78% during training attempts. These results highlight the benefits of multilingual pre-training for low-resource languages.

---

5. Limitations

Model training was constrained by limited computational resources available in the experimental environment. As a result, transformer fine-tuning was conducted with reduced data and limited epochs. Despite these constraints, meaningful performance improvements were observed.

6. Conclusion and Future Work

This study demonstrates that multilingual transformers can improve performance for low-resource Gujarati NLP tasks. Future work may explore data-centric approaches, improved annotation strategies, and more efficient model architectures to further enhance performance.