Emotion Detection in Persian Text using Machine Learning

Student Name
Shahid Beheshti University
Machine Learning Fundamentals – 3rd Assignment

June 2024

Abstract

This report presents a machine learning approach for detecting emotions in Persian text, categorized into five classes: happiness, sadness, anger, fear, and others. The methodology encompasses text preprocessing, feature engineering using TF-IDF, training with classical models such as logistic regression, and evaluating the results using precision, recall, and F1-score. The pipeline demonstrates the challenges and insights in multilingual NLP, especially on imbalanced data.

1 Introduction

Text-based emotion recognition is a crucial component in natural language understanding, with applications in sentiment analysis, mental health, and human-computer interaction. This report investigates a classical machine learning approach for detecting emotions in Persian texts. The dataset includes labeled samples with single-label classification into five emotional states.

2 Data Preprocessing

2.1 Loading and Normalization

The dataset was first loaded and Persian text normalization was applied to standardize characters and remove diacritics.

2.2 Cleaning

The following cleaning steps were applied:

- Removal of punctuation, numbers, and special characters.
- Tokenization into words.
- Removal of Persian stopwords.

2.3 Feature Engineering

Texts were vectorized using TF-IDF, and labels were encoded using `LabelEncoder`.

3 Exploratory Data Analysis (EDA)

The first step in EDA involved examining class distributions and word frequencies. As seen in Figure ??, the dataset is imbalanced, with most samples belonging to HAPPY and OTHER.

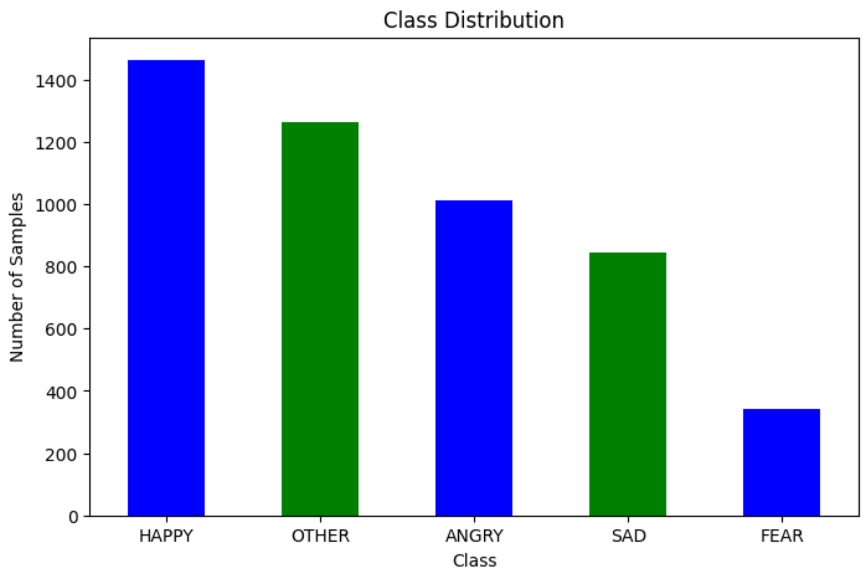

Figure 1: Class Distribution

The most frequent words (appearing more than 100 times) are shown in Figure ??, helping us understand linguistic patterns in emotional texts.

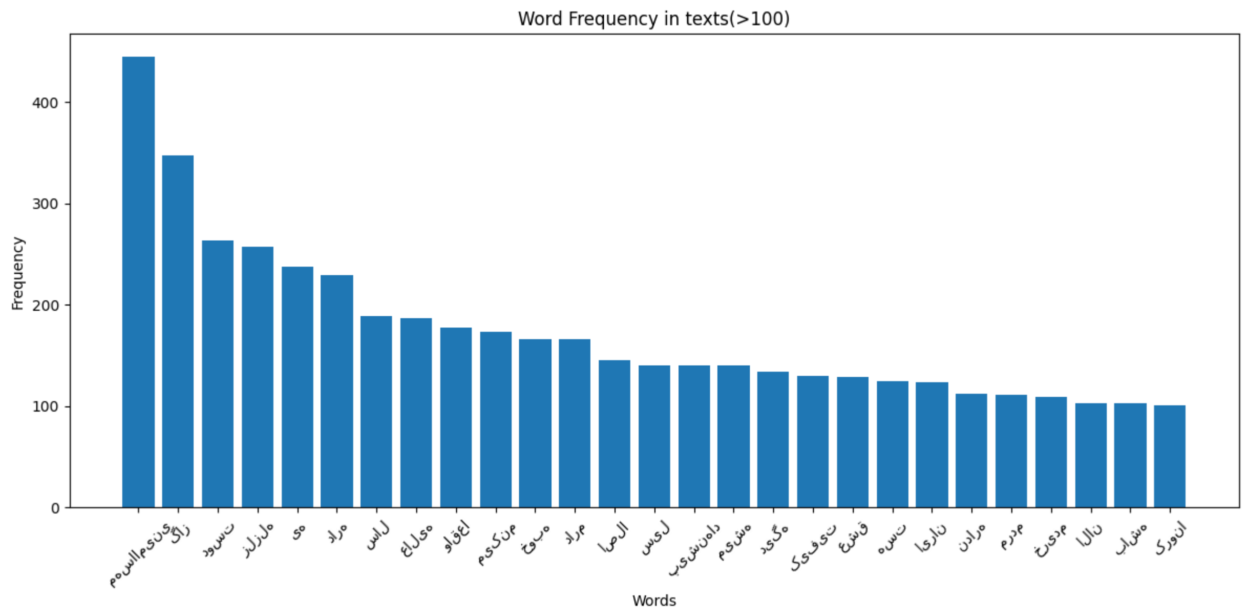

Figure 2: Word Frequency in Texts (Count ¿ 100)

4 Modeling and Classification

We chose logistic regression as our baseline classifier due to its efficiency on sparse data (TF-IDF). The dataset was split into training and test sets, and 5-fold cross-validation was used for evaluation stability. Further, other models (e.g., decision trees, XGBoost) were explored, but initial results focused on logistic regression for clarity.

5 Model Evaluation

The evaluation metrics include precision, recall, F1-score, and overall accuracy.

Table 1: Evaluation Report (Logistic Regression)

Class	**Precision**	**Recall**	**F1-score**	**Support**
ANGRY	0.00	0.00	0.00	185
FEAR	0.00	0.00	0.00	66
HAPPY	0.31	0.98	0.47	306
OTHER	0.71	0.04	0.08	267
SAD	0.00	0.00	0.00	161
Accuracy		0.32		
Macro Avg	0.20	0.21	0.11	985
Weighted Avg	0.29	0.32	0.17	985

6 Error Analysis

6.1 Class Imbalance

The classifier is biased towards the HAPPY class due to its dominance in training samples. Other classes like FEAR and SAD are significantly underrepresented.

6.2 Feature Limitation

TF-IDF does not capture sequence information or context well, which might limit performance on subtle emotional distinctions.

7 Model Monitoring and Boosting

To enhance model performance, boosting methods like XGBoost were explored. Training and validation error curves (Figure ??) show overfitting after several boosting rounds, emphasizing the importance of early stopping and regularization.

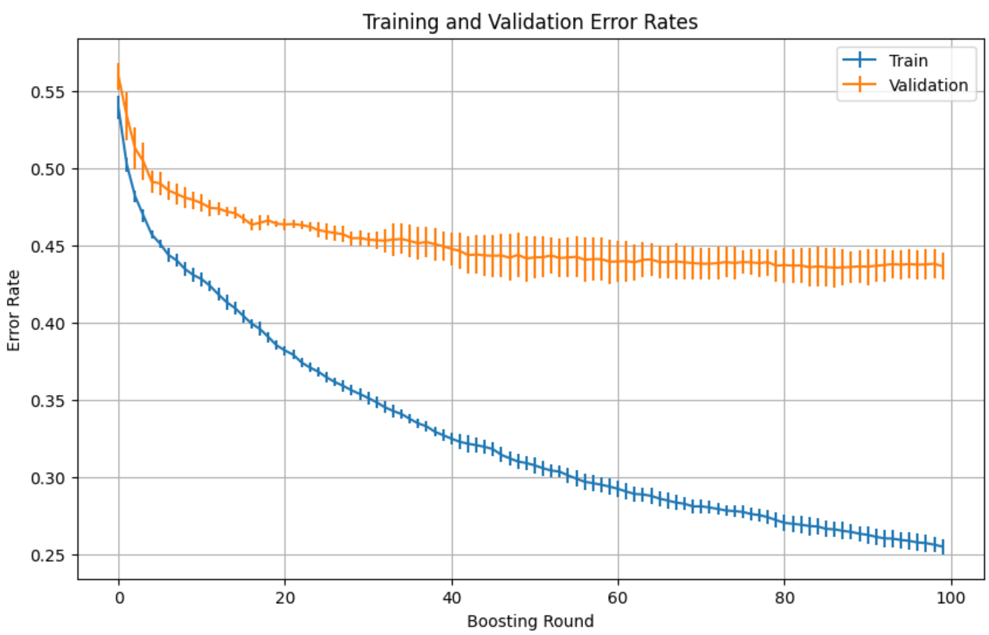

Figure 3: Training and Validation Error over Boosting Rounds

8 Conclusion and Future Work

This project implemented a basic yet insightful emotion detection model for Persian text using classical ML techniques. While initial results were limited by class imbalance and basic features, it laid a solid foundation. In future work, we aim to:

- Use transformer-based models like BERT with fine-tuning on Persian text.

- Apply SMOTE or other augmentation methods for balancing classes.
- Explore sequence models (e.g., RNN, LSTM) with word embeddings.

Appendix

- `output.png` – Word frequency histogram
- `output2.png` – Class distribution chart
- `output3.png` – Boosting error analysis