LEARNING FROM DATA – FINAL PROJECT REPORT

Multi-Class Classification of Consumer Complaints Using Machine Learning and Deep Learning Approaches

1. Introduction

1.1 Motivation

With the rapid growth of online platforms and digital customer services, large volumes of user-generated complaint texts are produced every day. Automatically categorizing these complaints is important for organizations to respond efficiently, prioritize requests, and improve customer satisfaction. Manual processing is time-consuming and error-prone, which makes machine learning-based text classification an effective solution.

This project focuses on building a multi-class text classification system that automatically assigns consumer complaints to predefined categories.

1.2 Problem Definition

The problem is defined as a multi-class text classification task. Given a complaint text in Turkish, the goal is to classify it into one of the following categories:

Banking

Telecommunication

Transportation

1.3 Project Objectives

The main objectives of this project are:

To collect and preprocess real-world complaint data through web scraping

To apply and compare multiple machine learning and deep learning models

To analyze model performance using appropriate evaluation metrics

To select the most robust model and analyze its generalization ability

2. Data Collection & Preprocessing

2.1 Data Collection

The dataset was collected through web scraping from an online complaint platform (Şikayetvar – sikayetvar.com) using requests and BeautifulSoup. The scraper iterates over category listing

pages and then visits each complaint's detail page to extract the complaint title and the full complaint description text. Each record was saved together with its source URL to ensure traceability and to enable duplicate removal.

A browser-like User-Agent header was used, and randomized waiting times were added between requests in order to reduce server load and avoid being blocked. Network timeouts were used to make the process robust against temporary connection problems. Duplicate records were removed using the Link field (URL) as a unique identifier.

During early data collection, an additional category (Kargo) was also scraped. However, since the number of Kargo samples was lower than other categories and could introduce class imbalance, this category was removed in the final version of the dataset. The final dataset contains 2382 complaint texts distributed across three balanced categories:

Category	Count
Banking	803
Telecommunication	802
Transportation	777

The dataset is balanced, which reduces the need for class imbalance handling techniques.

2.2 Data Preprocessing

Several preprocessing steps were applied to clean and normalize the Turkish text data:

Conversion to lowercase

Removal of numbers

Removal of punctuation and special characters

Removal of single-character tokens (e.g., "a", "b")

Stopword removal using a manually curated Turkish stopword list (embedded in the code)

Removal of very short and meaningless outputs after cleaning

The cleaned texts were stored in a separate column (Temiz_Metin) and used for model training.

3. Methodology

3.1 Feature Engineering

TF-IDF Representation

The main feature representation used in this project is TF-IDF (Term Frequency–Inverse Document Frequency) with unigrams and bigrams. The maximum number of features was limited to 5000 to control dimensionality:

ngram_range = (1, 2)

max_features = 5000

To prevent data leakage, TF-IDF was fitted only on the training set and then applied to validation and test sets using only transformation.

Dimensionality Reduction

For deep learning experiments, Truncated Singular Value Decomposition (TruncatedSVD) was applied to reduce dimensionality and improve computational efficiency:

n_components = 300

3.2 Models Used

Traditional Machine Learning Models

Logistic Regression

Linear Support Vector Machine (Linear SVM / LinearSVC)

Random Forest

Deep Learning Model

Multi-Layer Perceptron (MLP) using TF-IDF + SVD features

3.3 Training Strategy

Train / Validation / Test Split

The dataset was split using a stratified 70/15/15 split:

70% Training

15% Validation

15% Test

All splits were stratified to preserve class distribution and used a fixed seed for reproducibility (random_state = 42).

Cross-Validation & Learning Curve

To evaluate robustness, 5-fold stratified cross-validation was performed with shuffling:

StratifiedKFold(n_splits=5, shuffle=True, random_state=42)

Metric: Macro F1 (f1_macro)

A learning curve was also generated for the selected final model using Macro F1, showing how training and cross-validation performance changes as the training set size increases.

Evaluation Metrics

The primary evaluation metrics were:

Accuracy

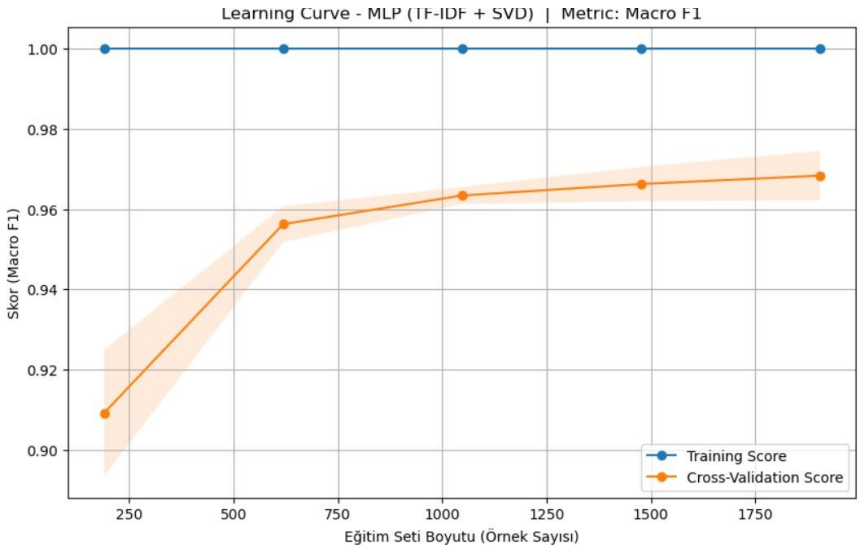
Macro F1-score (preferred for multi-class evaluation because it treats classes equally)

4. Results & Analysis

4.1 Baseline Model Results (Logistic Regression)

Metric	Value
Validation Accuracy	96.36%
Test Accuracy	94.41%
Train Accuracy	98.56%

The small gap between training and validation accuracy indicates that the baseline model does not suffer from severe overfitting.



4.2 Model Comparison

Model	Validation Accuracy	Validation Macro F1

Logistic Regression	0.9636	0.9638
Linear SVM	0.9748	0.9749
Random Forest	0.9776	0.9777
MLP (TF-IDF + SVD)	0.9832	0.9833

The MLP model achieved the best validation Macro F1 score and was selected as the final model.

4.3 Final Model Test Performance

Final Model: MLP (TF-IDF + SVD)

Test Accuracy: 95.81%

Test Macro F1: 0.9585

The confusion matrix indicates that most misclassifications occur between Banking and Telecommunication categories, which often share similar complaint language (e.g., billing problems, service dissatisfaction, customer support issues). Transportation complaints tend to contain more domain-specific terms (e.g., “sefer”, “bilet”, “iptal”), which improves separability.

4.4 Cross-Validation & Learning Curve

5-fold cross-validation produced an average Macro F1 score of 97.04% with a standard deviation of ±0.49%, indicating strong robustness and stable performance across folds.

The learning curve analysis shows that training and cross-validation scores converge as data size increases, suggesting a good bias–variance tradeoff and improved generalization with more data.

5. Error Analysis

Manual testing with unseen complaint texts demonstrated that the model correctly classifies clear domain-specific complaints. However, ambiguous complaints containing generic negative expressions or mixed-domain contexts may lead to misclassification, which is expected in real-world scenarios. Common error cases include short complaints with limited context and complaints that mention multiple services (e.g., telecom billing paid via bank card).

6. Discussion

6.1 Interpretation of Results

The experimental results show that the deep learning model (MLP) combined with TF-IDF and dimensionality reduction outperforms traditional machine learning approaches. A likely reason is

that SVD converts sparse TF-IDF vectors into a compact dense representation, enabling the MLP to learn more flexible, non-linear decision boundaries.

6.2 Bias–Variance Analysis

Learning curves and cross-validation results indicate that the selected model achieves a good balance between bias and variance. The model generalizes well without evidence of severe overfitting.

6.3 Limitations and Future Work

TF-IDF is limited in capturing deeper semantic meaning.

Incorporation of word embeddings (Word2Vec, GloVe, FastText) could improve semantic understanding.

Transformer-based models (e.g., BERT variants) could be explored for potentially higher accuracy.

More categories and larger datasets could be added to increase task complexity and test scalability.

7. Conclusion

This project demonstrates the application of machine learning and deep learning techniques for multi-class classification of Turkish consumer complaints. Complaint texts were collected through web scraping, cleaned using a structured preprocessing pipeline, and represented with TF-IDF features. Multiple models were trained and compared using stratified splitting, Macro F1 evaluation, cross-validation, and learning curves.

The selected final model, MLP with TF-IDF + SVD, achieved high validation performance and strong test results, showing that the proposed approach is robust and effective for real-world complaint classification.