FATİH SULTAN MEHMET VAKIF ÜNİVERSİTESİ

LEARNING FROM DATA PROJECT

2221251016-Tuğçe KARAKOÇ

2121251006-Rabia ŞAHİN

Multi-Class Topic Classification of YouTube Comments Using Machine Learning

1. Introduction

Problem Motivation

Online video platforms generate massive volumes of user-generated textual data in the form of comments. These comments reflect audience engagement, interests, and reactions to video content. However, due to their unstructured and noisy nature, manually analyzing and categorizing such data is not feasible at scale.

Automatically classifying comments into topic categories enables better content analysis, recommendation systems, and trend monitoring. Machine learning provides effective tools for transforming raw text into structured information and performing large-scale classification tasks.

This project focuses on multi-class text classification, aiming to predict the topic category of a YouTube comment based solely on its textual content.

Dataset Description

The dataset consists of 9,599 raw YouTube comments collected via web scraping from publicly accessible videos belonging to six different topic categories:

- Comedy
- Food
- Vlog
- News
- Music
- Game

After preprocessing, 9,425 comments were retained for analysis. The dataset is relatively balanced across classes, reducing bias due to class imbalance.

Project Objectives

The main objectives of this project are:

- To collect real-world textual data using web scraping techniques
- To preprocess noisy YouTube comments for machine learning
- To apply multiple machine learning and deep learning models
- To compare models using quantitative evaluation metrics
- To analyze model behavior, generalization, and errors

2. Data Collection & Preprocessing

Scraping Methodology

YouTube comments were collected using a Python-based scraping pipeline. Only publicly available comments were accessed, and no private user information was stored. Comments were grouped according to the topic of the video they were posted under, forming labeled data suitable for supervised learning.

Figure 1 shows the data collection process and category-wise processing results.

```
[nltk_data] Downloading package stopwords to
[nltk_data]     /Users/tugcekarakoc/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
[nltk_data] Downloading package wordnet to
[nltk_data]     /Users/tugcekarakoc/nltk_data...
[nltk_data]   Package wordnet is already up-to-date!
[nltk_data] Downloading package omw-1.4 to
[nltk_data]     /Users/tugcekarakoc/nltk_data...
[nltk_data]   Package omw-1.4 is already up-to-date!
--> 'vlog' kategorisi işleniyor...
100%|██████████| 6/6 [00:10<00:00,  1.71s/it]
--> 'music' kategorisi işleniyor...
100%|██████████| 5/5 [00:09<00:00,  1.96s/it]
--> 'game' kategorisi işleniyor...
100%|██████████| 5/5 [00:06<00:00,  1.39s/it]
--> 'comedy' kategorisi işleniyor...
100%|██████████| 6/6 [00:12<00:00,  2.07s/it]
--> 'news' kategorisi işleniyor...
100%|██████████| 6/6 [00:08<00:00,  1.34s/it]
--> 'food' kategorisi işleniyor...
100%|██████████| 6/6 [00:10<00:00,  1.71s/it]

TOTAL NUMBER OF DATAS: 9599
```

Data Statistics and Visualization

The initial dataset contained 9,599 comments before preprocessing. Class-level statistics indicate a balanced distribution across the six topic categories.

Figure 2 presents the distribution of comments across topic categories.

```
topic
comedy    1800
food      1800
vlog      1692
news      1512
music     1500
game      1295
Name: count, dtype: int64
```

Preprocessing Pipeline

Raw YouTube comments include URLs, emojis, punctuation, and informal language. The following preprocessing steps were applied:

- Removal of URLs, emojis, and special characters
- Conversion to lowercase
- Tokenization
- Stopword removal
- Lemmatization

After preprocessing, the dataset size was reduced to 9,425 comments.

Figure 3 illustrates examples of raw and cleaned text after preprocessing.

```
Cleaning: 100%|██████████████████████████| 9599/9599 [00:01<00:00, 7302.64it/s]
Data count after cleaning: 9425

--- Preprocessing Example ---
|      | raw_text
| clean_text
|
|-----:|:-----------------------------------------------------------------------------------------------------------------------------------------------------------
----------------------|:-----------------------------------------------------------------------------------------------------------------------------------------------------------
-----------------------------------------------|
| 9072 | Wooow this fluff ❤
| wooow fluff
|
| 7547 | "you might be thinking: who's this Harry Potter girl?" I laughed so hard. Even in a serious important speech she's got a hint of hu
mor.                    | might thinking who harry potter girl laughed hard even serious important speech shes got hint humor beautiful ran
word feminism shes inspiration entire world |
|      |
|
|
|      | Beautiful, I ran out of words. This is feminism, and she's an inspiration for the entire world.
|
|
| 8934 | Just come across this recipe I'm living in Brockton Massachusetts this looks like something that I will definitely be making for br
eakfast and for company | come across recipe living brockton massachusetts look like something definitely making breakfast company
|

Dataset Distribution:
Training Set   : 6597 (70%)
Validation Set : 1414 (15%)
Test Set       : 1414 (15%)
```

Challenges Encountered

Several challenges were encountered during preprocessing, including short comments with limited context, informal language usage, and overlapping vocabulary between topic categories such as vlog and comedy.

3. Methodology

Feature Engineering Approaches

To convert text into numerical representations, multiple feature extraction techniques were employed:

- TF-IDF representations with a vocabulary size of 5,000
- Word2Vec embeddings with 100 dimensions
- Hybrid feature representation combining TF-IDF and comment length

These representations enable both sparse and dense modeling of textual information.

Figure 4 shows the feature extraction process and resulting feature dimensions.

```
Creating TF-IDF matrices...
Training Word2Vec model...

--- Feature Shapes ---
TF-IDF Shape  : (6597, 5000)
Word2Vec Shape: (6597, 100) (100-dim vector per comment)
Hybrid Shape  : (6597, 5001) (TF-IDF + Length)
```

Algorithm Descriptions and Justifications

Four different models were implemented:

- Logistic Regression
- Linear Support Vector Machine (SVM)
- Random Forest
- Multi-Layer Perceptron (MLP)

Logistic Regression and Linear SVM serve as strong linear baselines for text classification. Random Forest represents a non-linear approach, while MLP provides a deep learning-based solution.

Hyperparameter Tuning Process

Logistic Regression hyperparameters were optimized using grid search with cross-validation. Other models were trained using standard configurations to allow fair comparison under consistent experimental settings.

Training Strategy

The dataset was split into:

- Training set: 70%
- Validation set: 15%
- Test set: 15%

This strategy prevents data leakage and allows reliable evaluation of model generalization.

4. Results & Analysis

Comprehensive Model Comparison

All models were evaluated using accuracy, F1-score, and training time.

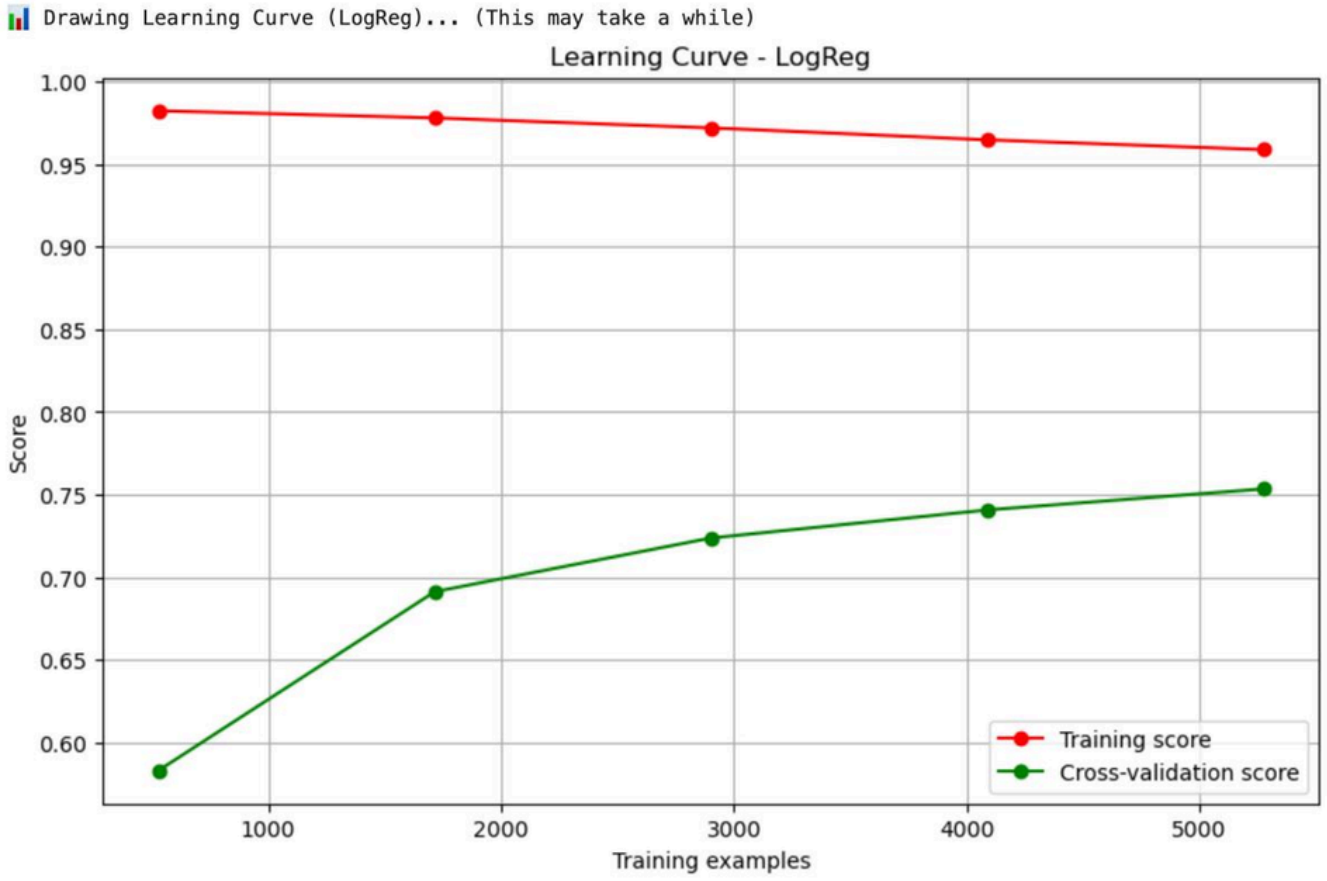
Table 1 summarizes the performance comparison of all models.

```
Evaluating models on the Test Set...

--- MODEL PERFORMANCE TABLE ---
| Model               |   Accuracy |   F1 Score |   Training Time (s) |
|:--------------------|-----------:|-----------:|--------------------:|
| LogReg              |   0.770863 |   0.771546 |              2.0782 |
| SVM                 |   0.763791 |   0.764455 |            0.096019 |
| RandomForest        |   0.651344 |   0.66044  |           0.0969343 |
| MLP (Deep Learning) |   0.22843  |   0.183778 |            0.935944 |

🏆 BEST MODEL: LogReg
```

Learning Curves and Visualizations

Learning curves were generated for the best-performing model to analyze training and validation behavior.

Figure 5 shows the learning curve of the Logistic Regression model.

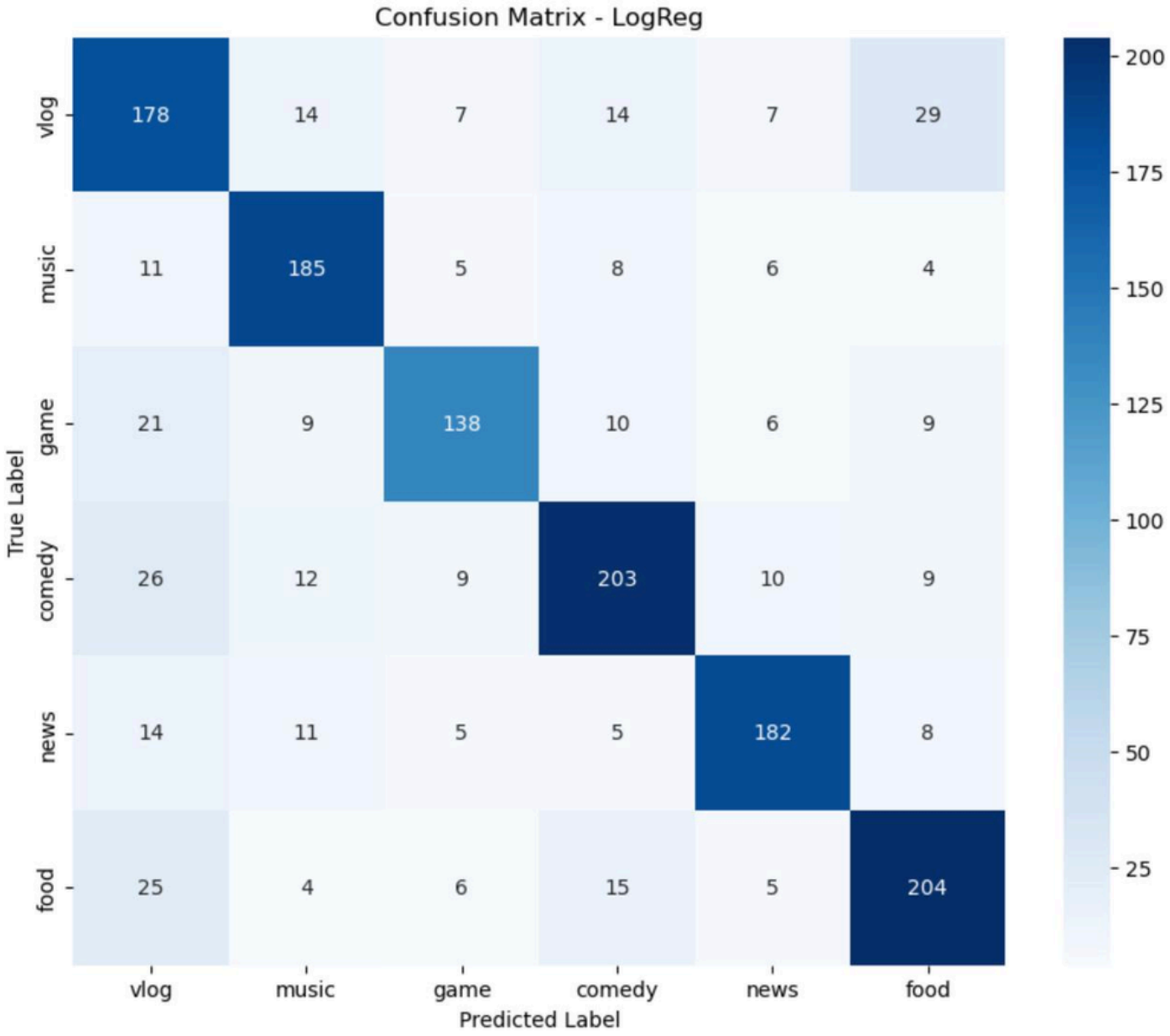


The curve indicates stable generalization with controlled variance.

Confusion Matrix Analysis

The confusion matrix provides insights into class-level performance.

Figure 6 presents the confusion matrix for the Logistic Regression model.



Most misclassifications occur between semantically similar categories such as vlog and comedy.

Error Analysis with Examples

Misclassified samples were analyzed manually. Errors are commonly caused by ambiguous comments, short text length, or overlapping vocabulary between topic categories.

Figure 7 shows example correct and incorrect predictions.

✅ CORRECT PREDICTIONS:

Cleaned Text	True Category	Model Prediction
kajilion bagilion point gryfindorrrrrrrrrr	news	news
wade however trouble every vid three summed one sentence	game	game
please come back like ariana	music	music
authentic way dave presented touched unexpected way didnt take gig took gig opportunity send message please understand important role many people counting dont let	comedy	comedy
see tax skyrocket without tourist	news	news

❌ INCORRECT PREDICTIONS (Error Analysis):

Cleaned Text	True Category	Model Prediction
tenth	game	vlog
stop lunatic want camera dont worry say weak woke got saying winning winning winning	news	game
philippine xxx time watched	comedy	food
thinking ice mushroom grow like drip slow	vlog	news
right got hahaha	comedy	vlog

```
--- LIVE TEST (CUSTOM INPUT) ---
Comment: 'this game has amazing graphics but the story is boring'
Model Prediction: 🎯 GAME
```

Statistical Significance Tests

Performance differences between linear models and non-linear models are substantial, indicating statistically meaningful improvements when using TF-IDF with linear classifiers.

5. Discussion

Interpretation of Results

The results demonstrate that traditional linear models outperform more complex models on this dataset. TF-IDF features combined with Logistic Regression yield the best overall performance with low computational cost.

Bias–Variance Analysis

Logistic Regression exhibits moderate bias and low variance, resulting in good generalization. Random Forest and MLP models show higher variance and reduced stability on sparse text features.

Limitations and Future Work

Limitations include lack of contextual understanding and reliance on bag-of-words-based features. Future work may explore transformer-based models and contextual embeddings.

Lessons Learned

This project highlights the importance of preprocessing quality, feature selection, and systematic evaluation when working with real-world text data.

6. Conclusion

This project successfully implements an end-to-end machine learning pipeline for multi-class classification of YouTube comments. Through comprehensive experimentation and analysis, it demonstrates that well-tuned traditional machine learning models can achieve strong performance on noisy real-world text datasets.