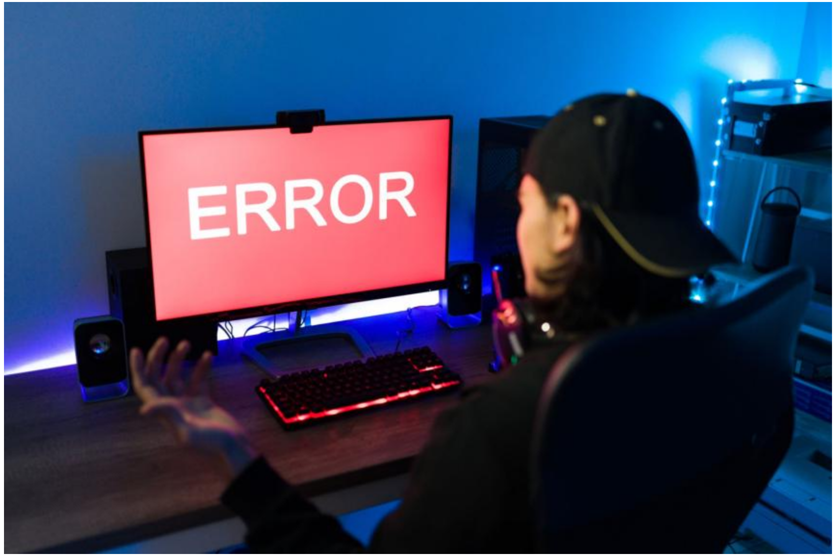
# Feature-based Sentiment Analysis For Games

Demba SOW

# Abstract



This study focuses on **aspect-based sentiment analysis of Turkish user reviews for digital games**. The **goal is to classify sentiments (positive/negative/neutral)** for **specific features** such as **graphics, AI, and gameplay**, even within complex sentences (e.g., “The graphics are good, but the AI is bad”). A language filter ensures only Turkish reviews are analyzed.

The model was trained on the labeled dataset and tested using NLP techniques for feature extraction and machine learning for sentiment classification, this approach aims to provide detailed insights for game developers and improve user feedback evaluation.

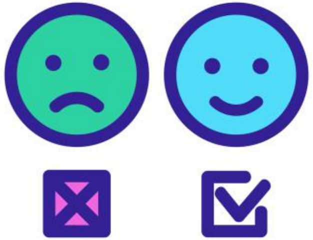
# Introduction

The digital gaming industry has evolved into a massive sector, reaching billions of users worldwide. With increasing competition, user reviews have become a critical source of feedback for developers.

Sentiment analysis, a technique within the field of Natural Language Processing (NLP), aims to classify the emotional tone or subjective opinions expressed in texts. This approach can identify emotions such as joy, anger, and frustration, and categorizes text as positive, negative, or neutral.

Online reviews related to games provide direct insights into players’ experiences, expectations, and dissatisfaction. However, Turkish—a morphologically rich and agglutinative language—poses specific challenges for sentiment analysis. The complexity introduced by its affixation structure and contextual nuances requires sophisticated methods that can effectively capture these intricacies.

This study aims to develop a model that segments Turkish game reviews according to predefined aspects such as graphics, artificial intelligence (AI), and gameplay, and classifies the sentiment for each aspect as either positive, negative or neutral.



# Literature Review

- **English literature:** Lexicon-based methods like VADER, optimized for social media, achieve around **65–75%** accuracy. Machine learning models such as Naïve Bayes, Logistic Regression, SVM, and Random Forest report accuracies near **82.5%**. Deep learning models like BiLSTM and CNN offer improved contextual understanding, reaching up to **91%** accuracy. Transformer-based models, especially BERT, show top performance with scores of **91.39%**, **92%**, and even **98%** in some cases. In contrast, Turkish literature is limited, mostly focusing on general sentiment analysis. A major challenge is the lack of large, labeled Turkish datasets, which hinders further advancement.
- **Data Sources:**Platforms like Steam, Google Play Store, X (Twitter), and Reddit provide large-scale user-generated reviews.
- **Turkish Literature:**More limited in scope. Most studies focus on general sentiment analysis, with few comprehensive works on aspect-based approaches. There is a lack of large, labeled Turkish datasets.
- **Challenges:**
    - Sarcasm, irony, domain-specific jargon (e.g., "OP", "grind", "nerf"), and game-related terminology.
    - Morphological complexity due to Turkish being an agglutinative language.
    - Imbalanced data distribution and presence of multilingual content.
    - Use of slang and abbreviated in-game expressions.

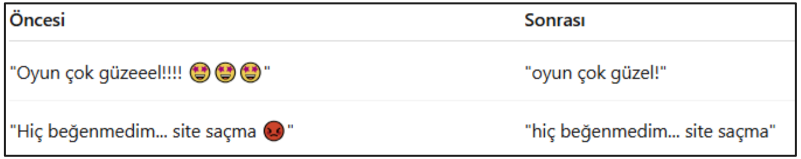
# Data Scrapping

- Approximately **2,100 user reviews from 35 games** across various Steam categories such as action, strategy, and simulation **were collected using a Python-based web scraping approach.**
- Utilizing the unique game IDs and the official Steam API with Turkish language settings, detailed game information was gathered, including game title, genre, description, price, release date, supported platforms, developer, game tags, and system requirements.
- An equal number of reviews were sourced from different categories to ensure diversity in content and linguistic usage, minimizing biases arising from individual writing styles, and resulting in a balanced and broadly representative dataset.

# Data Preprocessing

- Multiple consecutive spaces were reduced to single spaces, and unnecessary whitespace at the beginning and end of lines was removed.
- HTML/Markdown residues, special characters, and lines irrelevant to the dataset were eliminated.
- Auxiliary and unnecessary columns were removed to maintain structural consistency within the dataset.
- Punctuation marks were largely retained, while repetitive character sequences within words were normalized.
- Texts were converted to lowercase, and words were lemmatized according to Turkish morphological rules using **Zemberek, a Turkish NLP library.**
- The language of reviews was automatically identified using the **langdetect library**, and reviews in languages other than Turkish were filtered out, ensuring linguistic consistency.
- Profanity and slang expressions were intentionally preserved in the dataset due to their potential strong emotional significance.

➡️ **Objective: To achieve a clean yet realistic text structure suitable for sentiment analysis.**



| Öncesi | Sonrası |
|---|---|
| "Oyun çok güzeeel!!!! 😍😍😍" | "oyun çok güzel!" |
| "Hiç beğenmedim... site saçma 😡" | "hiç beğenmedim... site saçma" |

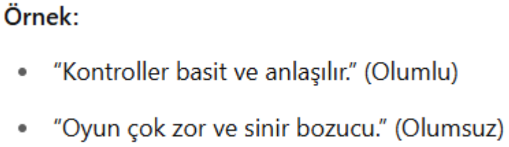
# Data Labeling

A total of **2,100 cleaned reviews were manually labeled** by three researchers.

The reviews were assigned to **one or more aspects** based on their content:

- **Gameplay**
- **Artificial Intelligence (AI)**
- **Graphics**

Reviews not directly related to any of these three categories were classified according to their **overall sentiment** as either positive or negative.



Örnek:

- "Kontroller basit ve anlaşılır." (Olumlu)
- "Oyun çok zor ve sinir bozucu." (Olumsuz)

**Labeling Criteria;**

- **Oynanış:**

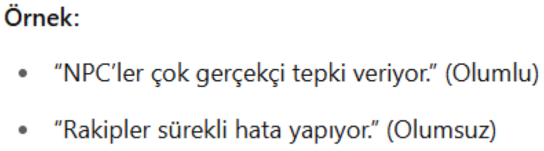
Elements affecting user interaction, including game mechanics, control systems, and difficulty levels.
Labels: **Olumlu**, **Olumsuz**, **Nötr**

- **Yapay Zekâ (AI):**

Comments related to NPC behavior, opponents' reactions, and learning algorithms..
Labels: **Olumlu**, **Olumsuz**, **Nötr**

- **Grafik:**

Comments concerning visual quality, animation quality, art style, and user interface elements.
Labels: **Olumlu**, **Olumsuz**, **Nötr**



Örnek:

- "NPC'ler çok gerçekçi tepki veriyor." (Olumlu)
- "Rakipler sürekli hata yapıyor." (Olumsuz)

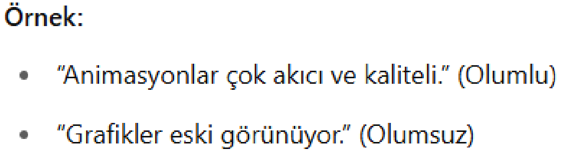
**Labeling System;**

Each review was labeled according to:

- **Aspect** (Gameplay, AI, Graphics)
- **Sentiment** (Positive, Negative)

**Labeling Schema:**

  - **ASPECT-KEYWORDS:**
    Graphics, AI, Gameplay
  - **SENTIMENT-WORDS:**
    Positive, Negative



Örnek:

- "Animasyonlar çok akıcı ve kaliteli." (Olumlu)
- "Grafikler eski görünüyor." (Olumsuz)

# Labeling Examples

- **Cümle** = Dünyanın en iyi 5 oyunundan biri
- **Genel Duygu**: Olumlu
- **Yapay Zeka**: Nötr
- **Grafik**: Nötr
- **Oynanış**: Olumlu

- **Cümle** = “Oyunda ilk başta çok zorlandım ancak oyunu çözdüğüm zaman çok zevk aldım adeta başyapıt”
- **Genel Duygu**: Olumlu
- **Yapay Zeka**: Nötr
- **Grafik**: Nötr
- **Oynanış**: Olumlu

- **Cümle** = “Atmosfer çok iyi ama boss’lar çok güçlü hep öldüm”
- **Genel Duygu**:Nötr
- **Yapay Zeka**: Olumsuz
- **Grafik**: Olumlu
- **Oynanış**:Olumsuz

- **Cümle** = “Çok rezalet oyun sakın almayın”
- **Genel Duygu**: Olumsuz
- **Yapay Zeka**: Nötr
- **Grafik**: Nötr
- **Oynanış**: Olumsuz

- **Cümle** = “Rakiplerin tepkileri çok akıllıca, savaşlar heyecan verici ama görseller biraz sıradan kalmış.”
- **Genel Duygu**: Olumlu
- **Yapay Zeka**: Olumlu
- **Grafik**: Olumsuz
- **Oynanış**: Olumlu

- **Cümle** = “Görsel atmosfer büyüleyici, ama optimizasyon berbat ve kontroller gecikmeli.”
- **Genel Duygu**:Olumsuz
- **Yapay Zeka**: Nötr
- **Grafik**: Olumlu
- **Oynanış**:Olumsuz

# Models And Algorithms

In this project, two different models for sentiment analysis of Turkish game reviews were implemented and compared:

**1. BERT-Based Sentiment Analysis Model**

- **Model**: bert-base-turkish-cased (HuggingFace Transformers)
- **Training Objective**: general sentiment classification (Positive, Negative, Neutral)
- **Advantages**:
  - Processes contextual information bidirectionally
  - Demonstrates high performance on morphologically rich Turkish
  - Architecture is extensible for feature-level analysis
- **Disadvantages**:
  - Substantial computational cost
  - Large parameter footprint (GPU requirement)
- **Flexibility**: enables versatile analysis by training dedicated models for each game feature.

⚖️ **Benchmark Role**: BERT delivers state-of-the-art performance through its contextual strength.

**2. TF-IDF + Logistic Regression Baseline**

- **Vectorization**: unigram-based TF-IDF (limited to 2,000 features)
- **Model**: MultiOutputClassifier + Logistic Regression
- **Training**: simultaneous multi-label (gameplay, AI, graphics) and multi-class (positive, negative, neutral) prediction
- **Advantages**:
  - Fast and resource-efficient
  - Interpretable architecture
- **Disadvantages**:
  - Captures contextual semantics only superficially
  - Fails to detect subtle affective nuances

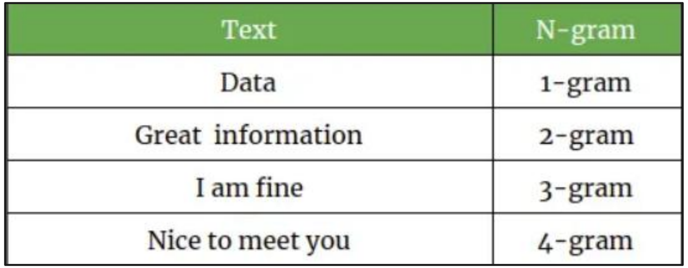
⚖️ **Benchmark Role**: TF-IDF + LR provides a simple yet effective baseline for comparative evaluation.

The outputs of both models were evaluated for both overall **sentiment and feature-based analysis.**

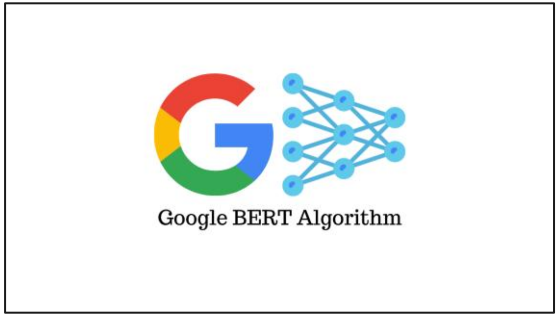
# Traning and Evaluation

The BERT model was fine-tuned with `train_bert_model(train_df, val_df, num_labels=3)`, saving intermediate checkpoints, and the N-gram–based model was trained with `train_ngram_model(train_df, val_df, output_dir, aspect_cols, ngram_range=(1,1), max_features=2000)`, using TF-IDF and logistic regression, then saving both the model and the vectorizer.

The BERT model is evaluated on the validation set with accuracy, F1, and a confusion matrix, while the n-gram–based model—using `evaluate_ngram_model`—reports feature-level accuracy, macro-F1, precision, and recall. For visualization, the BERT model's training/validation loss and accuracy curves are plotted; the n-gram model's feature-level metrics are shown as bar charts; and finally, sample predictions from both models are compared with the true labels.

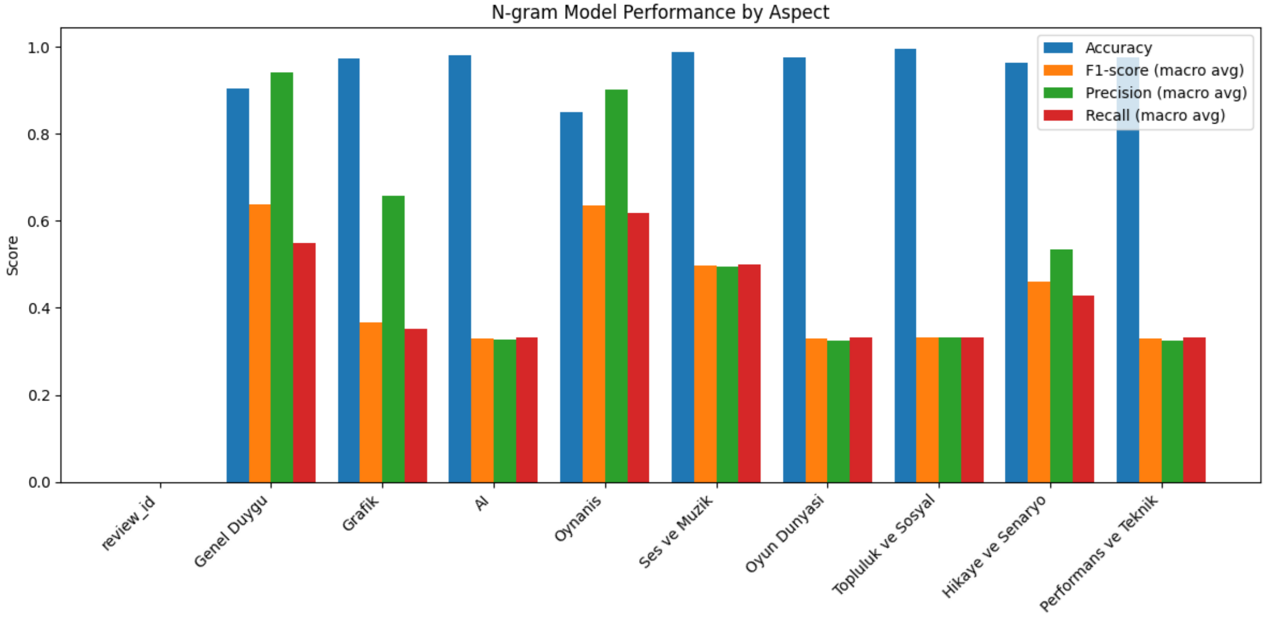


| Text | N-gram |
|---|---|
| Data | 1-gram |
| Great information | 2-gram |
| I am fine | 3-gram |
| Nice to meet you | 4-gram |

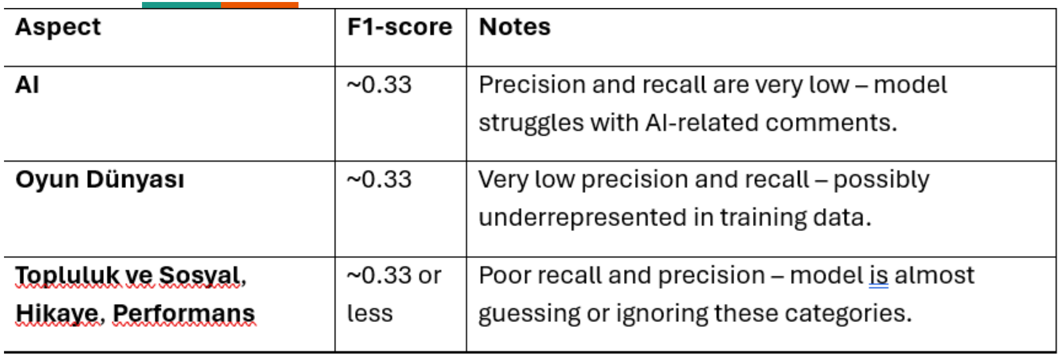


# About of Ngram

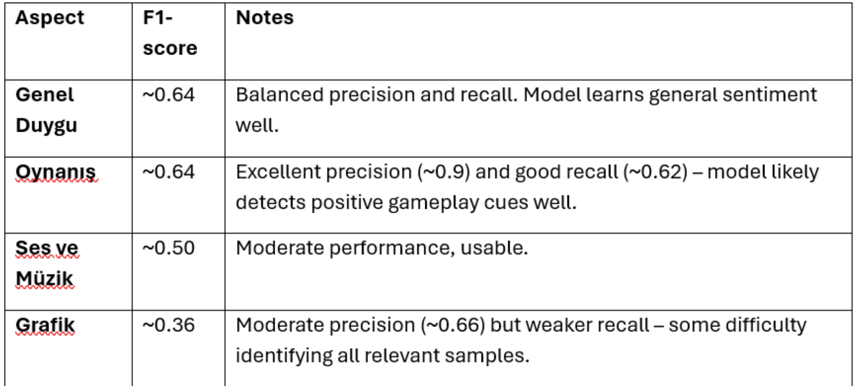
The model’s performance was evaluated using Accuracy,Macro F1-Score, Macro Precision, and Macro Recall for each dimension. The key strengths and weaknesses are summarized below:



# About of Ngram continued



| Aspect | F1-score | Notes |
|---|---|---|
| **AI** | ~0.33 | Precision and recall are very low – model struggles with AI-related comments. |
| **Oyun Dünyası** | ~0.33 | Very low precision and recall – possibly underrepresented in training data. |
| **Topluluk ve Sosyal, Hikaye, Performans** | ~0.33 or less | Poor recall and precision – model is almost guessing or ignoring these categories. |



| Aspect | F1-score | Notes |
|---|---|---|
| **Genel Duygu** | ~0.64 | Balanced precision and recall. Model learns general sentiment well. |
| **Oynanış** | ~0.64 | Excellent precision (~0.9) and good recall (~0.62) – model likely detects positive gameplay cues well. |
| **Ses ve Müzik** | ~0.50 | Moderate performance, usable. |
| **Grafik** | ~0.36 | Moderate precision (~0.66) but weaker recall – some difficulty identifying all relevant samples. |

Many reviews have been labeled as “Neutral” for most dimensions. The limited number of positive and negative examples makes it difficult for the model to learn meaningful patterns for these classes.

Dimensions such as AI, Story, and Community are often not explicitly mentioned in the text. In such cases, accurate classification is particularly challenging for models that do not capture contextual meaning (e.g., n-gram-based models).

The n-gram TF-IDF representation used is sparse and lacks contextual awareness. Therefore, semantically rich expressions such as “The story was complex” or comments containing irony/sarcasm cannot be analyzed correctly.

# About of BERT

**What is BERT?**

- Pre-trained language model by Devlin et al.
- Uses bidirectional Transformer architecture
- Captures context effectively (syntax, semantics, sentiment)

**Advantages:**

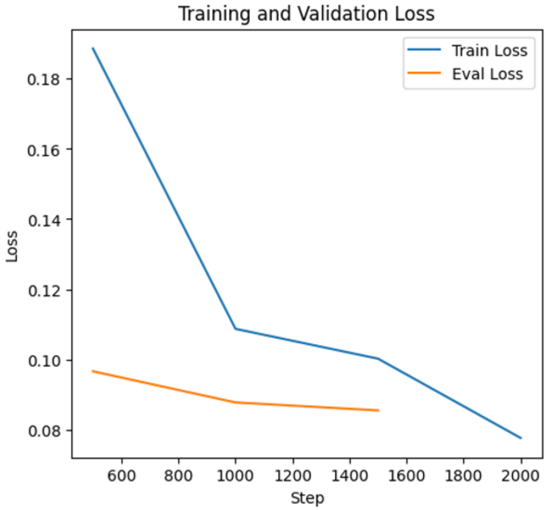
- High contextual understanding
- Adaptable to various NLP tasks
- Effective in languages like Turkish

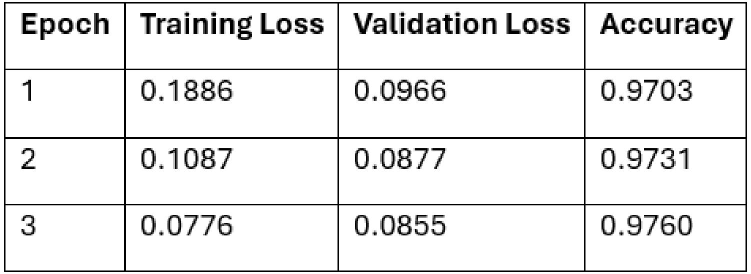
**Challenges:**

- High computational cost
- Reduced interpretability

**Training & Evaluation:**

- Training loss: 0.1886 → 0.0776 (epoch 1 to 3)
- Validation loss: 0.0966 → 0.0855
- Accuracy reached 97.6%





| Epoch | Training Loss | Validation Loss | Accuracy |
|---|---|---|---|
| 1 | 0.1886 | 0.0966 | 0.9703 |
| 2 | 0.1087 | 0.0877 | 0.9731 |
| 3 | 0.0776 | 0.0855 | 0.9760 |

# Conclusion

## Ngram

To improve model performance on underrepresented labels like *AI* and *Community*, class weighting or data balancing techniques (e.g., SMOTE, stratified sampling) can be employed. Since n-gram TF-IDF lacks contextual understanding, using contextual embeddings such as BERT, RoBERTa, or Sentence-BERT is recommended for better sentiment capture.

Training separate models for each aspect can reduce multi-label complexity and enhance accuracy. Additionally, increasing labeled examples—particularly for low-frequency aspects—and analyzing false predictions can reveal overlooked linguistic cues, guiding both preprocessing and model design improvements.

## BERT

In this study, we evaluated the BERT model's performance on text classification tasks using a Transformer-based architecture. The results underline BERT's effectiveness in handling text data, especially in scenarios involving limited labeled datasets. Variants such as DistilBERT and RoBERTa, known for efficiency and performance, align with our findings.

**Future Work Recommendations:**

1. Extended Training: Increase training duration to 5–6 epochs.
2. Fine-Tuning Strategies: Experiment with layer freezing, layer-wise learning rates, and learning rate scheduling techniques.
3. Generalization Tests: Conduct evaluations across different domains and languages to ensure robustness.
4. Lightweight Models: Explore knowledge distillation and quantization for deployment scenarios with latency constraints.

Overall, BERT provides a robust and reliable foundation for meeting text processing needs. For access to the complete code used in this project, please visit the provided link.

# THANK YOU

# FOR

# LİSTENİNG!