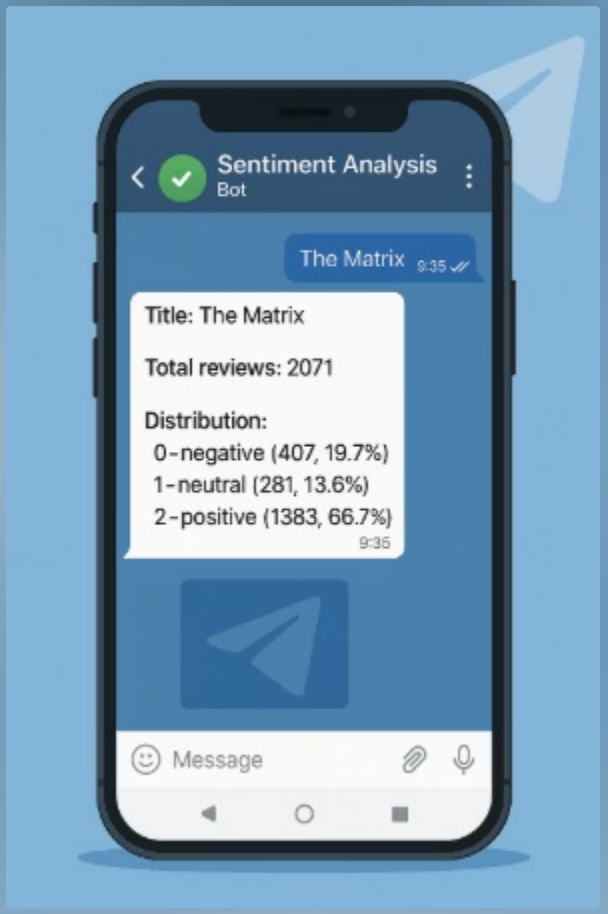
Sentiment Analysis of Russian Movie Reviews



Project Overview

Goal

Classify Russian movie reviews into three classes (positive, neutral and negative)

Data

Corpus of over 130,000 Russian movie reviews

Preprocessing

Cleaning, tokenization, and stemming applied for text normalization

Metrics

Evaluated using accuracy, precision, recall, and F1-score

Dataset

Total number of reviews: 131583

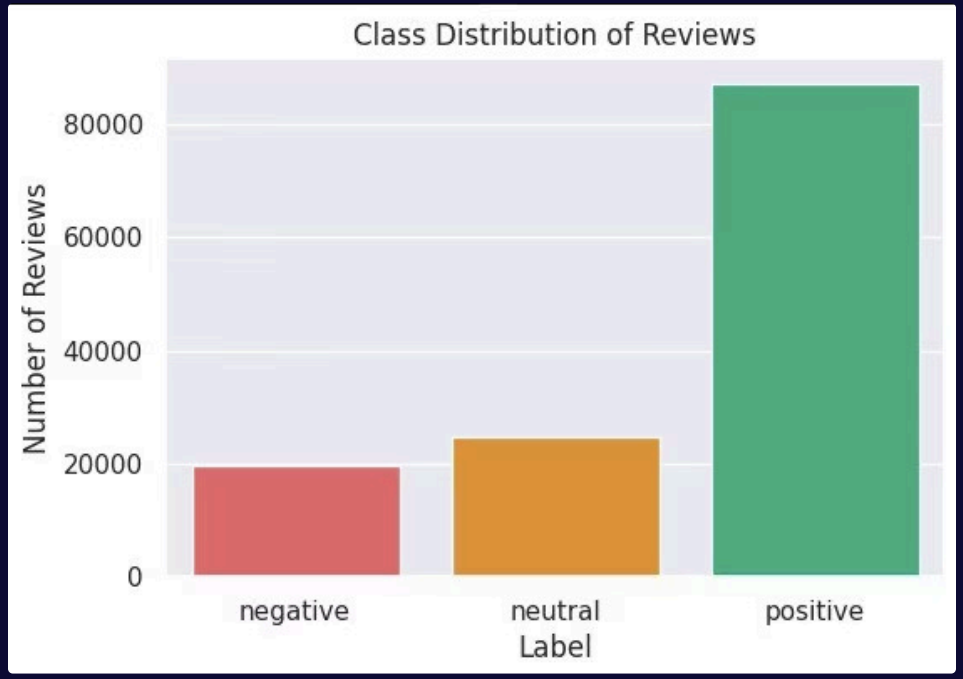
Positive: 87101

Negative: 19804

Neutral: 24678

Language: Russian

Subject: movie review



Feature Extraction: Embeddings

TF-IDF

Sparse, baseline embedding capturing word importance

Dim: 2000

BERT (RuBERT)

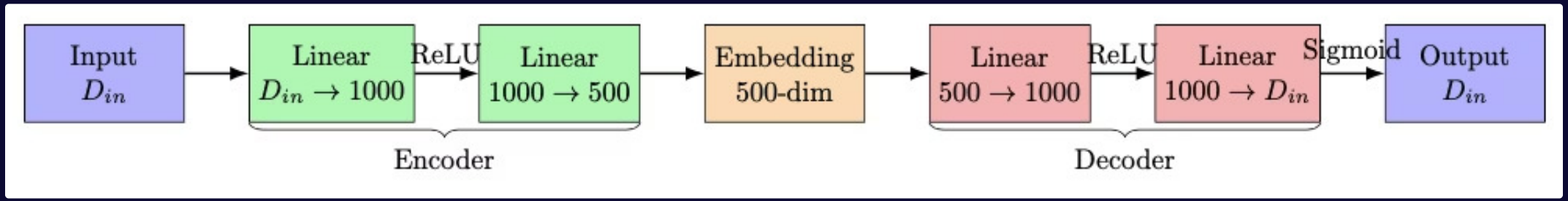
Contextual embeddings capturing semantic nuances; fine-tuned

Dim: 312

Autoencoder

Latent space reduces dimensions; extracts key features efficiently

Dim: 500



Baseline Predictive Models

- Logistic Regression
- Decision Tree
- **Simple Neural Network**

  Three hidden layers with ReLU activations

Advanced Predictive Models

BiLSTM

Bidirectional LSTM captures text sequence dependencies

- **Stacked Bidirectional LSTM** layers with:
  - **Residual connections** to mitigate vanishing gradients.
  - **Layer normalization** between LSTM steps.

BiLSTM + CNN

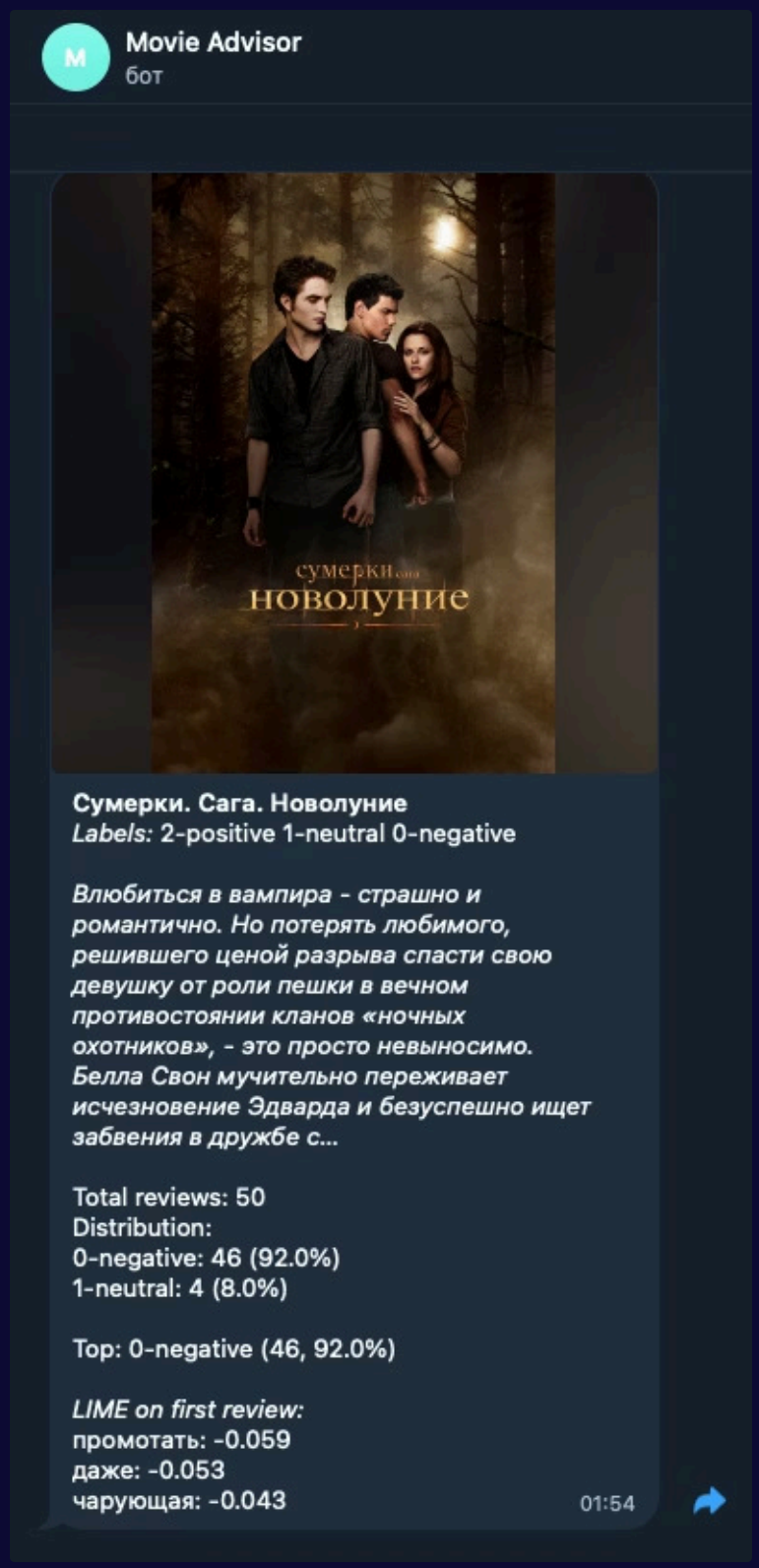
Combines sequential context with local feature extraction

Key Components

1. **BiLSTM Layer**:
   - Processes input embeddings bidirectionally to model long-range dependencies.
   - Output: Contextualized features (`hidden_dim=512`, bidirectional → output size `2*hidden_dim`).
2. **Multi-Filter CNN**:
   - Applies **parallel 1D convolutions** with varying filter sizes (`[1, 3, 5]`) to detect local n-gram patterns.
   - Uses `n_filters=100` per filter size, padding to preserve sequence length.

Results: Model Performance

Model	Embedding	Accuracy	Precision	Recall	F1-Score
Logistic Regression	TF-IDF	0.68	0.72	0.68	0.69
Decision Tree	TF-IDF	0.55	0.57	0.55	0.56
Simple NN	TF-IDF	0.68	0.69	0.68	0.68
Logistic Regression	BERT	0.66	0.73	0.66	0.68
Decision Tree	BERT	0.57	0.57	0.57	0.57
Simple NN	BERT	0.64	0.69	0.64	0.65
BiLSTM	BERT	0.70	0.68	0.70	0.69
BiLSTM+CNN	BERT	0.73	0.74	0.71	0.72
Logistic Regression	Autoencoder	0.66	0.73	0.66	0.69
Decision Tree	Autoencoder	0.48	0.64	0.48	0.52
Simple NN	Autoencoder	0.65	0.73	0.65	0.67
BiLSTM	Autoencoder	0.67	0.73	0.67	0.69



Telegram Bot

- **Enter movie title**
  “The user types a film name—any title works.”
- **Search via API**
  “The bot queries the KinoPoisk API for matching titles and displays the top options as buttons.”
- **Fetch up to 50 reviews**
  “Once a movie is selected, up to 50 recent user reviews are downloaded automatically.”
- **Classify & display results**
  “Each review is embedded and classified. Finally, the bot sends back a summary showing how many reviews fell into each sentiment category, plus the most frequent label.”

Analysis and Insights

Embedding Impact

BERT embeddings notably boost model performance.

Best Model

BiLSTM+CNN with BERT achieved 73% accuracy.

Error Patterns

Common misclassifications linked to ambiguous reviews.

Conclusion and Future Work

1

Summary

Advanced models and embeddings substantially improve sentiment prediction.

2

Future Directions

- Explore more pre-trained BERT models with higher dimension
- Experiment with diverse autoencoder designs
- Try another class imbalance handling technique for better accuracy