UNIVERSITAS
GADJAH MADA
Sentiment Analysis of Steam Reviews Using BERT and XLNet
Pasquale Dominic Ligamen Saleppang, Prof. Drs. Edi Winarko, M.Sc.,Ph.D
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

ABSTRACT

The rapid growth of the video game industry and the increasing volume of user generated reviews on digital platforms such as Steam have made automated sentiment analysis an essential tool for extracting meaningful insights from large scale player feedback. Traditional machine learning approaches, including Logistic Regression, Support Vector Machine (SVM), and Naïve Bayes, have been widely applied to this task but are limited by their reliance on bag-of-words feature representations that cannot capture contextual meaning, sarcasm, negation, and gaming specific language. Transformer based models offer a promising advancement, yet a direct comparison between BERT and XLNet — two models with fundamentally different pre-training strategies — on Steam game reviews has not previously been conducted.

This study performs a comparative analysis of BERT and XLNet for binary sentiment classification of Steam reviews, evaluating both models against three traditional machine learning baselines. A dataset of 813,062 English language reviews was collected from 60 games released between 2020 and 2024 using the Steam Web API. Sentiment labels were derived from Steam's built in binary recommendation flag. Both transformer models were fine-tuned using identical hyperparameters to ensure a fair comparison. Model performance was evaluated using accuracy, precision, recall, and F1-score, complemented by a qualitative error analysis of cases where the two models produced divergent predictions.

The results show that both transformer models substantially outperformed all baseline models. BERT achieved the highest overall performance with an accuracy of 93.55% and an F1-score of 95.37%, while XLNet achieved 93.47% accuracy and a 95.31% F1-score, with the difference between the two falling below 0.1% across all metrics. Both models reduced total misclassification errors by approximately 40% compared to the best baseline model, Logistic Regression (89.10% accuracy, 92.29% F1-score). The qualitative error analysis revealed that despite near identical aggregate performance, the two models exhibit complementary error patterns:

UNIVERSITAS
GADJAH MADA
Sentiment Analysis of Steam Reviews Using BERT and XLNet
Pasquale Dominic Ligamen Saleppang, Prof. Drs. Edi Winarko, M.Sc.,Ph.D
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

XLNet demonstrates superior handling of adversative conjunctions, long reviews with delayed sentiment resolution, and gaming specific sarcasm, while BERT demonstrates stronger performance on explicit framing, idiomatic expressions, and profanity used in positive contexts. These findings suggest that both architectures are equally suitable for general purpose Steam review sentiment classification, and that model selection should be guided by the specific linguistic phenomena most prevalent in the target application.