**Building of Informatics, Technology and Science (BITS)**
Volume 8, No 1, June 2026 Page: 151−160
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9794

# Sentiment Classification on Indonesian Game Sequels: A Comparative Analysis of SVM and Naive Bayes on Coffee Talk Franchise Reviews

**Nanda Yuris Riziq, Edy Mulyanto***

Faculty of Computer Science, Informatics Engineering Study Program, Universitas Dian Nuswantoro, Semarang, Indonesia
Email: [1]111202416099@mhs.dinus.ac.id, [2,*]edymulyanto@dsn.dinus.ac.id
Correspondence Author Email: edymulyanto@dsn.dinus.ac.id
Submitted: **14/04/2026**; Accepted: **02/06/2026**; Published**: 05/06/2026**

**Abstract**−User reviews on Steam are a critical source of feedback for game developers, yet manual sentiment analysis at scale is impractical. This study aims to compare Support Vector Machine (SVM), Multinomial Naive Bayes (MNB), and Complement Naive Bayes (CNB) for binary sentiment classification and to analyze sequel reception patterns through cross-game evaluation. Reviews were preprocessed with negation-aware stopword removal and WordNet lemmatization, then vectorized with TF-IDF unigram and bigram features. Four scenarios were evaluated: two within-game baselines, a cross-game generalization, and a combined evaluation. Class imbalance was handled at the model level via class weighting for SVM and the CNB variant. Macro-averaged F1-Score was the primary metric. SVM consistently outperformed both Naive Bayes variants, achieving macro-F1 of 0.81 within-game and 0.75 cross-game. MNB collapsed to majority-class prediction across all scenarios; in S2, all three models also failed on the minority class due to the small test partition (n=6). The cross-game result indicates that sentiment patterns transfer reasonably from the original game to its sequel, with the performance drop concentrated in the minority class. These findings offer practical guidance for Indonesian game developers monitoring sequel reception through automated sentiment analysis.

**Keywords:** Sentiment Analysis; Support Vector Machine; Naive Bayes; Steam Reviews; Class Imbalance

## 1. INTRODUCTION

The Indonesian game development industry has undergone substantial growth in recent years and has become an increasingly visible contributor to the country's creative economy. Industry figures place Indonesia's gaming market at approximately US$1.1 billion in 2023, supported by more than 174 million domestic players, and the government has issued Presidential Regulation No. 19 of 2024 specifically to accelerate the development of local game studios [1]. Within this landscape, Coffee Talk, a coffee-shop visual novel developed by Toge Productions, has emerged as one of the most internationally recognized titles produced by an Indonesian studio. The original Coffee Talk, released in 2020, surpassed two hundred thousand global copies sold and demonstrated that domestically produced narrative games could attain commercial success on global digital storefronts [2]. Its 2023 sequel, Coffee Talk Episode 2: Hibiscus & Butterfly, continued the franchise on Steam, and a third installment, Coffee Talk Tokyo, has been announced for 2025 [3]. The franchise therefore offers a rare and well-suited case study for examining how user sentiment evolves between an original game and its sequel within a single Indonesian-developed title.

On digital distribution platforms such as Steam, user reviews function as a critical form of user-generated content because they directly reflect market satisfaction and contain richer feedback than simple star ratings, especially when the reviewer is dissatisfied [4]. The Coffee Talk franchise alone has accumulated thousands of English-language reviews, with the original game collecting 5,247 reviews and the sequel 1,146 reviews in this study's collected dataset. Manually analyzing such a corpus is inefficient, time-consuming, and prone to subjective bias. Sentiment analysis, a sub-field of natural language processing, addresses this problem by automatically classifying the polarity of textual opinions, and has been applied to domains ranging from movie and product reviews to hotel and social-media text [5], [6]. With the release of Coffee Talk Episode 2, a practical question arises for developers and analysts: how can the sentiment of a sequel's audience be measured efficiently and compared to that of the original game in order to assess whether the franchise is maintaining audience satisfaction?

Two challenges complicate the task. First, Steam reviews of highly successful games are heavily skewed toward the positive class. The Coffee Talk dataset exhibits a positive-to-negative ratio greater than 17:1 in the original game and 37:1 in the sequel, which directly mirrors the games' "Overwhelmingly Positive" Steam status. Most existing comparative studies in the Indonesian sentiment-analysis literature instead use artificially balanced datasets, which masks the practical difficulty that real-world skewed data presents to standard classifiers [7]. Second, gaming reviews carry domain-specific linguistic features such as gaming abbreviations (DLC, GOTY, OST), numeric ratings ("10/10"), comparative phrasing typical of sequel reviews ("better than the first," "lost the charm of the original"), and emotionally expressive fan-driven language. These features are known to affect classifier performance and warrant careful preprocessing [6].

Several recent Indonesian studies have compared Support Vector Machine (SVM) and Naive Bayes (NB) for sentiment classification on Tokopedia product reviews [8], Twitter marketplace data [9], hotel reviews [10], Google Play Store reviews [11], and the PUBG Mobile mobile game [12]. The results of these studies are not unanimous. Several works find that SVM consistently outperforms NB on TF-IDF features when the dataset is large or moderately imbalanced [9], [11], [12], [13], [14] while a smaller body of work reports the opposite outcome on certain Indonesian-language Twitter and Tokopedia corpora [8], [15]. International comparative studies on non-English review corpora confirm that model choice and preprocessing significantly affect performance: Michailidis [16] found that

**Building of Informatics, Technology and Science (BITS)**
Volume 8, No 1, June 2026 Page: 151−160
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9794

transformer-based models (BERT, GPT-4) substantially outperformed traditional ML classifiers including SVM and Naive Bayes on Greek e-commerce reviews, reinforcing the motivation for careful algorithm selection in the present study. In the Steam game-review context specifically, prior Indonesian work has applied Naive Bayes to a single game [17], used SVM with lexicon features for opinion-spam detection [18], and applied sentiment analysis with topic modeling to a single Indonesian indie game [19]. Complement Naive Bayes (CNB), introduced specifically to remedy multinomial NB's known bias on skewed data [20], has been shown to outperform standard MNB on imbalanced Indonesian text [21], yet has rarely been compared against SVM on game-review corpora.

Three gaps remain in the existing literature. First, no prior study, to the authors' knowledge, has performed a comparative SVM-versus-Naive-Bayes evaluation on an Indonesian-developed game franchise across multiple installments. Second, most existing comparisons rely on rebalanced datasets that obscure how each classifier behaves on the natural class distribution that developers actually face on Steam. Third, the question of whether a sentiment model trained on an original game generalizes to its sequel, that is, whether sentiment patterns transfer across installments of the same franchise, has not been examined for an Indonesian local title. The present study addresses these gaps by comparing SVM and two Naive Bayes variants on the full natural distribution of Steam reviews from the Coffee Talk franchise, by adopting macro-averaged F1-Score as the primary metric so that the minority class is weighted equally with the majority [7], and by adding an explicit cross-game scenario that trains on the original game and tests on the sequel.

The objective of this research is therefore twofold. The technical objective is to measure and compare the performance of SVM and Naive Bayes algorithms on naturally imbalanced gaming review data using macro-averaged F1-Score, in order to determine which algorithm is more suitable for game review sentiment classification. The practical objective is to analyze user sentiment patterns toward the sequel game compared to its predecessor, through both within-game and cross-game evaluation, and to provide data-driven insight to local game developers regarding product consistency. The contributions of this work are: (1) the first sentiment-analysis study on an Indonesian-developed game sequel, addressing a gap in the local research landscape; (2) a comparative empirical evaluation of SVM, Multinomial Naive Bayes, and Complement Naive Bayes on a naturally imbalanced gaming review dataset, with explicit handling of class imbalance through model-level techniques and macro-F1 reporting; and (3) a cross-game evaluation that quantifies how well a sentiment model trained on an original game transfers to its sequel.

# 2. RESEARCH METHODOLOGY

## 2.1 Research Stages

The research follows a sequential pipeline consisting of seven stages, as shown in Figure 1. The pipeline starts from raw Steam review data and ends with the comparative evaluation of three classification models across four experimental scenarios. Each stage is described in the following subsections.

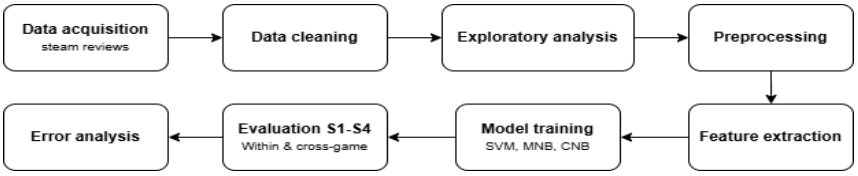



**Figure 1.** Research methodology pipeline.

Stage 1 is data acquisition from steam, and proceeds to Stage 2, quality cleaning, which removes null, duplicate, and near-empty reviews and verifies the language tag. Stage 3 is exploratory data analysis (EDA), which produces descriptive statistics on class distribution, review length, and vocabulary. Stage 4 is text preprocessing, applying domain-term preservation, case folding, negation-aware stopword removal, and WordNet lemmatization. Stage 5 is feature extraction with TF-IDF on unigram and bigram features. Stage 6 defines four experimental scenarios that include both within-game baselines and cross-game generalization. In this stage, the data was also trained on the three classification models on each scenario's training partition. Stage 7 evaluates each trained model with accuracy, per-class precision, recall, F1-Score, and macro-averaged F1-Score, supplemented by qualitative error analysis on misclassified reviews. While Stage 8 is error analysis based on the training and evaluation results.

## 2.2 Dataset and Class Distribution

The dataset consists of English-language Steam reviews of two games developed by the Indonesian studio Toge Productions: Coffee Talk (Steam App ID 914800, released 2020) and Coffee Talk Episode 2: Hibiscus & Butterfly (App ID 1663220, released 2023). Reviews were collected using the publicly available Toge Productions Steam Reviews Tool, with the language filter set to English at scrape time. Each review record includes the review text, a binary Recommended / Not Recommended indicator from the reviewer, total play time, and a timestamp. The Recommended flag was used as a proxy for sentiment label, where Recommended is mapped to the positive class and Not Recommended to the negative class. This proxy is a known approximation [4] but is the de-facto labeling standard for Steam-review sentiment studies.

**Building of Informatics, Technology and Science (BITS)**
Volume 8, No 1, June 2026 Page: 151−160
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9794

Table 1 reports the raw and post-cleaning class distribution. Both games exhibit a heavily skewed distribution toward the positive class, consistent with their "Overwhelmingly Positive" aggregate Steam status. Coffee Talk has a positive-to-negative ratio of approximately 16:1 after cleaning, while Episode 2 has a ratio of approximately 32:1. The combined dataset is therefore representative of real Steam-review distributions for highly rated indie games, rather than the artificially balanced corpora common in prior comparative studies.

**Table 1.** Class Distribution Before and After Data Cleaning

| Game | Raw Total | Cleaned | Positive | Negative |
|---|---|---|---|---|
| Coffee Talk (914800) | 5,247 | 4,444 | 4,186 (94.2%) | 258 (5.8%) |
| Episode 2 (1663220) | 1,146 | 964 | 935 (97.0%) | 29 (3.0%) |
| Combined | 6,393 | 5,408 | 5,121 (94.7%) | 287 (5.3%) |

Data cleaning removed entries with null review text (14 in Coffee Talk and 6 in Episode 2), reviews shorter than five whitespace tokens (769 and 174 respectively), exact duplicate texts (4 and 0), and entries containing no alphabetic characters (16 and 2). A random sample of 200 reviews per game was verified with the langdetect library; the proportion of non-English reviews in each sample remained below 3.5 percent, indicating that Steam's language tag is reliable enough that no further automatic filtering was warranted. Median review length, measured in whitespace-delimited tokens, is 29 tokens for positive reviews and 85 tokens for negative reviews of Coffee Talk, and 27 versus 110 tokens for Episode 2; negative reviews are substantially longer in both games, mirroring the observation in [4] that dissatisfied users typically write more detailed feedback than satisfied users.

### 2.3 Text Preprocessing

The preprocessing pipeline was designed with two principles in mind. First, gaming-domain tokens that carry sentiment information must be preserved rather than discarded. Second, negation must be retained because removing it inverts the sentiment of phrases such as "not good" or "never recommend." The full pipeline executes the following operations in order: (i) domain-term replacement for gaming abbreviations (DLC, GOTY, OST, indie game, visual novel) and numeric rating patterns (10/10, 8/10, 4/5) that are mapped to canonical tokens such as RATING_PERFECT and visual_novel; (ii) lowercasing; (iii) removal of URLs, email addresses, and HTML entities; (iv) removal of non-alphanumeric characters except apostrophes and underscores; (v) tokenization with the NLTK word tokenizer; (vi) part-of-speech tagging followed by WordNet lemmatization with the appropriate POS, which preserves real, interpretable words and avoids the over-stemming common to Porter or Snowball stemmers [6] and (vii) negation-aware stopword removal, in which the standard NLTK English stopword list is reduced by the set {not, no, never, nor, neither, without, n't} so that negation tokens survive into the feature space.

Lemmatization is preferred over stemming on the basis that lemmas remain in dictionary form, which improves both downstream interpretability and bigram quality. For example, the bigram "not recommend" is preserved verbatim by the present pipeline, whereas Porter stemming would reduce surface variants such as "recommended," "recommends," and "recommendation" to a single root form, collapsing semantically distinct forms into one feature. The 'NOT_' negation-tagging heuristic popularized by Pang et al. [5] is not directly applied here; instead, retaining negation tokens combined with bigram features in TF-IDF achieves an equivalent effect, because the bigram "not good" becomes a distinct feature from "good."

### 2.4 Feature Extraction with TF-IDF

Term Frequency-Inverse Document Frequency (TF-IDF) is used to convert preprocessed text into numerical feature vectors. TF-IDF combines the local term frequency in a document with the inverse of the term's frequency across the corpus, so that terms appearing in many documents are downweighted while discriminative terms are emphasized [22], [23]. The weight of term t in document d is computed as in (1):

$$TF\text{-}IDF(t, d) = (1 + \log(tf(t, d))) \times \log(N / df(t)) \quad (1)$$

where tf(t, d) is the raw frequency of term t in document d, N is the total number of documents, and df(t) is the number of documents containing t. The sublinear term-frequency variant 1 + log(tf) is enabled to dampen the effect of very high term frequencies, which is standard practice for review text. Equation (1) presents the textbook un-smoothed formulation. In practice, scikit-learn applies a smoothed variant: 1 is added to both N and df(t) inside the IDF logarithm, and 1 is added to the final IDF value to avoid zero weights for terms appearing in every document [24]. The vectorizer is configured with an n-gram range of (1, 2) so that both unigrams and bigrams are extracted, a minimum document frequency of 2 to ignore terms that appear in only a single review, and a maximum vocabulary size of 5,000 to bound the dimensionality of the feature space. The use of bigrams is supported by prior work showing that multi-word features capture sentiment-bearing constructions such as "highly recommend" and "not worth" that unigrams alone cannot represent [25], [26].

The vectorizer is fitted independently per scenario on the training partition only, and the same fitted vectorizer is then used to transform the test partition. This protocol prevents test-set vocabulary leakage into the training feature space.

**Building of Informatics, Technology and Science (BITS)**
Volume 8, No 1, June 2026 Page: 151−160
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9794

### 2.5 Classification Algorithms

Three classification algorithms are compared. Linear Support Vector Machine (SVM) is implemented with scikit-learn's LinearSVC, which is known to be effective on high-dimensional sparse text features [24]. Only the linear kernel is used because TF-IDF features are already high-dimensional and tend to be linearly separable in practice; non-linear kernels such as RBF rarely improve text-classification performance under these conditions and were therefore deferred to future work in favor of training-time efficiency. The class_weight parameter is set to balanced, meaning that the loss for each class is inversely proportional to its frequency in the training data, which provides an explicit cost-sensitive correction for the heavy class imbalance [7]. The maximum number of iterations is set to 2,000 to ensure convergence on the larger feature matrices.

Multinomial Naive Bayes (MNB) is included as a probabilistic baseline. It models word occurrence counts under a multinomial distribution and is one of the most widely used algorithms for text classification [5], [24]. No class weighting or resampling is applied to MNB; this asymmetry is intentional and lets the experiment characterize the practical behavior of standard MNB on naturally imbalanced data, which is the default configuration adopted in many prior Indonesian comparative studies.

Complement Naive Bayes (CNB) was introduced by Rennie et al. [20] specifically to remedy the bias that affects standard multinomial NB on skewed data. Rather than estimating each class's parameters from the documents that belong to it, CNB estimates them from the complement, that is, all documents that do not belong to it, which corrects the unbalanced training-data effect on parameter estimates. CNB is therefore included as an algorithmic alternative for handling imbalance without resorting to resampling [21].

### 2.6 Experimental Scenarios

Four experimental scenarios are defined to evaluate within-game performance, cross-game generalization, and the combined dataset, as summarized in Table 2. In Scenario S1 the model is trained on 80 percent of Coffee Talk and tested on the remaining 20 percent, providing the within-game baseline for the original game. Scenario S2 mirrors S1 on Episode 2. Scenario S3 is the cross-game scenario: the model is trained on the full Coffee Talk corpus and tested on the full Episode 2 corpus, which directly evaluates whether a sentiment model trained on an original game generalizes to its sequel. Scenario S4 trains and evaluates on the combined dataset using an 80/20 split, providing the standard pooled-evaluation comparison.

**Table 2.** Experimental Scenarios and Their Train/Test Configurations

| ID | Training Data | Testing Data | Purpose |
|---|---|---|---|
| S1 | Coffee Talk (80%) | Coffee Talk (20%) | Within-game baseline (CT1) |
| S2 | Episode 2 (80%) | Episode 2 (20%) | Within-game baseline (CT2) |
| S3 | Coffee Talk (full) | Episode 2 (full) | Cross-game generalization |
| S4 | Combined (80%) | Combined (20%) | Standard pooled evaluation |

All splits use stratified sampling to preserve the natural class distribution in both partitions, with a fixed random seed of 42 to ensure full reproducibility. The post-split class ratios are 16.3:1 and 16.1:1 for the S1 train and test sets, 32.5:1 and 31.2:1 for S2, and 17.8:1 and 18.0:1 for S4. In S3 the train ratio is 16.2:1 and the test ratio is 32.2:1, which reflects the natural difference in class composition between the two games rather than a flaw in stratification.

### 2.7 Evaluation Metrics

Model performance is evaluated using accuracy, per-class precision, recall, and F1-Score, plus macro-averaged and weighted-averaged F1-Score. Macro-averaged F1-Score is adopted as the primary metric because it computes the F1-Score independently for each class and then averages them with equal weight, which avoids the well-documented pitfall in which a model that simply predicts the majority class on heavily skewed data can attain accuracy above 0.94 while completely failing to detect the minority class [7]. Per-class metrics are derived from the confusion matrix following the standard definitions in (2)-(4):

$$Precision = TP / (TP + FP) \quad (2)$$

$$Recall = TP / (TP + FN) \quad (3)$$

$$F1\text{-}Score = 2 \times Precision \times Recall / (Precision + Recall) \quad (4)$$

where TP, FP, and FN denote true positives, false positives, and false negatives respectively for the class under evaluation. The negative class F1-Score is reported separately because, under heavy imbalance, it is the most diagnostic indicator of whether a classifier can detect dissatisfied users, which is precisely the actionable subset for game developers.

### 2.8 Implementation Environment

All experiments are implemented in Python 3.10 using the scikit-learn library [24] for vectorization and modeling, NLTK for tokenization, POS tagging, and WordNet lemmatization, and pandas/NumPy for data manipulation.

**Building of Informatics, Technology and Science (BITS)**
Volume 8, No 1, June 2026 Page: 151−160
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9794

Visualizations are produced with matplotlib and seaborn. The langdetect library is used solely for language verification on samples. All random operations use a fixed seed of 42 to ensure reproducibility. The complete preprocessing and modeling pipeline is documented at the cell level in the accompanying Jupyter notebook.

# 3. RESULT AND DISCUSSION

## 3.1 Overall Comparative Results

Table 3 presents the full per-scenario, per-model evaluation. Each row reports accuracy, precision, recall, and F1-Score for both classes, together with macro-averaged and weighted-averaged F1, and the support count for each class in the test set.

**Table 3.** Full Evaluation Metrics for Each Scenario × Model Combination

| Sc. | Model | Acc. | P-Neg | R-Neg | F1-Neg | F1-Pos | Macro-F1 | Wtd-F1 |
|---|---|---|---|---|---|---|---|---|
| S1 | SVM | 0.963 | 0.732 | 0.577 | 0.645 | 0.980 | 0.813 | 0.961 |
| S1 | MNB | 0.939 | 0.000 | 0.000 | 0.000 | 0.969 | 0.484 | 0.912 |
| S1 | CNB | 0.939 | 0.472 | 0.327 | 0.386 | 0.968 | 0.677 | 0.934 |
| S2 | SVM | 0.964 | 0.000 | 0.000 | 0.000 | 0.982 | 0.491 | 0.951 |
| S2 | MNB | 0.969 | 0.000 | 0.000 | 0.000 | 0.984 | 0.492 | 0.954 |
| S2 | CNB | 0.938 | 0.000 | 0.000 | 0.000 | 0.968 | 0.484 | 0.938 |
| S3 | SVM | 0.973 | 0.560 | 0.483 | 0.519 | 0.986 | 0.752 | 0.972 |
| S3 | MNB | 0.970 | 0.000 | 0.000 | 0.000 | 0.985 | 0.492 | 0.955 |
| S3 | CNB | 0.945 | 0.200 | 0.276 | 0.232 | 0.972 | 0.602 | 0.949 |
| S4 | SVM | 0.960 | 0.625 | 0.614 | 0.620 | 0.979 | 0.799 | 0.960 |
| S4 | MNB | 0.946 | 0.000 | 0.000 | 0.000 | 0.973 | 0.486 | 0.921 |
| S4 | CNB | 0.941 | 0.422 | 0.333 | 0.373 | 0.969 | 0.671 | 0.938 |

**Note:** P-Neg = Precision (Negative class), R-Neg = Recall (Negative class). Bold rows highlight the best macro-F1 model in each scenario where a meaningful comparison is possible (S1, S3, S4).

Figure 2 visualizes the macro-F1 scores across all twelve scenario-and-model combinations. The chart makes the performance gap between SVM and the two Naive Bayes variants immediately apparent, and the dashed line at 0.50 marks the random baseline for reference.

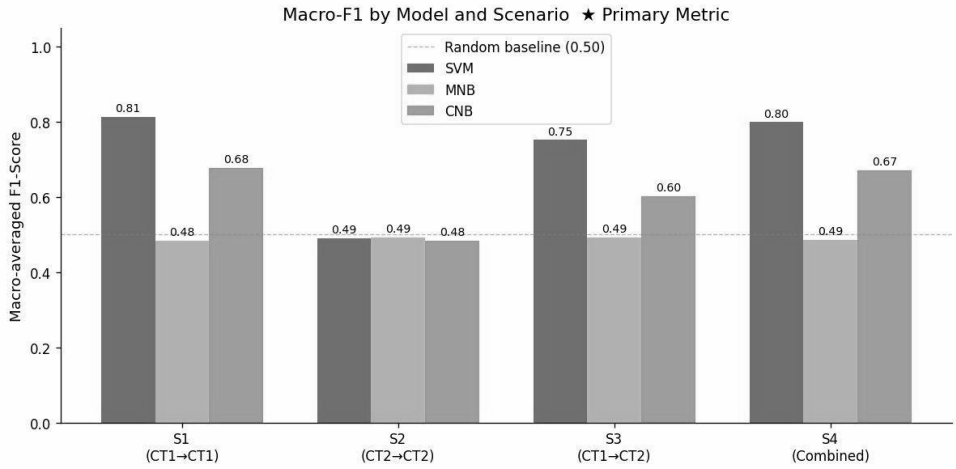



**Figure 2.** Macro-Averaged F1-Score by Model and Scenario. The dashed line indicates the random baseline at 0.50.

Three findings stand out. First, SVM achieves the highest macro-F1 in three of the four scenarios. In S1 (within-game on Coffee Talk) SVM reaches 0.813, in S3 (cross-game) it reaches 0.752, and in S4 (combined) it reaches 0.799. The advantage of SVM over MNB is substantial in every case, ranging from +0.26 in S3 to +0.33 in S1. Second, standard Multinomial Naive Bayes essentially fails to function as a sentiment classifier in this experimental setting: its macro-F1 sits at 0.49 in every scenario, almost exactly the random baseline. Third, Complement Naive Bayes occupies an intermediate position. CNB consistently outperforms MNB by 0.11 to 0.19 macro-F1 across S1, S3, and S4, but it never reaches SVM's level.

## 3.2 Minority-Class Detection

Macro-F1 alone can mask why MNB fails. Figure 3 shows the negative-class F1-Score, which is the most diagnostic measure under heavy imbalance, because it captures the model's ability to identify dissatisfied users.

**Building of Informatics, Technology and Science (BITS)**
Volume 8, No 1, June 2026 Page: 151−160
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9794

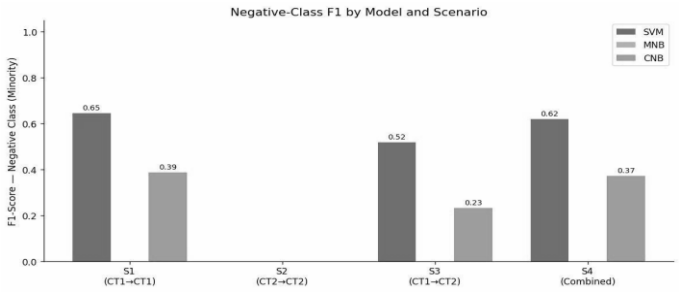
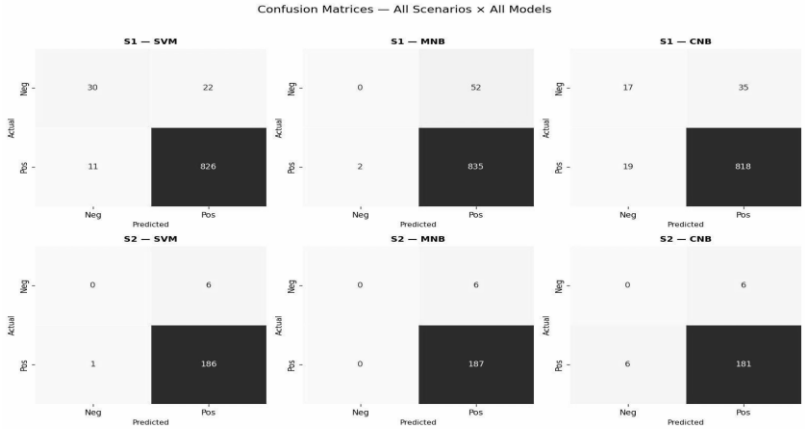


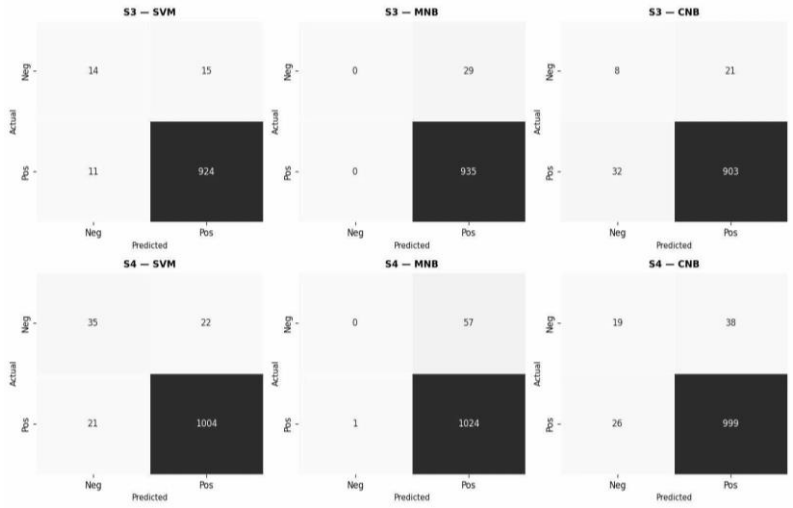

**Figure 3.** Negative-Class F1-Score by Model and Scenario. Bars at zero indicate that the model failed to detect any true negative review.

The negative-class F1 of MNB is 0.00 in every scenario. The confusion matrices in Figure 4 explain the cause: in S1, MNB classifies all 52 true-negative reviews as positive; in S3, it classifies all 29 true-negative reviews as positive; and in S4, it classifies 57 of 57 true-negative reviews as positive. In other words, MNB collapses to majority-class prediction. Its high accuracy (0.94 to 0.97) is therefore a direct artifact of the class imbalance: a constant predictor that always outputs "Positive" would achieve essentially the same accuracy. This finding is fully consistent with the well-known prior bias of Multinomial NB on skewed training data described by Rennie et al. [20] and corroborates the claim that accuracy is misleading on imbalanced sentiment data [7].

SVM with class_weight='balanced' detects the negative class with reasonable but not high quality. Its negative-class F1 reaches 0.645 in S1, 0.519 in S3, and 0.620 in S4. CNB sits between MNB and SVM, with negative-class F1 of 0.386 in S1, 0.232 in S3, and 0.373 in S4. The result indicates that the algorithmic Bayesian correction in CNB is effective, but the explicit cost-sensitive learning in class-weighted SVM is more effective in this setting.




(a)




(b)

**Figure 4.** Confusion Matrices for All Scenarios × Models. Rows are actual labels and columns are predicted labels. The vertical column corresponding to MNB shows that the predicted-Negative cell is empty (or near empty) in every scenario, confirming majority-class collapse.

**Building of Informatics, Technology and Science (BITS)**
Volume 8, No 1, June 2026 Page: 151−160
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9794

### 3.3 Scenario S2: Statistical Reliability and Limitations

Scenario S2 (within-game on Episode 2) deserves a separate note. The 20 percent test partition contains only six negative reviews because Episode 2 has only 29 negative reviews after cleaning, and even with stratified sampling the resulting test minority is too small to support any meaningful comparison. All three models scored 0.00 on negative-class F1 in S2; this is not a model failure per se but a consequence of having only six chances to be right or wrong on the minority class. A single misclassification produces a substantial swing in the metric. For this reason, the comparative claims in this paper are based on Scenarios S1, S3, and S4, while S2 is retained for completeness and reported transparently.

### 3.4 Cross-Game Generalization (S3)

Scenario S3 directly answers the practical question of whether a sentiment model trained on the original Coffee Talk generalizes to its sequel. Table 4 reports the SVM performance delta between S1 (within-game on Coffee Talk) and S3 (Coffee Talk to Episode 2).

**Table 4.** SVM Performance Delta Between Within-Game (S1) and Cross-Game (S3) Scenarios

| Metric | S1 (CT1→CT1) | S3 (CT1→CT2) | Delta |
|---|---|---|---|
| Accuracy | 0.963 | 0.973 | +0.010 |
| Macro-F1 | 0.813 | 0.752 | −0.061 |
| F1 (Negative) | 0.645 | 0.519 | −0.127 |
| F1 (Positive) | 0.980 | 0.986 | +0.006 |
| Precision (Negative) | 0.732 | 0.560 | −0.172 |
| Recall (Negative) | 0.577 | 0.483 | −0.094 |

The transferability of the sentiment model is partial. Macro-F1 drops by 0.061 from S1 to S3, and negative-class F1 drops by 0.127. Positive-class F1 is essentially unchanged, indicating that the vocabulary used by satisfied players is consistent across the two games, while the vocabulary used by dissatisfied players exhibits more drift. Nevertheless, the absolute level of cross-game macro-F1 remains 0.752, which is far above MNB's 0.49 within-game performance and even outperforms CNB's within-game S1 score of 0.677. From a practical standpoint, the result indicates that sentiment patterns transfer reasonably well between an original game and its narrative sequel within the same franchise. Developers can therefore deploy a sentiment classifier trained on a predecessor's reviews to obtain useful early signal on a sequel's reception, with the caveat that minority-class detection should be expected to degrade by approximately 0.10 to 0.15 F1.

### 3.5 SVM Per-Class Behavior

Figure 5 shows SVM's precision, recall, and F1 broken out by class across all four scenarios. The positive class (right panel) is classified almost perfectly in every scenario, with all three metrics at or above 0.97. The negative class (left panel) shows a more variable picture: precision is the strongest metric in S1 at 0.73 but drops to 0.56 in S3 and 0.62 in S4, while recall is uniformly lower than precision. Higher precision than recall on the negative class means that when SVM does predict "negative," it is usually correct, but it tends to leave many true negatives uncaught. The implication is that SVM is conservative on the negative class, that is, it is more likely to miss a negative review than to mislabel a positive review as negative. For developer-facing applications where the cost of missing a complaint is higher than the cost of reviewing a false alarm, the decision threshold can be tuned to trade some precision for additional recall.

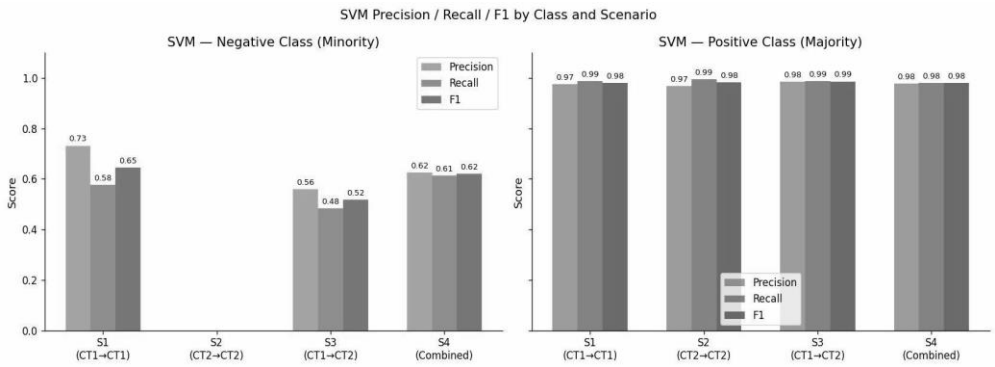



**Figure 5.** SVM Precision, Recall, and F1-Score by Class and Scenario.

### 3.6 Qualitative Error Analysis

To go beyond aggregate metrics, a manual error analysis was performed on the 43 misclassified reviews from S4 SVM, the most representative scenario. Of these, 22 were false negatives (reviews labeled by the user as Not

**Building of Informatics, Technology and Science (BITS)**
Volume 8, No 1, June 2026 Page: 151−160
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9794

Recommended but predicted as Positive by the model) and 21 were false positives (Recommended but predicted as Negative). Each misclassified review was tagged with the dominant linguistic phenomenon that plausibly caused the misclassification. Table 5 reports the resulting category distribution.

**Table 5.** Distribution of Error Categories in S4 SVM (n = 43 Misclassifications)

| Error Category | False Neg. | False Pos. | Total |
|---|---|---|---|
| Mixed sentiment | 12 | 3 | 15 |
| Negation pattern | 0 | 8 | 8 |
| Short / ambiguous | 2 | 3 | 5 |
| Comparison with predecessor | 0 | 1 | 1 |
| Other | 1 | 0 | 1 |

The dominant error pattern is mixed sentiment, which accounts for 15 of 43 misclassifications (35 percent). A representative false negative reads: "I was drawn to the art and overall vibes here and really liked the conversation flows as good as you can make coffee idea. However, the writing is just too off ... the game becomes unenjoyable." The review contains positive descriptors of the art, ambience, and the game's core idea, followed by a substantive negative judgment that ultimately drives the user's Not-Recommended verdict. A bag-of-words model based on TF-IDF cannot weight the rhetorical structure of "draw... like... however... unenjoyable," so it sums positively-charged tokens and predicts Positive. This pattern is well known in the sentiment-analysis literature [5] and has been shown to remain a challenge even for models that incorporate negation scope detection.

The second most frequent category is the negation pattern, accounting for 8 false positives. Several positive reviews contain negation tokens whose scope is local but which the bigram representation reads as global, for example "it's not what I expected, but I love it." Although negation-aware preprocessing preserved the negation tokens, the bigram "not expected" carries enough negative weight in the trained model to flip the prediction. This corroborates findings by HaCohen-Kerner et al. [6] that simplistic preprocessing decisions can have non-trivial effects on sentiment classification accuracy, and aligns with the long-standing observation that negation scope is a primary source of sentiment-classification error. Short or ambiguous reviews account for five errors and are intrinsically hard to classify because they offer little textual evidence. Direct comparison with the predecessor ("not as good as the first one") appears in only one error in this sample, but anecdotally is more common in cross-game scenarios and may partially explain the negative-class drop in S3.

### 3.7 Discussion

Three substantive points emerge from the experimental results. The first concerns the choice of algorithm for naturally imbalanced gaming review data. Class imbalance can also be addressed at the data level through resampling techniques such as SMOTE [27], which generates synthetic minority instances, or through synthetic text generation approaches [28]. this study deliberately addresses imbalance at the model level via class weighting and the Complement formulation [20] in order to preserve the natural Steam-review distribution. The Indonesian comparative literature on SVM versus Naive Bayes is split: studies on PUBG Mobile reviews [12], Indonesian Twitter marketplace data [9], religious application reviews [13], aspect-based gadget reviews [14], Google Play Store reviews [11], and hotel reviews [10] all report SVM advantages, while studies on Tokopedia product reviews [8] and Indonesian Twitter data [15] report the opposite. The present study's finding that SVM with balanced class weights substantially outperforms standard MNB joins the first group, but it adds a critical methodological clarification: the prior MNB-wins results are typically obtained on artificially balanced datasets, and standard MNB without class weighting is fundamentally unsuitable for the natural Steam-review distribution. When CNB, the imbalance-aware NB variant, is included, the gap between Bayesian methods and SVM narrows but does not close. The recommendation for game-review sentiment analysis on Steam is therefore SVM with class weighting as the primary choice, with CNB as a viable alternative when training-time efficiency is critical, and standard MNB only when the training data has been explicitly rebalanced.

The second point concerns the cross-game generalization result. The SVM macro-F1 drop from 0.813 to 0.752 between S1 and S3 indicates a moderate domain shift between the two installments. The drift is concentrated in the negative class while positive-class F1 remains essentially unchanged at 0.98 in both scenarios. The asymmetry suggests that satisfied players evaluate the franchise on a consistent basis, while dissatisfied players are more likely to develop sequel-specific complaints, including direct comparisons with the predecessor that surfaced in the qualitative error analysis. This is consistent with prior empirical observations on Steam that negative reviews carry more detailed and shifting feedback than positive reviews [4].

The third point concerns the practical reliability of the methodology for monitoring sequel reception. Because S3 macro-F1 (0.752) remains substantially above the random baseline and above the within-game performance of MNB on either game, an SVM model trained on a predecessor's reviews can serve as a useful first-line analytical tool when a sequel launches and only a small volume of fresh review data is available. As soon as the sequel's own review corpus reaches a size sufficient for stratified training with adequate minority-class support, a within-game model becomes preferable to the cross-game transfer. The minimum corpus size at which a within-game model surpasses a transferred predecessor model is a relevant question for future work.

**Building of Informatics, Technology and Science (BITS)**
Volume 8, No 1, June 2026 Page: 151−160
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9794

Several limitations bound these conclusions. First, the Recommended / Not Recommended flag is used as a sentiment proxy; some recommended reviews contain substantive criticism and vice versa, as documented in [4] and confirmed by the error analysis. Second, the negative class in Episode 2 (29 reviews) is small in absolute terms, which constrains S2's statistical reliability and contributes to the negative-class F1 drop in S3. Third, the comparison is restricted to two classical algorithms (SVM and Naive Bayes); modern transformer-based models such as BERT have been applied to Steam reviews in recent work [29].

# 4. CONCLUSION

This study compared Support Vector Machine, Multinomial Naive Bayes, and Complement Naive Bayes for binary sentiment classification on 6,393 English-language Steam reviews of the Indonesian-developed Coffee Talk franchise, across four scenarios that included within-game baselines, cross-game generalization from the original to the sequel, and a combined evaluation. The results show that SVM with balanced class weights consistently outperformed both Naive Bayes variants under the natural Steam-review imbalance, achieving macro-averaged F1-Scores of 0.813, 0.752, and 0.799 in scenarios S1, S3, and S4, respectively. Standard Multinomial Naive Bayes collapsed to majority-class prediction in every scenario, confirming that accuracy is misleading on imbalanced sentiment data, while Complement Naive Bayes partially mitigated the imbalance but did not match SVM. The cross-game scenario further demonstrates that sentiment patterns transfer reasonably between an original game and its sequel, with macro-F1 dropping by only 0.061, indicating that local game developers can deploy a predecessor-trained classifier as an early-warning tool for sequel reception. Several limitations bound these findings: the Recommended/Not Recommended flag is used as a sentiment proxy and may not capture nuanced opinions; the negative class in Episode 2 (n=29) is small, which limits Scenario S2's statistical reliability; only two classical algorithms were compared, excluding transformer-based models such as BERT; and findings are based on a single Indonesian franchise, so generalization to other genres or developers requires further validation.

# REFERENCES

[1] ANTARA News, “Indonesia’s game industry levels up,” ANTARA News. Accessed: May 06, 2026. [Online]. Available: https://en.antaranews.com/news/321747/indonesias-game-industry-levels-up
[2] D. W. Nugraha, ““Coffee Talk” dan Asa Industri Gim Lokal,” Kompas.id. Accessed: May 07, 2026. [Online]. Available: https://www.kompas.id/artikel/coffee-talk-dan-asa-industri-gim-lokal
[3] Y. Pratomo and M. F. Nurhappy, “Toge Productions Umumkan Game ‘Coffee Talk Tokyo’, Rilis 2025,” Kompas.com. Accessed: May 07, 2026. [Online]. Available: https://tekno.kompas.com/read/2024/08/28/21000027/toge-productions-umumkan-game-coffee-talk-tokyo-rilis-2025
[4] D. Lin, C.-P. Bezemer, Y. Zou, and A. E. Hassan, “An empirical study of game reviews on the Steam platform,” *Empirical Software Engineering*, vol. 24, no. 1, pp. 170–207, Feb. 2019, doi: 10.1007/s10664-018-9627-4.
[5] B. Pang, L. Lee, and S. Vaithyanathan, “Thumbs up? Sentiment classification using machine learning techniques,” in *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - EMNLP ’02*, Morristown, NJ, USA: Association for Computational Linguistics, 2002, pp. 79–86. doi: 10.3115/1118693.1118704.
[6] Y. HaCohen-Kerner, D. Miller, and Y. Yigal, “The influence of preprocessing on text classification using a bag-of-words representation,” *PLOS ONE*, vol. 15, no. 5, p. e0232525, May 2020, doi: 10.1371/journal.pone.0232525.
[7] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk, and F. Herrera, *Learning from Imbalanced Data Sets*. Cham: Springer International Publishing, 2018. doi: 10.1007/978-3-319-98074-4.
[8] M. Aulia and A. Hermawan, “Analisis Perbandingan Algoritma SVM, Naïve Bayes, dan Perceptron untuk Analisis Sentimen Ulasan Produk Tokopedia,” *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 7, no. 4, p. 1850, Oct. 2023, doi: 10.30865/mib.v7i4.6839.
[9] I. Kurniawan, A. L. Hananto, S. Hilabi, A. Hananto, B. Priyatna, and A. Y. Rahman, “Perbandingan Algoritma Naive Bayes Dan SVM Dalam Sentimen Analisis Marketplace Pada Twitter,” *JATISI (Jurnal Teknik Informatika dan Sistem Informasi)*, vol. 10, no. 1, pp. 731–740, Mar. 2023.
[10] I. Guspian and M. H. Basri, “Comparison of Naive Bayes and SVM Algorithms in Sentiment Analysis for the Optimization of Hotel Operational Services in Central Bangka Regency,” *Jutisi : Jurnal Ilmiah Teknik Informatika dan Sistem Informasi*, vol. 14, no. 2, p. 1315, Aug. 2025, doi: 10.35889/jutisi.v14i2.3107.
[11] L. B. Ilmawan and M. A. Mude, “Perbandingan Metode Klasifikasi Support Vector Machine dan Naïve Bayes untuk Analisis Sentimen pada Ulasan Tekstual di Google Play Store,” *ILKOM Jurnal Ilmiah*, vol. 12, no. 2, pp. 154–161, Aug. 2020, doi: 10.33096/ilkom.v12i2.597.154-161.
[12] P. R. Sari, D. R. Indah, E. Rasywir, M. A. Firdaus, and G. Athalina, “Comparison of Naive Bayes and SVM Algorithms for Sentiment Analysis of PUBG Mobile on Google Play Store,” *SISTEMASI*, vol. 13, no. 6, p. 2767, Nov. 2024, doi: 10.32520/stmsi.v13i6.4814.
[13] Heti Aprilianti, Khothibul Umam, and Maya Rini Handayani, “Comparative Study of SVM, KNN, and Naïve Bayes for Sentiment Analysis of Religious Application Reviews,” *Journal of Applied Informatics and Computing*, vol. 9, no. 3, pp. 920–927, Jun. 2025, doi: 10.30871/jaic.v9i3.9482.
[14] J. W. Iskandar and Y. Nataliani, “Perbandingan Naïve Bayes, SVM, dan k-NN untuk Analisis Sentimen Gadget Berbasis Aspek,” *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, no. 6, pp. 1120–1126, Dec. 2021, doi: 10.29207/resti.v5i6.3588.
[15] M. I. Fikri, T. S. Sabrila, and Y. Azhar, “Perbandingan Metode NaÃ¯ve Bayes dan Support Vector Machine pada Analisis Sentimen Twitter,” *SMATIKA JURNAL*, vol. 10, no. 02, pp. 71–76, Dec. 2020, doi: 10.32664/smatika.v10i02.455.

**Building of Informatics, Technology and Science (BITS)**
Volume 8, No 1, June 2026 Page: 151−160
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9794

[16] P. D. Michailidis, "A Comparative Study of Sentiment Classification Models for Greek Reviews," *BDCC*, vol. 8, no. 9, p. 107, Sep. 2024, doi: 10.3390/bdcc8090107.
[17] A. Pangestu, Y. Tajul Arifin, and R. Ade Safitri, "Analisis Sentimen Review Publik Pengguna Game Online Pada Platform Steam Menggunakan Algoritma Naïve Bayes," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 7, no. 6, pp. 3106–3113, Jan. 2024, doi: 10.36040/jati.v7i6.8829.
[18] R. Taqiuddin, F. A. Bachtiar, and W. Purnomo, "Opinion Spam Classification on Steam Review using Support Vector Machine with Lexicon-Based Features," *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, Nov. 2021, doi: 10.22219/kinetik.v6i4.1323.
[19] M. Y. Febrianta, S. Widiyanesti, and S. R. Ramadhan, "Analisis Ulasan Indie Video Game Lokal pada Steam Menggunakan Analisis Sentimen dan Pemodelan Topik Berbasis Latent Dirichlet Allocation," *J. Animation and Games stud.*, vol. 7, no. 2, pp. 117–144, Oct. 2021, doi: 10.24821/jags.v7i2.5162.
[20] J. Rennie, L. Shih, and J. Teevan, *Tackling the Poor Assumptions of Naive Bayes Text Classifiers*. 2003.
[21] Azwan Triyadi, "Public Sentiment Analysis About Neuralink from Twitter Using Naïve Bayes: Multinomial, Gaussian and Complement," *The Indonesian Journal of Computer Science*, vol. 13, no. 5, Oct. 2024, doi: 10.33022/ijcs.v13i5.4278.
[22] K. Sparck Jones, "A Statistical Interpretation Of Term Specificity And Its Application In Retrieval," *Journal of Documentation*, vol. 28, no. 1, pp. 11–21, Jan. 1972, doi: 10.1108/eb026526.
[23] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing & Management*, vol. 24, no. 5, pp. 513–523, Jan. 1988, doi: 10.1016/0306-4573(88)90021-0.
[24] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 2008. doi: 10.1017/CBO9780511809071.
[25] S. Ranjan and S. Mishra, "Comparative Sentiment Analysis of App Reviews," Jun. 2020.
[26] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*. 2026.
[27] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.
[28] C. Suhaeni and H.-S. Yong, "Enhancing Imbalanced Sentiment Analysis: A GPT-3-Based Sentence-by-Sentence Generation Approach," *Applied Sciences*, vol. 14, no. 2, p. 622, Jan. 2024, doi: 10.3390/app14020622.
[29] J. Al Mursyidy Fadhlurrahman, N. A. Herawati, H. R. Widya Aulya, I. Puspasari, and N. P. Utama, "Sentiment Analysis of Game Reviews on STEAM using BERT, BiLSTM, and CRF," in *2023 International Conference on Electrical Engineering and Informatics (ICEEI)*, Bandung, Indonesia: IEEE, Oct. 2023, pp. 1–6. doi: 10.1109/ICEEI59426.2023.10346219.