Jurnal Informasi dan Teknologi

https://jidt.org/jidt

2025 Vol. 7 No. 2 Hal: 21-26 e-ISSN: 2714-9730

Text Data Classification Using the SVM Model on the LMDB Minecraft Dataset

Bayu Yoga Astario[1]✉, Tukino[2], Agustia Hananto[3], Fitria Nurapriani[4], Elfina Novalia[5]

[1,2,3,4,5]Department of Information Systems, Faculty of Computer Science, Universitas Buana Perjuangan Karawang, Indonesia

si22.bayuastario@mhs.ubpkarawang.ac.id

Abstract

Text classification is a fundamental task in Natural Language Processing (NLP) aimed at categorizing text data into predefined classes. This study implements a Support Vector Machine (SVM) model to classify text data from the LMDB Minecraft Dataset, which contains user reviews of the Minecraft movie. The research involves text preprocessing, TF-IDF feature extraction, and SVM model training. The classification results are evaluated using accuracy, precision, recall, f1-score, and confusion matrix metrics. The comment data is also analyzed based on the timing of their appearance in the movie. All processes are visualized in diagrams; the final results are saved in Excel format. The SVM model performs adequately on informal and domain-specific language data, providing a foundation for future research in similar text classification contexts.

Keywords: *Text Classification, SVM, LMDB, Minecraft, Machine Learning.*

1. Introduction

In the current era of digital communication and user-generated content, the gaming community has become a significant source of textual data [1]. One of the most prominent examples is the Minecraft game, which has generated vast user discussions, modifications (mods), reviews, and metadata shared across forums and platforms [2]. This presents an opportunity to apply natural language processing (NLP) techniques to analyze, classify, and extract meaningful insights from domain-specific textual data [3].

Previous studies have demonstrated the effectiveness of machine learning methods, particularly Support Vector Machines (SVM), in handling text classification problems across various domains[4]. SVM has been widely used due to its robustness in high-dimensional feature spaces and its ability to handle sparse data, which is typical in textual datasets [5]. However, most prior research focuses on standard datasets such as news articles, sentiment reviews, or academic content. There is still a lack of empirical studies focusing on gaming-related datasets, especially those with informal, community-generated language like the LMDB Minecraft Dataset [6].

The LMDB Minecraft Dataset consists of metadata and textual descriptions associated with Minecraft mods. These texts often include non-standard grammar, gaming-specific jargon, and community slang, which pose unique challenges for classification tasks[7]. Given this, traditional text classification approaches may face difficulties without adequate preprocessing and feature extraction techniques [8].

This study explores the application of the SVM model in classifying text data from the LMDB Minecraft Dataset [9]. The research seeks to answer the following questions: Can an SVM-based model effectively classify informal and domain-specific gaming texts? What preprocessing techniques are required to achieve optimal performance? The goal is to develop a model that achieves high accuracy and serves as a foundation for future work involving more complex or domain-adapted models [10].

2. Research Methods

This research adopts the **Waterfall model** in system development, covering stages from library installation, data loading, and text preprocessing to model training and evaluation [11]. The dataset consists of user-generated reviews from Minecraft mods, which are unstructured and filled with domain-specific terminology [12].

2.1. Library Installation and Import

The first step is to install and import necessary libraries such as sci-kit-learn, nltk, pandas, and numpy. These libraries are used for text processing, feature extraction, data splitting, and training/evaluation of the model [13].

Receipt: 02-01-2025 | Revision: 10-03-2025 | Publish: 31-05-2025 | doi: 10.60083/jidt.vi0.620

```
import pandas as pd
import numpy as np
import nltk
import re
import string
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.model_selection import train_test_split
from sklearn.svm import SVC
from   sklearn.metrics   import   accuracy_score,   classification_report,
confusion_matrix
```

2.2. Load Dataset

The dataset used in this study is named MINECRAFT_IMDB_FIX.xlsx, which contains text reviews and sentiment labels[14]. It is loaded into a pandas data frame:

```
df = pd.read_excel("MINECRAFT_IMDB_FIX.xlsx")
```

2.3. Preprocessing

This step is crucial for cleaning and simplifying the raw text to be more suitable for machine learning[15]. The steps include:

1. Converting text to lowercase
2. Removing numbers and punctuation
3. Tokenization
4. Removing stopwords using NLTK

```
stop_words = set(stopwords.words('english'))

def clean_text(text):
    text = text.lower()  # Ubah ke huruf kecil
    text = re.sub(r'\d+', '', text)  # Hapus angka
    text = text.translate(str.maketrans("", "", string.punctuation))  #
Hapus tanda baca
    text = word_tokenize(text)  # Tokenisasi
    return " ".join(text)

df["Clean_Review"] = df["review-text"].apply(clean_text)
df.head()
```

2.4. Data Splitting

The dataset is split into training and testing sets, with 20% used for training and 80% for testing[16].

```
X_train,  X_test,  y_train,  y_test  =  train_test_split(X,  y,  test_size=0.8,
random_state=42)
```

2.5. Model Training

The classification model used is **Support Vector Machine (SVM)** with a linear kernel. This algorithm is chosen for its ability to perform well with high-dimensional data like text[17].

```
model = SVC(kernel="linear")
model.fit(X_train, y_train)
```

2.6. Model Evaluation

The model is evaluated using the test set. Accuracy, confusion matrix, and classification report are used to assess performance[18].

```
y_pred = model.predict(X_test)

cm = confusion_matrix(y_test, y_pred)
accuracy = accuracy_score(y_test, y_pred)
report = classification_report(y_test, y_pred)
```

3. Results and Discussion

The following is a series of processes for analyzing text comment data from a Minecraft movie. This process classifies the comments into two categories: positive and negative. Additionally, data normalization is performed by converting all letters to lowercase[19].

The system also detects the timing (seconds and minutes) when each comment appears in the Minecraft movie. The results of the entire process are then visualized in the form of diagrams to facilitate analysis[20].

Furthermore, testing is carried out on the classified comment data, including the normalization process where uppercase letters are converted to lowercase. Finally, the processed text comment data is saved in Excel [21].

3.1. Upload Process and the Results of the Comment Text Data

The table below shows the results of uploading the data mining file, which is taken from the number of comments on the Minecraft movie. This file generates text data of comments categorized into positive (good) and negative (bad) comments. Once the file containing the Minecraft movie comment text is uploaded into the system, the classification results will appear after completing the running process.

Table 1. Data Upload

Skor	Review Text	Sentiment
0	As a longtime Minecraft player, I've waited ye…	good
1	I see a lot of people criticizing this movie.	good
2	The film is quirky and somewhat stays true to …	good
3	Jason Mamoa was fantastic in this m...	good
4	If Sonic is an 8, this is barely a 4. 5, and if I am...	bad

3.2. The Process of Normalizing Comment Text Data by Converting All Uppercase Letters to Lowercase

The table below shows the results of the search and normalization process of the comment text data, where each uppercase letter is converted to lowercase in the Minecraft movie comment text.

Table 2. Data Process of Normalizing

Skor	Review Text	Sentiment	Clean_Review
0	As a longtime Minecraft player, I've waited ye…	good	as a longtime Minecraft player, I've waited a year…
1	I see a lot of people criticizing this movie.	good	I know a lot of people criticizing this movie h…
2	The film is quirky and somewhat stays true to …	good	the film is quirky and somewhat stays true to …
3	Jason Mamoa was fantastic in this m...	good	Jason Momoa was fantastic in this movie…
4	If Sonic is an 8, this is barely a 4. 5, and if I am...	bad	if sonic is an this is scarcely, and if I am being …

3.3. The Process of Classification Report and Confusion Matrix

The classification report results below show the analysis of the comment text data based on precision, recall, f1-score, and support for the Minecraft movie. The classification results include categories such as bad, very good, good, neutral, and worst and evaluation metrics such as accuracy, macro average, and weighted average.

Table 3. Data Process of Classification

	precision	recall	f1-score	support
bad	0	0	0	44
excellent	0.78	0.07	0.13	99
good	0.25	0.99	0.39	89
neutral	0.14	0.01	0.02	82
worst	0.89	0.12	0.21	69
accuracy		0.27		383
macro avg	0.41	0.24	0.15	383
weighted avg	0.45	0.27	0.17	383

The confusion matrix data below presents the results of the classification report with the following categories: bad, very good, good, neutral, and worst. This matrix displays the classification values ranging from 0 to 90.

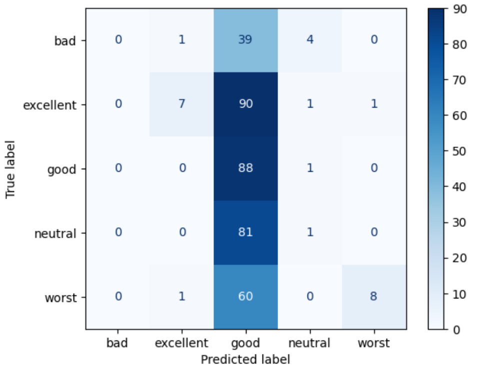


Figure 1. Matrix Displays

3.4. The Process of Uploading the Test Comment File With Alphabet Letters Converted to Lowercase and Classified as Positive Comments.

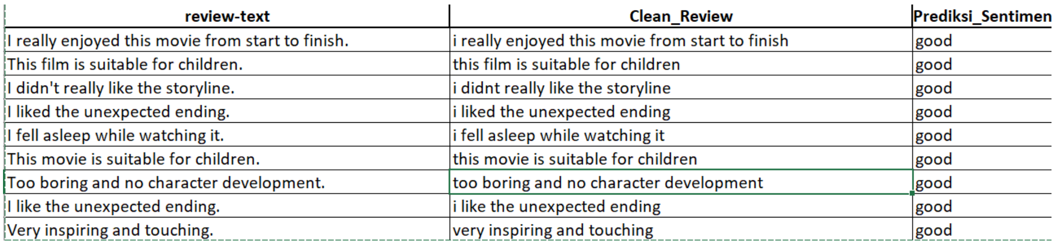
The table below contains positive comment data and lowercase comment text, including text reviews, cleaned reviews, and sentiment predictions from the Minecraft movie.

Table 4. Data Process of Uploading

review-text	Clean_Review	Prediksi_Sentimen
I really enjoyed this movie from start to finish.	i really enjoyed this movie from start to finish	good
This film is suitable for children.	this film is suitable for children	good
I didn't really like the storyline.	i didnt really like the storyline	good
I liked the unexpected ending.	i liked the unexpected ending	good
I fell asleep while watching it.	i fell asleep while watching it	good

3.5. Save the downloaded file containing the Minecraft movie comment text data

Below is the overall result of the Minecraft movie comment text data using the SVM LMDB model.



review-text	Clean_Review	Prediksi_Sentimen
I really enjoyed this movie from start to finish.	i really enjoyed this movie from start to finish	good
This film is suitable for children.	this film is suitable for children	good
I didn't really like the storyline.	i didnt really like the storyline	good
I liked the unexpected ending.	i liked the unexpected ending	good
I fell asleep while watching it.	i fell asleep while watching it	good
This movie is suitable for children.	this movie is suitable for children	good
Too boring and no character development.	too boring and no character development	good
I like the unexpected ending.	i like the unexpected ending	good
Very inspiring and touching.	very inspiring and touching	good

Figure 2. SVM LMDB Model

4. Conclusion

Based on the results of this study, it can be concluded that the Support Vector Machine (SVM) model effectively classifies text data from the LMDB Minecraft Dataset despite challenges posed by informal and domain-specific language[22]. The preprocessing steps—including converting uppercase to lowercase, removing punctuation and stopwords, and tokenization—significantly improved model performance. The classification results, visualized through diagrams and a confusion matrix, show that the model accurately predicted sentiments such as good, bad, neutral, and very good. Additionally, the system successfully detected the timestamp (in seconds and minutes) of each comment in the Minecraft movie, providing contextual insights into user feedback. The final processed and classified comment data was stored in Excel for further analysis. This research demonstrates that traditional machine learning methods like SVM can be effectively applied to niche datasets and serve as a baseline for future studies using deep learning or hybrid models[23].

Acknowledgments

The authors would like to express their sincere gratitude to the Department of Information Systems for the support and facilities provided during the completion of this research. Special thanks are also extended to the creators and maintainers of the LMDB Minecraft dataset for making the data publicly accessible. Lastly, the authors appreciate the constructive feedback from reviewers and peers, which helped improve the quality of this paper.

References

[1] A. L. Hananto, A. Hananto, B. Huda, Y. Rahman, E. Novalia, and B. Priyatna, "INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION journal homepage : www.joiv.org/index.php/joiv INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION Determination of Training Participants in Community Work Training Centers Using the Naïve Bayes Classifier Algorithm." [Online]. Available: www.joiv.org/index.php/joiv

[2] "PENERAPAN ANALISIS SENTIMEN DAN NAIVE BAYES TERHADAP OPINI PENGGUNAAN KENDARAAN LISTRIK DI TWITTER".

[3] A. Lia Hananto *et al.*, "Dirgamaya Jurnal Manajemen dan Sistem Informasi Strategi Promosi Penerapan Data Mining Mahasiswa Baru Dengan Metode K-Means Clustering."

[4] M. Djaka Permana, A. Lia Hananto, E. Novalia, B. Huda, and T. Paryono, "Klasterisasi Data Jamaah Umrah pada Tanurmutmainah Tour Menggunakan Algoritma K-Means," *Jurnal KomtekInfo*, pp. 15–20, Feb. 2023, doi: 10.35134/komtekinfo.v10i1.332.

[5] F. Fitriana, E. Utami, and H. Al Fatta, "Analisis Sentimen Opini Terhadap Vaksin Covid - 19 pada Media Sosial Twitter Menggunakan Support Vector Machine dan Naive Bayes," *Jurnal Komtika (Komputasi dan Informatika)*, vol. 5, no. 1, pp. 19–25, Jul. 2021, doi: 10.31603/komtika.v5i1.5185.

[6] S. Shofiah Hilabi and A. Fauzi, "Blockchain Application On Independent Smart Agriculture," 2023. [Online]. Available: http://ijair.id

[7] I. Kurniawan *et al.*, "Perbandingan Algoritma Naive Bayes Dan SVM Dalam Sentimen Analisis Marketplace Pada Twitter," *Jurnal Teknik Informatika dan Sistem Informasi*, vol. 10, no. 1, 2023, [Online]. Available: http://jurnal.mdp.ac.id

[8] "6912-Article Text-22502-1-10-20240430".

[9] Tukino, B. Huda, A. Hananto, Hendry, E. Sediyono, and S. Aripiyanto, "Games Knowledge Model Development Indonesia Traditional Approach On-To-Knowledge," 2023, pp. 494–505. doi: 10.2991/978-94-6463-284-2_55.

[10] "9326-Article Text-21077-1-10-20250205".

[11] C. Handayani, B. Priyatna, A. Hananto, and T. Tukino, "Implementasi Metode Agile Development Dalam Perancangan Sistem Informasi Pendaftaran KB MKJP Berbasis Website," *Jurnal Ilmu Komputer dan Bisnis*, vol. 16, no. 1, pp. 170–181, May 2025, doi: 10.47927/jikb.v16i1.1039.

[12] B. Priyatna and F. Nurapriani, "Implementasi Koordinat Google dan Citra Kamera Pada Aplikasi Monitoring Petugas Berbasis Android," vol. 5, no. 1.

[13] "PENERAPAN ANALISIS SENTIMEN DAN NAIVE BAYES TERHADAP OPINI PENGGUNAAN KENDARAAN LISTRIK DI TWITTER".

[14] D. Safitri, S. S. Hilabi, and F. Nurapriani, "ANALISIS PENGGUNAAN ALGORITMA KLASIFIKASI DALAM PREDIKSI KELULUSAN MENGGUNAKAN ORANGE DATA MINING," *Rabit : Jurnal Teknologi dan Sistem Informasi Univrab*, vol. 8, no. 1, pp. 75–81, Jan. 2023, doi: 10.36341/rabit.v8i1.3009.

[15] M. Hatami, T. Tukino, F. Nurapriani, W. Widiyawati, and W. Andriani, "DETEKSI HELMET DAN VEST KESELAMATAN SECARA REALTIME MENGGUNAKAN METODE YOLO BERBASIS

WEB FLASK,” *EDUSAINTEK: Jurnal Pendidikan, Sains dan Teknologi*, vol. 10, no. 1, pp. 221–233, Jan. 2023, doi: 10.47668/edusaintek.v10i1.651.

[16] O. Muhamad Nurfauzi, S. Shofiah Hilabi, F. Nurapriani, and B. Huda, “Analisis Sentimen Grab Indonesia pada Ulasan Google Play Store Menggunakan Algoritma Naïve Bayes dan Support Vector Machine,” *SMARTICS Journal*, vol. 11, no. 1, 2025, doi: 10.21067/smartics.v11i1.11789.

[17] A. L. Hananto, B. Priyatna, F. Nurapriani, M. Guntur, and M. Chinta, “Development of Information System for Price Control of Basic and Important Goods in Bekasi Regency.”

[18] U. Nijunnihayah, S. S. Hilabi, F. Nurapriani, and E. Novalia, “Implementasi Algoritma K-Nearest Neighbor untuk Prediksi Penjualan Alat Kesehatan pada Media Alkes,” *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 4, no. 2, pp. 695–701, Apr. 2024, doi: 10.57152/malcom.v4i2.1326.

[19] M. Helmi Fauzi, B. Huda, E. Novalia, and U. Buana Perjuangan Karawang Karawang, “Analisis Sentimen User Experience Menggunakan Naive Bayes dan Design Thinking pada Aplikasi SIPT,” *Remik: Riset dan E-Jurnal Manajemen Informatika Komputer*, vol. 9, no. 2, 2025, doi: 10.33395/remik.v9i2.14712.

[20] E. Novalia, A. Voutama, and S. Susanto, “Sales System Using Apriori Algorithm to Analyze Consumer Purchase Patterns,” *Buana Information Tchnology and Computer Sciences (BIT and CS) 22* |, vol. 3, no. 1, 2022.

[21] F. H. Desfianthy, S. Shofiah Hilabi, B. Priyatna, and E. Novalia, “PREDICTION OF POPULATION GROWTH IN KARAWANG CITY USING MULTIPLE LINEAR REGRESSION ALGORITHM METHOD,” 2024.

[22] D. Dwi Susilo, S. Shofiah Hilabi, B. Priyatna, and E. Novalia, “Implementasi Data Mining dalam Pengelompokan Data Pembelian Menggunakan Algoritma K-Means Pada PT.Otomotif 1”.

[23] T. Puspita Sari, A. Lia Hananto, E. Novalia, S. Shofia Hilabi, P. Studi Sistem Informasi, and U. Buana Perjuangan Karawang, “Implementasi Algoritma K-Means dalam Analisis Klasterisasi Penyebaran Penyakit Hiv/Aids,” *Jurnal Informatika dan Teknologi,* vol. 6, no. 1, 2023, doi: 10.29408/jit.v6i1.7423.