MIFORTEKH (Jurnal Manajemen Informatika & Teknologi)
p-ISSN : 2808-7550 (print) e-ISSN : 2798-0235 (online)
Vol. 6, No. 2, Oktober 2026
https://journal.stiestekom.ac.id/index.php/mifortekh

Analisis Sentimen dan Identifikasi Topik Kritik Pembaruan *Game Umamusume* Menggunakan SVM Dan LDA

Muhammad Rafi Mauludin[1], Novita Anggraini[2], Harry Pribadi Fitrian[3]
[1,3]Universitas Teknologi Digital, Bandung, Indonesia
[2]Universitas Teknologi Digital Komputer, Bandung, Indonesia
Email author: muhammad20122021@digitechuniversity.ac.id [1], novitaanggraini@digitechuniversity.ac.id [2], harrypribadi@digitechuniversity.ac.id [3]

Article Info

Article history:

Received July 19, 2026
Revised August 26, 2026
Accepted October 22, 2026

Keywords:

Sentiment Analysis,
LDA,
Topic Modeling,
SVM,
Umamusume.

ABSTRACT

Whenever a new update or patch is released for a gacha game, the player community typically responds in various ways. Consequently, this study focuses on sentiment analysis and the identification of key topics on the social media platform X regarding updates to the game *Umamusume: Pretty Derby*. Data was collected via scraping, resulting in 139 cleaned data points after a preprocessing phase. Following TF-IDF weighting, the data was classified using a Support Vector Machine (SVM) model, which achieved an accuracy of 64.29%. Evaluation results indicate that the model performs reasonably well in identifying positive and neutral sentiments; however, it struggles to detect negative sentiments, partly due to an imbalanced dataset and the frequent use of sarcasm. Meanwhile, the Latent Dirichlet Allocation (LDA) method successfully identified three primary topics: reactions to patch notes, and discussions concerning characters and gameplay. It is hoped that these findings will provide valuable input for developers to evaluate future updates. Although further improvements are needed, this level of accuracy demonstrates the potential of the SVM approach for text classification within the gacha game domain.

Corresponding Author:

Muhammad Rafi Mauludin,
Universitas Teknologi Digital
Jl. Cibogo indah III – Bodogol Rt.08/05 Kel. Mekarjaya Kec. Rancasari Kota Bandung – Jawa Barat 40292
Email: muhammad20122021@digitechuniversity.ac.id

1. INTRODUCTION

Industri *game* seluler berbasis model *freemium* atau *gacha* berkembang sangat pesat dalam beberapa tahun terakhir. Kelangsungan hidup dan keberlanjutan pendapatan dari *game* jenis ini sangat tergantung pada tingkat retensi pemain (*player retention*), yaitu seberapa lama pemain bertahan dan aktif bermain dalam jangka panjang (Hamari, J., Hanner, N., & Koivisto, J.., 2020). Salah satu strategi utama yang diterapkan oleh pihak pengembang (*developer*) untuk menjaga antusiasme komunitas adalah dengan merilis pembaruan konten (*patch update*) secara rutin. Pembaruan ini dapat berupa

penambahan karakter baru, penyesuaian keseimbangan *gameplay* (*buff/nerf*), perbaikan *bug*, hingga peluncuran acara (*event*) berkala.

Umamusume: Pretty Derby, sebuah *game* taktik dan simulasi pacuan kuda fiksi yang dikembangkan oleh Cygames, merupakan salah satu judul populer yang secara konsisten melakukan pembaruan konten. Setiap kali *patch* baru dirilis, komunitas pemain akan memberikan beragam respons, kritik, dan masukan di berbagai platform media sosial, khususnya X (dahulu Twitter). Masukan dari komunitas ini menyimpan potensi informasi yang sangat berharga bagi pengembang untuk mengevaluasi kualitas pembaruan yang telah dirilis (Liu, B, 2012).

Namun, kendala utama terletak pada sifat dari data ulasan tersebut. Opini yang diutarakan oleh netizen umumnya berbentuk teks tidak terstruktur (*unstructured text*), menggunakan bahasa informal, serta berjumlah sangat masif. Menganalisis ribuan cuitan secara manual tentu tidak efisien dan rentan terhadap bias subjektif. Oleh karena itu, pendekatan berbasis ilmu komputasi dan *Machine Learning* menjadi solusi yang krusial. Beberapa penelitian terdahulu telah membuktikan keandalan *Machine Learning* dalam melakukan analisis sentimen pada ulasan *game*, seperti penelitian ulasan Roblox (Dewi, N. K. F. P., Sudipa, I. G. I., Sunarya, I. W., Dewi, N. W. J. K., & Kusuma, A. S., 2025), Grand Theft Auto V (Saputra, A. K., Handayani, M. R., Wibowo, N. C. H., & Umam, K, 2025), maupun Genshin Impact (Sulistyo, D. A., & Fahrudillah, M. F, 2025). Analisis sentimen juga terbukti efektif saat diterapkan pada evaluasi layanan finansial digital dan produk *e-commerce* (Anggraini, N., & Suroyo, H, 2019), (Huynh, T, 2025), (Reka, K., Raja, G., & Malarchelvi Pd, S. K, 2025).

Meskipun demikian, sebagian besar penelitian analisis sentimen terdahulu hanya berfokus pada pengelompokan polaritas kalimat (positif, netral, negatif) tanpa menggali lebih dalam mengenai akar masalah atau topik spesifik apa yang memicu timbulnya sentimen tersebut. Padahal, pemodelan topik telah sukses diterapkan untuk membedah isu-isu publik di media berita (Rakhmawati, N. A., Cisatra, A., Ansori, D. D. M., Akmal, D. N. F. A., & Ramadhani, S, 2024) maupun persepsi terhadap teknologi baru (Ridat, A. F. I., & Isnain, A. R, 2025).

Untuk mengisi celah penelitian (*research gap*) tersebut, penelitian ini mengusulkan pendekatan hibrida (*hybrid approach*) dengan mengombinasikan algoritma *Support Vector Machine* (SVM) untuk klasifikasi sentimen dan *Latent Dirichlet Allocation* (LDA) untuk pemodelan topik. Pendekatan gabungan SVM-LDA ini sebelumnya pernah terbukti efektif pada ulasan aplikasi komunitas *gamer* (Pratama, M. M. A., Kurniawan, D., Rifai, A., & Tania, K. D, 2026). Dengan mengintegrasikan kedua algoritma ini, penelitian diharapkan tidak hanya dapat memetakan proporsi sentimen pemain terhadap pembaruan *game Umamusume:Pretty Derby*, tetapi juga mampu mengidentifikasi Klaster topik kritik utama yang melatarbelakangi sentimen tersebut secara lebih komprehensif. SVM dipilih karena efektivitasnya dalam menangani data teks berdimensi tinggi dengan margin pemisah yang jelas, sementara LDA dipilih karena kemampuannya mengidentifikasi topik tersembunyi tanpa data label topik.

2. METHOD

2.1 Kerangka Kerja Penelitian

Penelitian ini dijalankan dengan mengadopsi kerangka kerja standar *Cross-Industry Standard Process for Data Mining* (CRISP-DM) yang meliputi enam tahapan utama: pemahaman bisnis (*business understanding*), pemahaman data (*data understanding*), persiapan data (*data preparation*), pemodelan (*modeling*), evaluasi (*evaluation*), dan penyebaran (*deployment*) (Wirth, R., & Hipp, J, 2020). Seluruh pemrosesan komputasi dilakukan menggunakan bahasa pemrosesan Python (Sri, M, 2025) dengan memanfaatkan pustaka utama seperti pandas, scikit-learn, nltk, dan gensim (Gupta, P., & Sehgal, N. K., 2021).

2.2 Pengumpulan dan Preprocessing Data

Data ulasan dikumpulkan dari platform media sosial X melalui teknik *web scraping* menggunakan kata kunci (*keywords*) yang berkaitan dengan pembaruan *game Umamusume: Pretty Derby*. Dari proses ekstraksi awal, diperoleh sebanyak 474 cuitan mentah. Setelah dilakukan penyaringan manual, didapatkan dataset bersih sebanyak 139 cuitan yang siap diolah. Data yang

dibuang terdiri dari 150 cuitan data duplikat, 100 cuitan yang tidak relevan dengan pembaruan game, dan 85 cuitan yang hanya berisi *link* atau *retweet.*

Untuk meningkatkan kualitas data sebelum masuk ke tahap pemodelan, dilakukan lima tahapan *preprocessing* teks (Rahmi Hidayati, H. P. F., Ikhwan Ruslianto, 2026):

1. **Cleaning:** Menghapus elemen non-teks seperti URL, hastag, angka, tanda baca (*punctuation*), emoji, dan *mention* (@username).
2. **Case Folding:** Mengubah seluruh karakter huruf dalam teks menjadi huruf kecil (*lowercase*).
3. **Tokenization:** Memecah deretan kalimat menjadi potongan kata individual (*tokens*).
4. **Stopword Removal:** Mengeliminasi kata-kata umum yang tidak memiliki bobot informasi signifikan (seperti "yang", "di", "ke", "dan").
5. **Stemming:** Mengubah kata berimbuhan menjadi bentuk kata dasarnya menggunakan pustaka Bahasa Indonesia.

2.3 Ekstraksi Fitur (TF-IDF)

Teks hasil *preprocessing* kemudian diubah menjadi representasi matriks angka menggunakan metode *Term Frequency-Inverse Document Frequency* (TF-IDF). Metode ini menghitung seberapa penting suatu kata t dalam sebuah dokumen d relatif terhadap keseluruhan korpus dokumen D. Nilai pembobotan TF-IDF dihitung menggunakan persamaan matematis berikut:

$$\text{TF-IDF}(t, d, D) = \text{TF}(t, d) \times \log\left(\frac{|D|}{|\{d \in D: t \in d\}|}\right)$$

Matriks TF-IDF yang dihasilkan berukuran 139 x N, di mana N adalah jumlah kata unik setelah proses preprocessing.

2.4 Pemodelan SVM dan LDA

Pada tahap klasifikasi sentimen, digunakan algoritma *Support Vector Machine* (SVM) berbasis pembelajaran statistik (Vapnik, V, 1995). SVM bekerja dengan cara mencari *hyperplane* optimal yang memisahkan data antarkelas sentimen (positif, netral, negatif) dengan margin maksimal dalam ruang dimensi tinggi (Joachims, T, 1998).
Persamaan keputusan dasar linier pada SVM dinyatakan sebagai berikut:

$$f(x) = \text{sign}(w \cdot x + b)$$

Di mana w merupakan vektor bobot (*weight*), x adalah vektor fitur masukan (hasil TF-IDF), dan b adalah nilai bias.

Sementara itu, untuk pemodelan topik, digunakan algoritma berbasis distribusi probabilistik *Latent Dirichlet Allocation* (LDA) (Blei, D. M., Ng, A. Y., & Jordan, M. I., 2023). Parameter LDA yang digunakan: alpha = 'auto', beta = 0.01, jumlah iterasi = 300, dan jumlah topik diuji pada rentang 2-5 untuk menentukan coherence score tertinggi. LDA mengasumsikan bahwa setiap dokumen merupakan campuran dari beberapa topik tersembunyi (*latent topics*), dan setiap topik terdiri dari distribusi kata-kata tertentu (Hidayat, F. M., Cahyadi, Sanjaya, H., Purnomo, D., & Wiranto, H, 2026). Metode ini digunakan untuk mengelompokkan kata-kata yang sering muncul bersamaan dalam dokumen ke dalam klaster isu yang bermakna.

3. RESULT DAN ANALISIS

Bagian ini menjabarkan hasil dari pengujian model, baik untuk sentimen, pemetaan topiknya, serta diskusi secara terintegrasi.

3.1. Hasil Klasifikasi Sentimen dengan SVM

Berdasarkan proses pelabelan manual terhadap 139 data cuitan bersih, diperoleh distribusi kelas sentimen yang tidak seimbang (*imbalanced data*). Komposisi dataset terdiri dari sentimen Positif sebanyak 62 data (44,6%), sentimen Netral sebanyak 53 data (38,1%), dan sentimen Negatif sebanyak 24 data (17,3%).

Tabel 1. Distribusi Dataset Sentimen Ulasan Umamusume

Kelas Sentimen	**Jumlah Data (Tweet)**	**Persentase (%)**
Positif	62	44,6%
Netral	53	38,1%
Negatif	24	17,3%
Total	**139**	**100,0%**

Setelah dilakukan pengujian klasifikasi menggunakan algoritma SVM dengan skema pembagian data *training* dan *testing* sebesar 80:20, model berhasil memperoleh tingkat akurasi sebesar **64,29%**. Evaluasi kinerja model SVM secara rinci disajikan pada Tabel 2.

Tabel 2. Matriks Evaluasi Kinerja Model SVM per Kelas Sentimen

Kelas Sentimen	**Precision**	**Recall**	**F1-Score**
Positif	0.68	0.73	0.70
Netral	0.62	0.67	0.64
Negatif	0.50	0.33	0.40
Rata-rata (Macro)	**0.60**	**0.58**	**0.58**

Berdasarkan Tabel 2, model SVM mencatatkan kinerja terkuat pada klasifikasi sentimen positif (F1-score 0,70) dan netral (F1-score 0,64). Namun, kinerjanya menurun signifikan pada sentimen negatif dengan *Recall* sebesar 0,33 dan *F1-score* 0,40. Penurunan kinerja pada kelas negatif dipicu oleh dua faktor utama:

1. **Ketidakseimbangan Data (*Data Imbalance*):** Proporsi sampel kelas negatif yang hanya sebesar 17,3% membuat model SVM kurang mendapatkan contoh pola kalimat yang cukup selama proses pelatihan (*training phase*).
2. **Kompleksitas Bahasa dan Sarkasme:** Karakteristik cuitan di platform X sangat didominasi oleh gaya bahasa informal, penyingkatan kata, serta penggunaan tuturan ironi atau sarkasme. Sebagai contoh, kalimat *"Wah hebat banget update kali ini bikin game-nya makin sering crash"*

secara harafiah mengandung kata bermakna positif ("hebat"), namun secara kontekstual bermakna kritik negatif. Model SVM berbasis linier memiliki keterbatasan dalam membedakan konteks kalimat sindiran tersebut.

3.2. Hasil Identifikasi Topik dengan LDA

Pengujian algoritma LDA berhasil memetakan kata-kata utama dari korpus teks menjadi 3 klaster topik yang dominan. Penentuan jumlah K=3 topik ini menghasilkan nilai *coherence score* yang paling optimal. Nilai coherence score untuk K=3 adalah 0.52, yang merupakan nilai tertinggi dibandingkan K=2 (0.41) dan K=4 (0.45). Ringkasan kata kunci dan representasi isu dari tiap topik disajikan pada Tabel 3.

Tabel. 3 Hasil Pemodelan Topik Menggunakan LDA

Topik	Kata Kunci Utama (Top Keywords)	Interpretasi Isu / Topik Obrolan
Topik 1	*update, patch, rilis, bagus, mantap, event, gacha*	**Reaksi Terhadap Patch Notes:** Ekspresi antusiasme dan respon awal pemain terhadap fitur/event baru yang dirilis.
Topik 2	*global, server, inggris, kapan, tunggu, beda, jadwal*	**Ekspektasi Versi Global:** Keinginan komunitas agar game meluncurkan server global berbahasa Inggris dan menyamakan jadwal *patch*.
Topik 3	*karakter, Tanino Gimlet, skill, nerf, buff, strategi, lari*	**Karakter dan Gameplay:** Diskusi teknis terkait performa karakter tertentu dan penyesuaian mekanik *gameplay*.

4. DISCUSSION

Penggabungan metode SVM dan LDA pada penelitian ini menunjukkan potensi untuk memberikan wawasan yang lebih komprehensif bagi pengembang game dibandingkan pendekatan sentimen konvensional. Jika ditelusuri lebih lanjut melalui analisis silang, cuitan negatif mayoritas berasal dari Topik 2 (kekecewaan versi global) dan Topik 3 (kritik terhadap nerf/buff karakter), yang menunjukkan bahwa isu-isu ini menjadi pemicu utama sentimen negatif.

Capaian akurasi SVM sebesar 64,29% menegaskan bahwa pengelompokan teks pada media sosial berdimensi tinggi memiliki tantangan tersendiri, terutama terkait keterbatasan volume dataset dan ketidakseimbangan kelas. Dari sudut pandang praktis, hasil ekstraksi LDA memberikan tiga implikasi manajerial yang krusial bagi pengembang Cygames:

1. **Evaluasi Fitur Patch (Topik 1):** Topik 1 mendominasi sebanyak 40% dari total cuitan, menunjukkan bahwa konten event baru mendapat perhatian dan respons positif yang signifikan dari komunitas. Hal ini menunjukkan bahwa strategi *gameplay loop* harian yang ditawarkan patch saat ini sudah sesuai dengan ekspektasi mayoritas pemain.
2. **Kebutuhan Lokalisasi (Topik 2):** Munculnya isu server global sebagai salah satu topik utama mengindikasikan adanya potensi pasar internasional yang sangat besar yang belum tergarap optimal. Kesenjangan waktu rilis versi Jepang dan rencana rilis global menjadi pemicu frustrasi utama bagi pemain non-Jepang.
3. **Keseimbangan Mekanik Game (Topik 3):** Diskusi intensif mengenai karakter spesifik (seperti Tanino Gimlet) menunjukkan bahwa komunitas sangat peka terhadap perubahan

statistik atau skill karakter. Pengembang perlu lebih berhati-hati dalam melakukan buff/nerf agar tidak merusak meta permainan.

Sebagai langkah perbaikan untuk penelitian mendatang, keterbatasan akurasi akibat *data imbalance* dapat diatasi dengan menerapkan teknik manipulasi sintesis data seperti *Synthetic Minority Over-sampling Technique* (SMOTE). Selain itu, penggunaan arsitektur berbasis *Deep Learning* atau *Transformer* (seperti BERT/IndoBERT) sangat disarankan untuk meningkatkan pemahaman kontekstual terhadap gaya bahasa sarkastik.

5. CONCLUSION

Penelitian ini berhasil mengimplementasikan pendekatan hibrida SVM-LDA untuk analisis sentimen dan topik pada ulasan game Umamusume: Pretty Derby. Model SVM mencapai akurasi 64,29%, dengan kelemahan utama pada deteksi sentimen negatif akibat ketidakseimbangan data dan kompleksitas bahasa sarkastik. LDA memetakan tiga isu utama: reaksi patch notes (Topik 1), ekspektasi versi global (Topik 2), dan diskusi gameplay (Topik 3). Pendekatan ini memberikan wawasan operasional bagi pengembang, namun perlu ditingkatkan dengan penyeimbangan kelas, penambahan data, dan eksplorasi arsitektur Deep Learning untuk penelitian selanjutnya.

ACKNOWLEDGEMENTS

Terima kasih kepada Program Studi Informatika, Universitas Teknologi Digital atas dukungan fasilitas yang diberikan selama penelitian ini berlangsung. Penulis juga mengucapkan terima kasih kepada komunitas pemain Umamusume: Pretty Derby yang secara tidak langsung telah menyediakan kumpulan data berharga melalui diskusi publik di media sosial.

REFERENCES

Anggraini, N., & Suroyo, H. (2019). Comparison of sentiment analysis against digital payment "t-cash and go-pay" in social media using orange data mining. *Journal of Information Systems and Informatics*, *1*(2), 152–163. https://doi.org/10.33557/journalisi.v1i2.21

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Dewi, N. K. F. P., Sudipa, I. G. I., Sunarya, I. W., Dewi, N. W. J. K., & Kusuma, A. S. (2025). Sentiment analysis of roblox game reviews using support vector machine method. *Sinkron : Jurnal Dan Penelitian Teknik Informatika*, *9*(4), 1863–1876. https://doi.org/10.33395/sinkron.v9i4.15272

Gupta, P., & Sehgal, N. K. (2021). *Introduction to machine learning in the cloud with python: Concepts and practices*. Springer International Publishing. https://doi.org/10.1007/978-3-030-71270-9

Hamari, J., Hanner, N., & Koivisto, J. (2020). "Why pay premium in freemium services?" A study on purchasing behavior in virtual worlds. *International Journal of Information Management*, *51*, 102032.

Hidayat, F. M., Cahyadi, Sanjaya, H., Purnomo, D., & Wiranto, H. (2026). Pemodelan topik berita nasional indonesia menggunakan latent dirichlet allocation. *INFOTECH Journal*, *12*(1), 33–39. https://doi.org/10.31949/infotech.v12i1.16926

Huggett, K. N., Quesnelle, K. M., & Jeffries, W. B. (Eds.). (2022). *An introduction to medical teaching: The foundations of curriculum design, delivery, and assessment* (Vol. 20). Springer International Publishing. https://doi.org/10.1007/978-3-030-85524-6

Huynh, T. (2025). *Sentiment analysis of Amazon's product reviews*. https://aaltodoc.aalto.fi/handle/123456789/140072

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. *European Conference on Machine Learning*, 137–142.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, *5*(1), 1–167.

Pratama, M. M. A., Kurniawan, D., Rifai, A., & Tania, K. D. (2026). Sentiment and topic analysis of digital

community application gamer reviews using svm-lda and crisp-dm. *Sistemasi: Jurnal Sistem Informasi*, *15*(1), 181–193. https://doi.org/10.32520/stmsi.v15i1.5746

Rahmi Hidayati, H. P. F., Ikhwan Ruslianto. (2018). Implementasi metode naive bayes classifier untuk aplikasi filtering email spam dengan lemmatization berbasis web. *Coding Jurnal Komputer Dan Aplikasi*, *6*(2). https://doi.org/10.26418/coding.v6i2.25487

Rakhmawati, N. A., Cisatra, A., Ansori, D. D. M., Akmal, D. N. F. A., & Ramadhani, S. (2024). Identifikasi topik hangat di media berita menggunakan latent dirichlet allocation. *JIEET (Journal of Information Engineering and Educational Technology)*, *8*(1), 14–17. https://doi.org/10.26740/jieet.v8n1.p14-17

Reka, K., Raja, G., & Malarchelvi Pd, S. K. (2025). Ai and deep reinforcement learning for sentiment analysis of customer reviews. *Web Intelligence*, *23*(2), 213–221. https://doi.org/10.1177/24056456241304010

Ridat, A. F. I., & Isnain, A. R. (2025). Pemodelan topik dengan lda untuk memahami persepsi publik terhadap apple vision pro melalui analisis berita twitter. *JIPI (Jurnal Ilmiah Penelitian Dan Pembelajaran Informatika)*, *10*(4), 3233–3246. https://doi.org/10.29100/jipi.v10i4.6500

Saputra, A. K., Handayani, M. R., Wibowo, N. C. H., & Umam, K. (2025). Sentiment analysis of user reviews on the game gta v using support vector machine. *Jurnal Sisfokom (Sistem Informasi Dan Komputer)*, *14*(3), 284–290. https://doi.org/10.32736/sisfokom.v14i3.2368

Sri, M. (2021). *Practical natural language processing with python: With case studies from industries using text data at scale*. Apress. https://doi.org/10.1007/978-1-4842-6246-7

Sulistyo, D. A., & Fahrudillah, M. F. (2025). Analisis sentimen ulasan pemain genshin impact menggunakan kombinasi tf-idf, lexicon, dan support vector machine. *JURNAL FASILKOM*, *15*(3), 580–593. https://doi.org/10.37859/jf.v15i3.10553

Vapnik, V. (1995). *The nature of statistical learning theory*. Springer-Verlag.

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*, 29–39.