PENGEMBANGAN SISTEM PELABELAN OTOMATIS UNTUK *ASPECT-BASED SENTIMENT ANALYSIS* PADA ULASAN *VIDEO GAME* MENGGUNAKAN *LARGE LANGUAGE MODELS*

1st Raditya Naufal Wicaksono
S1 Sistem Informasi
Telkom University
Bandung, Indonesia
raditya.nw@gmail.com

2nd Nur Ichsan Utama, S.T., M.T, Ph.D.
S1 Sistem Informasi
Telkom University
Bandung, Indonesia

3rd Nia Ambarsari, S.Si., M.T.
S1 Sistem Informasi
Telkom University
Bandung, Indonesia

***Abstrak*—**Industri video game berkembang pesat dengan jutaan ulasan pengguna yang dihasilkan setiap tahun, namun volume dan kompleksitas opini multi-aspek dalam ulasan tersebut menyulitkan pengembang untuk memperoleh ringkasan yang terstruktur. Penelitian ini bertujuan untuk merancang dan mengembangkan sebuah sistem prototipe fungsional berupa *dashboard* untuk mengotomatiskan proses *Aspect-Based Sentiment Analysis* (ABSA) pada ulasan video game. Metodologi penelitian yang digunakan adalah pendekatan hibrida yang mengintegrasikan CRISP-DM untuk fase riset data dan Model Waterfall untuk fase pengembangan sistem. Sebuah dataset *ground truth* berkualitas tinggi yang terdiri dari 896 ulasan Steam berhasil dibangun melalui proses anotasi manual yang ketat dan divalidasi dengan reliabilitas antar-anotator yang sangat tinggi (nilai Fleiss' Kappa rata-rata > 0.85). Eksperimen evaluasi komprehensif dilakukan untuk menemukan kombinasi strategi *prompt engineering* dan *Large Language Model* (LLM) yang paling optimal. Hasil penelitian menunjukkan bahwa strategi prompt *4-Shot + Guideline* merupakan yang paling efektif. Pada evaluasi perbandingan model, Llama 4 Maverick menunjukkan kinerja tertinggi untuk tugas *Aspect Category Detection* (ACD) dengan Macro F1-Score 0.9055 dan tugas *Aspect Category Sentiment Classification* (ACSC) dengan Macro F1-Score 0.8373. Temuan ini diimplementasikan dalam sebuah prototipe *dashboard* fungsional, yang menunjukkan kelayakan kombinasi model LLM *state-of-the-art* dengan strategi prompt yang tepat sebagai pendekatan yang menjanjikan untuk menghasilkan analisis opini pemain yang terstruktur secara otomatis.

***Kata kunci*—** analisis sentimen berbasis aspek*, large language model, prompt engineering*, ulasan video game, pengembangan sistem.

I. PENDAHULUAN

Industri video game terus menunjukkan pertumbuhan pesat, dengan nilai pasar global mencapai $196 miliar pada tahun 2023 dan diproyeksikan tumbuh 6% per tahun hingga 2028 [1]. Seiring pertumbuhan ini, ulasan dari pemain di platform seperti Steam menjadi komponen krusial yang tidak hanya memengaruhi keputusan pembelian [2], tetapi juga keterlibatan jangka panjang pemain [3]. Namun, volume ulasan yang masif, seperti pada Cyberpunk 2077 yang mencapai lebih dari 700.000 ulasan, menyulitkan pengembang dan pemain untuk memproses dan memahami opini yang terkandung secara manual [4].

Analisis sentimen konvensional yang hanya menilai sentimen secara keseluruhan dinilai kurang efektif untuk ulasan video game yang multi-dimensi [5]. Sebagai solusi, Aspect-Based Sentiment Analysis (ABSA) dikembangkan untuk mengidentifikasi sentimen pada aspek spesifik seperti gameplay atau grafis [6]. Kemunculan Large Language Models (LLM) seperti GPT-4 membuka pendekatan baru dalam ABSA melalui prompt engineering, yang memungkinkan analisis zero-shot dan few-shot tanpa pelatihan ulang [7]. Berbagai penelitian menunjukkan bahwa efektivitas LLM dalam tugas ABSA sangat bergantung pada kualitas prompt dan pemilihan model [8].

Oleh karena itu, penelitian ini berfokus pada perancangan dan pengembangan sebuah sistem *dashboard* fungsional untuk otomatisasi ABSA pada ulasan video game. Untuk mencapai tujuan tersebut, penelitian ini melalui tahap evaluasi sistematis untuk menentukan kombinasi model LLM dan teknik *prompt engineering* yang paling andal, yang hasilnya kemudian diimplementasikan sebagai inti dari sistem yang dibangun.

II. KAJIAN TEORI

Menyajikan dan menjelaskan teori-teori yang berkaitan dengan variabel-variabel penelitian. Poin subjudul ditulis dalam abjad.

A. Analisis Sentimen Berbasis Aspek (ABSA)

Analisis sentimen adalah teknik untuk menganalisis dan mengkategorikan ekspresi sentimen dalam sebuah teks [9]. ABSA merupakan tingkatan analisis yang lebih mendetail, yang bertujuan memahami hubungan antara aspek, opini, dan polaritas sentimen dalam sebuah kalimat. Penelitian ini berfokus pada dua subtugas utama ABSA:

1. *Aspect Category Detection* (ACD): Proses mengidentifikasi kategori aspek yang telah ditentukan sebelumnya dari dalam sebuah kalimat. Kategori ini biasanya bersifat spesifik domain [10].
2. *Aspect Category Sentiment Classification* (ACSC): Proses mengklasifikasikan polaritas sentimen (positif, negatif, netral) terhadap kategori aspek yang telah diidentifikasi [11].

Pendekatan ini dipilih karena paling sesuai untuk domain ulasan *video game* yang memiliki aspek-aspek terstandarisasi dan cocok untuk skenario zero/few-shot menggunakan LLM.

B. Large Language Models dan Prompt Engineering

Natural Language Processing (NLP) adalah cabang kecerdasan buatan yang dirancang untuk membuat komputer memahami bahasa manusia [12]. Perkembangan terkini dalam NLP adalah Large Language Models (LLM), yaitu model yang dilatih pada data teks dalam jumlah masif menggunakan arsitektur transformer untuk memahami dan menghasilkan bahasa manusia [13]. Karena telah dilatih pada korpus data yang sangat besar, LLM dapat diarahkan untuk menyelesaikan berbagai tugas, termasuk ABSA, hanya dengan instruksi berbasis teks (prompt) tanpa perlu pelatihan ulang. Teknik merancang dan mengoptimalkan prompt ini dikenal sebagai prompt engineering, yang memainkan peran

krusial dalam memandu LLM untuk mengenali aspek dan sentimen secara akurat [7].

C. Analisis Ulasan Video Game

Ulasan pemain pada platform digital seperti Steam telah menjadi sumber data yang sangat berharga. Ulasan tekstual terbukti memiliki pengaruh paling signifikan terhadap penjualan [14]. Namun, ulasan video game memiliki kompleksitas tinggi; sering kali mengandung opini positif dan negatif dalam satu kalimat atau perbandingan dengan game lain, yang menyebabkan analisis sentimen konvensional menjadi kurang efektif [15]. Beberapa penelitian sebelumnya telah mencoba mengatasi tantangan ini menggunakan berbagai metode seperti kombinasi BiLSTM dan CRF [16], pendekatan unsupervised dengan SBERT dan clustering [17], hingga Random Forest [18]. Namun, penelitian-penelitian ini masih berfokus pada model konvensional dan belum mengevaluasi potensi penuh dari LLM state-of-the-art dengan prompt engineering canggih.

D. Metodologi Riset dan Pengembangan Sistem

Untuk menjembatani riset data dan pengembangan sistem, penelitian ini menggunakan pendekatan hibrida.

1. CRISP-DM: *Cross-Industry Standard Process for Data Mining* adalah metodologi standar untuk proyek data science. Kerangka kerja ini membagi siklus hidup proyek menjadi enam fase utama: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, dan Deployment. Penekanannya pada pemahaman tujuan bisnis dan proses yang iteratif membuatnya sangat cocok untuk tahap riset dan analisis data dalam penelitian ini [19].
2. Model Waterfall: Dikenal juga sebagai Linear Sequential Life Cycle Model, metodologi ini bersifat sekuensial di mana setiap fase pengembangan perangkat lunak (seperti Requirement Analysis, System Design, Implementation, Testing) harus selesai sebelum fase berikutnya dimulai. Model ini sangat cocok untuk proyek dengan ruang lingkup yang sudah terdefinisi dengan jelas dan tuntutan dokumentasi yang tinggi, sesuai dengan karakteristik proyek Tugas Akhir ini[20].

Kombinasi kedua metodologi ini memastikan bahwa proses riset data dijalankan secara sistematis dan proses rekayasa perangkat lunak untuk membangun prototipe dilakukan secara terstruktur.

III. METODE

Penelitian ini mengadopsi metodologi hibrida yang mengintegrasikan CRISP-DM untuk fase riset data dan Model *Waterfall* untuk pengembangan sistem.

A. Tahap Riset dan Analisis Data (CRISP-DM)

Fase ini mencakup seluruh kegiatan untuk menemukan kombinasi *prompt* dan model LLM yang paling optimal.

1. *Business Understanding*: Analisis kebutuhan di industri *video game* dilakukan melalui studi literatur untuk mengidentifikasi aspek-aspek kunci yang memengaruhi persepsi pemain. Enam aspek utama— *Gameplay, Graphics, Story, Technical, World,* dan *Multiplayer*— diidentifikasi dan didefinisikan secara operasional [21].

TABEL 1
(Definisi Setiap Aspek Berdasarkan Literatur)

Aspek	Definisi Ringkas	Sumber
Gameplay	Tindakan dan interaksi pemain dengan tantangan dalam game.	[22]
Graphics Fidelity	Kualitas elemen visual, estetika, dan tampilan grafis.	[23]
Story	Alur cerita, narasi, karakter, dan pengembangan plot.	[24]
Technical	Kinerja, *bug*, *crash*, dan isu kompatibilitas perangkat keras.	[25]
World	Desain lingkungan, struktur dunia, dan atmosfer dalam *game*.	[24]
Multiplayer	Interaksi antar pemain, fitur multiplemain	[26]

2. *Data Understanding*: Data ulasan mentah dikumpulkan dari Steam API, kemudian dibersihkan untuk menghapus ulasan spam atau non-informatif. Pra-pelabelan dengan model DeepSeek V3 dilakukan untuk memahami karakteristik awal data.
3. *Data Preparation*: Tahap ini berfokus pada pembangunan *dataset ground truth* berkualitas tinggi. Prosesnya meliputi perancangan pedoman anotasi, anotasi manual oleh tiga anotator, dan uji reliabilitas menggunakan Fleiss' Kappa (κ). Hasil uji menunjukkan tingkat kesepakatan "Hampir Sempurna" (rata-rata $\kappa > 0.85$). *Dataset* kemudian disempurnakan lebih lanjut melalui validasi dengan pendekatan "LLM-as-a-Judge" untuk meninjau ambiguitas.

TABEL 2
(Hasil Uji Reliabilitas Antar-Annotator Menggunakan Fleiss' Kappa)

Aspek Sentimen	Nilai Fleiss' Kappa (κ)
Gameplay	0.955
Graphics Fidelity	0.881
Story	0.923
Technical	0.96
World	0.877
Multiplayer	0.896

4. Pada tahap *Modelling*, dua eksperimen utama dirancang. Pertama, evaluasi pengaruh *prompt engineering* dengan menguji variasi prompt (*Minimal, Guideline, 2/4/6-shot + Guideline*) pada model *baseline* GPT-4.1. Kedua, evaluasi komparatif beberapa model LLM (GPT-4.1, Llama 4 Maverick, DeepSeek R1, dll.) menggunakan *prompt* terbaik dari tahap pertama. Kinerja model diukur menggunakan metrik *Macro F1-Score* untuk tugas ACD dan ACSC, yang dipilih karena kemampuannya menangani data tidak seimbang secara adil[27].

$$\text{Macro F1} = \frac{1}{k}\sum_{c=1}^{k} F1_c$$

B. Tahap Pengembangan Sistem (*Waterfall*)

Fase ini berfokus pada rekayasa perangkat lunak untuk membangun prototipe *dashboard* berdasarkan hasil dari tahap riset.

1. *Requirement Analysis*: Kebutuhan fungsional (misalnya, menambah *game*, memulai anotasi, visualisasi *insight*) dan non-fungsional (misalnya, pemrosesan asinkron, keamanan kunci API) didokumentasikan.
2. *System Design*: Arsitektur sistem dirancang, dan artefak desain seperti diagram UML (Use Case, Activity, Sequence, Class) dan ERD dibuat untuk memvisualisasikan struktur dan alur kerja sistem.
3. Proses pengkodean frontend dan backend dilakukan. Model dan prompt LLM terbaik diintegrasikan ke dalam logika backend.
4. Testing: Pengujian fungsional (*Black-Box*) dan struktural (*White-Box*) dilakukan untuk memverifikasi setiap fitur dan memastikan sistem berjalan sesuai harapan.

IV. HASIL DAN PEMBAHASAN

A. Kinerja Model dan Validasi Prompt

Tahap pertama mengevaluasi bagaimana struktur *prompt* memengaruhi kinerja GPT-4.1. Hasil untuk subtugas *Aspect Category Detection* (ACD) menunjukkan bahwa kelengkapan *prompt* secara konsisten meningkatkan performa. Kinerja puncak dicapai oleh *4-Shot + Guideline* dengan rata-rata *Macro F1-Score* 0.8856, menunjukkan bahwa 4 contoh adalah titik optimal untuk tugas deteksi aspek.

TABEL 3
(Perbandingan Performa Prompt untuk Subtugas ACD pada GPT-4.1 (Macro F1))

Variasi *Prompt*		*MP*	*GP*	*2G*	*4G*	*6G*
Aspek	*Gameplay*	0.5004	0.5865	0.7071	**0.7582**	0.7505
	Graphics Fidelity	0.9091	0.899	0.909	**0.9202**	0.9134
	Story	0.9315	0.9232	0.923	**0.9358**	0.9313
	Technical	0.8239	0.857	0.8374	0.849	**0.8632**
	World	0.8489	0.8641	0.8831	**0.8977**	0.8909
	Multiplayer	**0.9579**	0.9543	0.9566	0.9525	0.9543
	Macro-F1 Average	0.8286	0.8474	0.8694	**0.8856**	0.8839

Keterangan: MP: *Minimal Prompt*; GP: *Guideline Prompt*; 2G/4G/6G: *Prompt Few-shot* (2/4/6 contoh) + *Guideline.*

Pada subtugas *Aspect Category Sentiment Classification* (ACSC) yang lebih kompleks, pentingnya konteks dalam *prompt* menjadi lebih jelas. Performa terbaik dicapai oleh varian *few-shot*, dengan *6-Shot + Guideline* (0.8276) dan 4-Shot + Guideline (0.8272) menunjukkan hasil yang hampir identik. Mengingat efisiensi, 4-Shot + Guideline dipilih sebagai *prompt* terbaik untuk tahap selanjutnya.

TABEL 4
(Perbandingan Performa Prompt untuk Subtugas ACSC pada GPT-4.1 (Macro F1))

Variasi *Prompt*		*MP*	*GP*	*2G*	*4G*	*6G*
Aspek	*Gameplay*	0.7705	0.7786	0.7854	0.7911	**0.8211**
	Graphics Fidelity	0.8146	0.8383	0.8906	0.8841	**0.9044**
	Story	**0.8699**	0.8687	0.8517	0.8664	0.8461
	Technical	0.7568	0.799	0.8328	**0.8412**	0.8379
	World	0.7084	0.7688	0.7736	0.7873	**0.7879**
	Multiplayer	0.8016	0.8312	0.7982	0.7934	0.7685
	Macro-F1 Average	0.7869	0.8141	0.8221	0.8272	**0.8276**

Menggunakan *prompt 4-Shot + Guideline*, beberapa model LLM dievaluasi. Pada tugas ACD, DeepSeek R1 menunjukkan kinerja tertinggi (Macro F1 0.9033), mengindikasikan bahwa sebagian besar LLM modern andal dalam mendeteksi topik dengan *prompt* yang baik.

TABEL 5
(Perbandingan Kinerja Model LLM untuk Tugas ACD)

Model		*GPT-4.1*	*Llama 4 Maverick*	*DeepSeek R1 0528*	*Gemini 2.5 Flash (Thinking)*	*Qwen 3 235B A22B*
Aspek	*Gameplay*	0.8061	**0.8317**	0.8241	0.7582	0.8087
	Graphics Fidelity	**0.9238**	0.9177	0.9124	0.9202	0.9019
	Story	**0.9445**	0.9401	0.9326	0.9358	0.9258
	Technical	**0.9178**	0.8946	0.8545	0.849	0.8777
	World	0.8711	0.8643	0.8845	**0.8977**	0.8569
	Multiplayer	**0.9564**	0.9315	0.9357	0.9525	0.9319
	Macro-F1 Average	**0.9033**	0.8967	0.8906	0.8856	0.8838

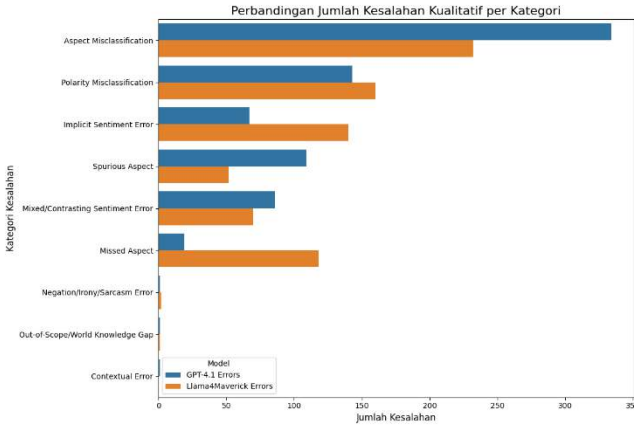
Namun, pada tugas ACSC yang lebih menuntut pemahaman nuansa, terjadi pergeseran peringkat yang signifikan. GPT-4.1 muncul sebagai model terunggul (Macro F1 0.8272), diikuti oleh Llama4Maverick (0.8163). DeepSeek R1, yang sebelumnya unggul, mengalami penurunan performa. Ini menegaskan bahwa kemampuan deteksi aspek tidak selalu berkorelasi dengan kemampuan klasifikasi sentimen, di mana model seperti GPT-4.1 dan Llama4Maverick menunjukkan keunggulan dalam penalaran konteks yang lebih dalam.

TABEL 6
(Perbandingan Kinerja Model LLM untuk Tugas ACSC)

Model	*GPT-4.1*	*Llama 4 Maverick*	*DeepSeek R1 0528*	*Gemini 2.5 Flash*	*Qwen 3 235B A22B*

					(Thinking)	
Aspek	*Gameplay*	0.7911	**0.7986**	0.7222	0.722	0.7266
	Graphics Fidelity	**0.8841**	0.851	0.7375	0.7793	0.7768
	Story	0.8664	**0.8901**	0.8627	0.831	0.8309
	Technical	**0.8412**	0.8199	0.7693	0.7801	0.7425
	World	0.7873	0.787	**0.7992**	0.7512	0.79
	Multiplayer	**0.7934**	0.7513	0.7539	0.7365	0.7046
	Macro-F1 Average	**0.8272**	0.8163	0.7741	0.7667	0.7619

Untuk memahami "perilaku" model, dilakukan analisis kesalahan kualitatif pada dua model teratas, GPT-4.1 dan Llama4Maverick.



GAMBAR 1
(Perbandingan Jumlah Kesalahan Kualitatif per Kategori)

Gambar di atas menunjukkan bahwa kedua model memiliki profil kelemahan yang berbeda. GPT-4.1 cenderung lebih "agresif" atau "kreatif", yang menyebabkannya lebih sering melakukan kesalahan klasifikasi aspek (Aspect Misclassification) dan kesalahan polaritas (Polarity Misclassification). Sebaliknya, Llama4Maverick lebih "konservatif", yang membuatnya lebih sering melewatkan aspek yang sebenarnya ada (Missed Aspect) dan gagal menangkap sentimen tersirat (Implicit Sentiment Error). Studi kasus spesifik mengonfirmasi pola ini; misalnya, GPT-4.1 salah mengklasifikasikan sentimen campuran sebagai negatif murni, sementara Llama4Maverick gagal mendeteksi kritik implisit mengenai bug teknis. Menariknya, kedua model menunjukkan performa sangat baik pada kasus sulit seperti sarkasme dan ironi, dengan jumlah kesalahan yang mendekati nol.

B. Pengujian dan Hasil Keluaran Sistem Prototipe

Prototipe *dashboard* yang dikembangkan telah melalui serangkaian pengujian fungsional (*Black-Box*) dan keamanan (*White-Box*). Hasil pengujian menunjukkan bahwa semua skenario, mulai dari penambahan game hingga pengunduhan dataset, berjalan sesuai dengan hasil yang diharapkan dan dinyatakan *PASSED*. Sebagai studi kasus, sistem dijalankan untuk menganalisis 100 ulasan Cyberpunk 2077. Protitipe berhasil menghasilkan halaman *insight* yang menyajikan ringkasan statistik, visualisasi distribusi aspek dan sentimen, serta tabel data mentah yang interaktif. Output ini mendemonstrasikan kemampuan sistem untuk mengubah data ulasan yang tidak terstruktur menjadi wawasan yang dapat ditindaklanjuti secara otomatis.

V. KESIMPULAN

Penelitian ini berhasil merancang dan membangun sebuah prototipe sistem *dashboard* fungsional untuk otomatisasi ABSA pada ulasan *video game*, dengan mengadopsi metodologi hibrida CRISP-DM dan Waterfall. Enam aspek analisis yang relevan (*Gameplay, Graphics, Story, Technical, World, Multiplayer*) telah diidentifikasi berdasarkan literatur, dan sebuah *dataset ground truth* berkualitas tinggi dengan 896 ulasan berhasil dibuat dengan validitas statistik yang sangat baik (Fleiss' Kappa > 0.85). Hasil eksperimen mengidentifikasi *4-Shot + Guideline* sebagai strategi *prompt* paling optimal. Di antara model LLM yang dievaluasi, Llama 4 Maverick terbukti menjadi yang paling unggul untuk kedua subtugas, dengan perolehan *Macro F1-Score* 0.9055 untuk ACD dan 0.8373 untuk ACSC. Kombinasi optimal ini telah berhasil diimplementasikan ke dalam prototipe yang teruji, membuktikan kelayakan pendekatan ini untuk menjembatani riset akademis dengan aplikasi industri praktis.

REFERENSI

[1] "Meet the Moment: How Gamers Are Changing the Game," 2024.

[2] T. Kraemer, W. H. Weiger, and S. Heidenreich, "Do all stars shine the same? Investigating the nonlinear effects of user and critic reviews on video game sales," *J Bus Res*, vol. 188, Feb. 2025, doi: 10.1016/j.jbusres.2024.115034.

[3] M. Philp and M. V. Nepomuceno, "How reviews influence product usage post-purchase: An examination of video game playtime," *J Bus Res*, vol. 172, Feb. 2024, doi: 10.1016/j.jbusres.2023.114456.

[4] A. Kosmopoulos, A. Liapis, G. Giannakopoulos, and N. Pittaras, "Summarizing Game Reviews: First Contact," 2020. [Online]. Available: www.metacritic.com

[5] Y. C. Hua, P. Denny, K. Taskova, and J. Wicker, "A Systematic Review of Aspect-based Sentiment Analysis: Domains, Methods, and Trends," Nov. 2023, doi: 10.1007/s10462-024-10906-z.

[6] I. A. Kandhro, F. Ali, M. Uddin, A. Kehar, and S. Manickam, "Exploring aspect-based sentiment analysis: an in-depth review of current methods and prospects for advancement," Jul. 01, 2024, *Springer Science and Business Media Deutschland GmbH*. doi: 10.1007/s10115-024-02104-8.

[7] P. F. Simmering and P. Huoviala, "Large language models for aspect-based sentiment analysis," no. 2022, pp. 1–12, 2023, [Online]. Available: http://arxiv.org/abs/2310.18025

[8] K. Scaria, H. Gupta, S. Goyal, S. A. Sawant, S. Mishra, and C. Baral, "InstructABSA: Instruction Learning for Aspect Based Sentiment Analysis," *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2024*, vol. 2, pp. 720–736, 2024, doi: 10.18653/v1/2024.naacl-short.63.

[9] H. T. Ismet, T. Mustaqim, and D. Purwitasari, "Aspect Based Sentiment Analysis of Product Review Using Memory Network," *Scientific Journal of Informatics*, vol. 9, no. 1, pp. 73–83, May 2022, doi: 10.15294/sji.v9i1.34094.

[10] W. Zhang, X. Li, Y. Deng, L. Bing, and W. Lam, "A Survey on Aspect-Based Sentiment Analysis: Tasks, Methods, and Challenges," *IEEE Trans Knowl Data Eng*, vol. 35, no. 11, pp. 11019–11038, 2023, doi: 10.1109/TKDE.2022.3230975.

[11] Y. Li *et al.*, "Better Queries for Aspect-Category Sentiment Classification," *19th Chinese National Conference on Computational Linguistic, CCL 2020*, no. c, pp. 1079–1088, 2020.

[12] D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural Language Processing : State of The Art , Current Trends and Challenges," 2023.

[13] H. Naveed *et al.*, "A Comprehensive Overview of Large Language Models," 2023, [Online]. Available: http://arxiv.org/abs/2307.06435

[14] F. Adıgüzel, "The Effect of YouTube Reviews on Video Game Sales," *Journal of Business Research - Turk*, vol. 13, no. 3, pp. 2096–2109, Sep. 2021, doi: 10.20491/isarder.2021.1249.

[15] M. Viggiato, D. Lin, A. Hindle, and C. P. Bezemer, "What Causes Wrong Sentiment Classifications of Game Reviews," *IEEE Trans Games*, 2021, doi: 10.1109/TG.2021.3072545.

[16] 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 *Proceedings of the International Conference on Electrical Engineering and Informatics*, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/ICEEI59426.2023.10346219.

[17] N. Khanh and H. Tran, "BACHELOR THESIS Development of a Machine Learning System for Aspect-Based Sentiment Analysis and Text Summarization of Video Game Reviews on Steam." [Online]. Available: https://store.steampowered.com/

[18] M. D. Purbolaksono, U. K. Dewi, R. L. Wicaksono, and A. P. Wibowo, "Sentiment Analysis of Game Review using Random Forest," vol. 10, no. 2, pp. 161–169, 2024, doi: 10.21108/ijoict.v10i2.1007.

[19] R. Wirth and J. Hipp, "CRISP-DM: Towards a Standard Process Model for Data Mining," 2000.

[20] S. Mudassar, "Waterfall Model Used in Software Development Reference: Software Requirements Engineering Waterfall Model," *ResearchGate*, no. June, pp. 2–4, 2023, doi: 10.13140/RG.2.2.29580.69764.

[21] S. Heidenreich, F. Handrich, and T. Kraemer, "Flawless victory! Investigating search and experience qualities as antecedent predictors of video game success," *Electronic Markets*, vol. 33, no. 1, Dec. 2023, doi: 10.1007/s12525-023-00647-2.

[22] E. Guardiola, "Gameplay definition: A game design perspective," *20th International Conference on Intelligent Games and Simulation, GAME-ON 2019*, no. October, pp. 5–10, 2019.

[23] M. S. Balalaa, "Video Game Graphics: A Comprehensive Analysis of Styles and Techniques," vol. 13, no. 3, 2023.

[24] E. Aarseth, "A narrative theory of games," *Foundations of Digital Games 2012, FDG 2012 - Conference Program*, no. October, pp. 129–133, 2012, doi: 10.1145/2282338.2282365.

[25] A. Wrześniewska and M. Skublewska-Paszkowska, "Video game performance analysis on selected operating systems Analiza wydajności gier komputerowych na wybranych systemach operacyjnych," vol. 29, no. June, pp. 317–324, 2023.

[26] L. Caroux, "Presence in video games: A systematic review and meta-analysis of the effects of game design choices," *Appl Ergon*, vol. 107, no. February, 2023, doi: 10.1016/j.apergo.2022.103936.

[27] J. Opitz, "A Closer Look at Classification Evaluation Metrics and a Critical Reflection of Common Evaluation Practice," *Trans Assoc Comput Linguist*, vol. 12, no. 2018, pp. 820–836, 2024, doi: 10.1162/tacl_a_00675.