Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 3, 2987-2992
2025
Publisher: Learning Gate
DOI: 10.55214/25768484.v9i3.5898

Enhancing emotion analysis through valence-arousal modeling with user and AI evaluations

Wookyaung Jin[1], Hyunjung Kim[2*]
[1]Konkuk University, The Graduate School of Information & Communication, Department of Convergence Information Technology (Artificial Intelligence Major), Seoul, Republic of Korea.
[2]Konkuk University, Sanghuh College and The Graduate School of Information & Communication, Department of Convergence Information Technology (Artificial Intelligence Major), Seoul, Republic of Korea; nygirl@konkuk.ac.kr (H.J.K.).

Abstract: With the advancement of text mining and natural language processing technologies, sentiment analysis has found widespread application across various fields. However, current research often emphasizes binary or multi-class classifications, which fail to capture the full spectrum of human emotions. To address this issue, the valence-arousal (VA) model has been proposed but encounters challenges such as data imbalance and subjective labeling. This study presents a novel approach that integrates large language models with user evaluations and employs relevant augmented generation techniques to enhance data quality and consistency. In addition, the VA data is visualized to assess its utility in multidimensional sentiment analysis. Future research will focus on expanding the dataset and conducting in-depth analyses to further validate the proposed approach.

Keywords: *Data augmentation, Large language models, Retrieval-augmented generation, Sentiment analysis, Valence-arousal.*

1. Introduction

Text mining and natural language processing are essential in various industries, including customer feedback analysis, trend monitoring, and mental health tracking [1-3]. Among these applications, sentiment analysis has become a significant technology for understanding human emotions and supporting decision-making processes. Traditional sentiment analysis methods rely on binary or multi-class classifications but often fail to capture the intricate spectrum of human emotions [4, 5].

To address these limitations, the valence-arousal (VA) model has been introduced [6]. This model quantifies emotions across two dimensions: valence (the positive-negative axis) and arousal (the activation axis), thus providing a more nuanced framework for analyzing the complex nature of emotions. Despite its promise, the VA model faces significant challenges. Accurately quantifying emotions requires precise data labeling. However, the inherent subjectivity and complexity of emotions lead to considerable variations among individual evaluations [7]. This subjectivity threatens data consistency, negatively affecting model performance and scalability [8]. Furthermore, existing emotion datasets tend to be biased toward high-arousal emotions (e.g., anger, joy), whereas low-arousal emotions (e.g., calmness, sadness) are often underrepresented [9]. This imbalance restricts the ability of models to effectively learn and predict low-arousal emotions.

This study proposes an innovative approach that integrates user evaluations with large language models (LLMs) to tackle existing challenges. By combining user-provided sentiment scores with automated evaluations generated by LLMs, this methodology enhances data reliability and consistency. In addition, the incorporation of retrieval-augmented generation (RAG) techniques enhances the contextual robustness and reliability of sentiment evaluations performed by Lewis, et al. [10]. To

History: Received: 28 January 2025; Revised: 20 March 2025; Accepted: 25 March 2025; Published: 29 March 2025
* Correspondence: nygirl@konkuk.ac.kr

validate its effectiveness, the study visualizes VA data, uncovering distinct patterns such as V- and U-shaped distributions [11, 12]. And focuses on optimizing model performance for broader applications.

2. Valence-Arousal-Based Bata Processing and Optimization Methodology

This study introduces a robust methodology for generating highly reliable datasets through iterative VA evaluations and emotion reassessment processes, as illustrated in Fig. 1. The methodology incorporates the RAG model during the VA reassessment and emotion evaluation phases to enhance the consistency of evaluation results and address data scarcity challenges. This integration not only supports existing evaluation methods but also improves the overall quality of emotion datasets. The detailed steps of the proposed approach are systematically outlined and analyzed.

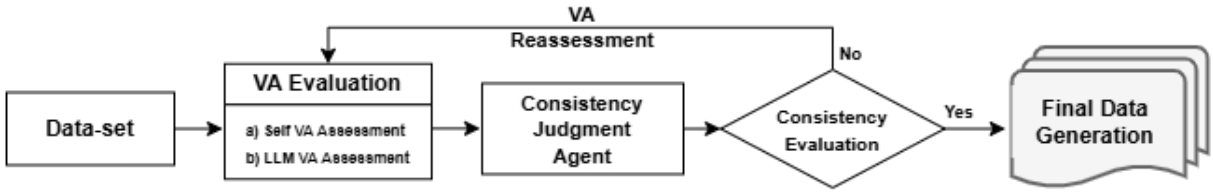

Figure 1.
VA evaluation work flow.

2.1. Datasets

The original dataset was sourced from the AI Hub Wellness Dialogue Scripts, which contains approximately 20,000 sentences related to mental health counseling, annotated with 50 distinct emotion tags. This dataset features Korean language sentences, primarily predicted to exhibit low-valence emotional states [13]. The inclusion of these data ensures diversity and enhances the reliability of emotion analysis, providing a solid foundation for sentiment evaluation and related research applications.

2.2. VA Evaluation

In this approach, both self-assessment and LLM-based evaluation methods are utilized to assess VA scores with consistency and efficiency.

2.2.1. Self-VA Assessment

Each sentence is subjectively evaluated for VA scores according to established VA guidelines (see Figure 2), following methodologies from prior studies [14]. Scores are assigned in increments of 0.5, and each sentence undergoes two consecutive evaluations to minimize variability. Following these evaluations, data with discrepancies exceeding a 0.5 threshold between the two scores is removed to ensure consistency in the labeling process. To confirm the reliability of the evaluations, the remaining data are then verified using a Cohen's Kappa score of 0.8 or higher. Cohen's Kappa is a statistical measure that assesses the consistency of a single evaluator's labeling across repeated assessments, ensuring robust and reliable data [15]. Once verified, the final score for Self-VA assessment is calculated by averaging the retained data.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 3: 2987-2992, 2025
DOI: 10.55214/25768484.v9i3.5898

Valence	Guidelines	Example
Strong Positive (+1)	The expression of strong positive emotions is clearly evident (e.g., joy, satisfaction, excitement).	"정말 멋진 일이에요!", "이렇게 기쁘기는 처음이에요."
Positive (+0.5)	Positive emotions are noticeable, but their intensity remains moderate.	"그거 참 좋은 생각이네요.", "나쁘지 않은데요."
Neutrality (0)	No distinct positive or negative emotions are conveyed. Statements are factual or involve simple inquiries.	"알겠어요.", "그렇군요."
Negative (-0.5)	Mild dissatisfaction, disappointment, or unease is present, but with low intensity.	"초조하네요.", "기대보다는 덜하네요."
Strong Negative (-1)	The expression of strong negative emotions is clearly evident, characterized by high intensity (e.g., frustration, irritation, anger).	"정말 짜증 나네요.", "다신 하고 싶지 않아요."

Arousal	Guidelines	Example
Strong Arousal (+1)	The expression of highly intense emotional energy is evident, characterized by a high activation level and strong emotional engagement.	"와, 이건 정말 대단한걸요!", "말도 안 돼요!"
Medium Arousal (+0.5)	Moderate levels of excitement or interest are present, but the intensity remains controlled and not overwhelming.	"오, 그거 좋네요.", "아, 재밌겠어요.."
Neutrality (0)	A state of emotional equilibrium, where no noticeable increase or decrease in energy levels is observed.	"알겠습니다.", "아무생각이 안나네요."
Low Arousal (-0.5)	Slightly reduced energy levels, often associated with mild fatigue, boredom, or low engagement.	"몸에 힘이 없어요.", "피곤하네요."
Very low Arousal (-1)	A state of emotional exhaustion or helplessness, characterized by minimal speech energy and a lack of motivation.	"암울하네요....", "무기력하네요...."

Figure 2.
Evaluation guidelines.

2.2.1. LLM VA Assessment

The LLM ChatGPT-4 assesses VA scores by following the same VA guidelines used in the self-assessment process. The model independently assigns scores within a range from −1 to 1, ensuring consistency with the established evaluation framework. This method produces a varied distribution of VA scores, capturing different emotional states and enabling a more comprehensive analysis of the emotional spectrum. The results of the evaluation are presented in JSON format (Fig. 3), which facilitates efficient data management and systematic analysis.

[{"sentence": "진짜 암울해", "valence": -0.9, "arousal": -0.6}, {"sentence": "오빠 때문에 침울해", "valence": -0.8, "arousal": -0.5} ...]

Figure 3.
JSON output.

2.3. Consistency Judgment Agent

If the error between the Self- and LLM-VA assessments remains within 0.5, the data is deemed consistent and progresses to the next processing stage. However, if the error exceeds the 0.5 threshold, the corresponding sentence is flagged to undergo the VA reassessment phase to ensure accuracy and reliability.

2.4. VA Reassessment

In this phase, the VA scores of data identified as inconsistent during the initial evaluation are reassessed. This process incorporates user evaluations, LLM evaluations, and existing emotion tags to enhance the robustness of the scoring. ChatGPT-4 is utilized to recalculate VA scores, employing prompts that include similar sentence data to capture subtle emotional differences and improve the accuracy of the initial evaluations (see Fig. 4).

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 3: 2987-2992, 2025
DOI: 10.55214/25768484.v9i3.5898

Category	Details
Sentence Data set	1. "과거생각하면 우울해지고 미래도 걱정돼." → VA(-0.8, -0.6) 2. "혼자 있으면 진짜 암울해서 담배가 땡겨" → VA(-0.5, -0.4) 3. "암울하니까 죽고 싶어" → VA(-1, -0.7) 4. "마음이 무너지는 것 같아" → VA(-0.9, -0.4) ...
Input Sentence	"너무 암울한데 어떻게 하면 좋을까"
VA Evaluation	나의 평가 : (-0.8, -0.9) ChatGPT 평가 : (-0.5, -0.6)
Similar Sentence	1. "암울하니까 죽고 싶어" → VA(-1, -0.7) 2. "과거 생각하면 우울해지고 미래도 걱정돼." → VA(-0.8, -0.6)
Reassess VA	Final VA value : (-0.9, -0.6)
Reason	**Valence** : The input sentence is similar to the VA(-1,-0.7) of "암울하니까 죽고 싶어" by using the direct and strong negative expression "너무 엄울하다" , but it also include the question **"어떻게 하면 좋을 까 "** Therefore, it is judged to be slightly more relaxed -0.9 than the extreme negative state (-1) **Arousal** : the input sentence emotionally expresses a gloomy state similar to the VA(-0.8,-0.6) of "과거 생각하면 우울해지고 미래도 걱정돼", but it also contains the will to solve "어떻게 하면", resulting in extremly low Arousal (-1) It is not at the level

Figure 4.
VA reassessment prompt.

Reassessment results undergo an additional consistency evaluation. In this phase, discrepancies between the Self- and LLM-VA evaluations are examined, and the reliability of each result is assessed based on supporting evidence. This process ensures that the reassessed data conforms to established standards for accuracy and reliability.

2.5. Final Data Generation

Data that successfully passes all consistency checks is used to generate optimized emotion datasets by combining sentence data with final VA evaluation results. The finalized data is systematically organized to ensure both reliability and efficiency, providing a strong foundation for emotion analysis. The optimized dataset improves the accuracy of emotion models and supports the development of consistent and reliable analytical processes.

3. Emotion Data Analysis via VA Weights

A dataset was constructed following the proposed workflow, and VA values derived from weighted evaluations were visualized to analyze the emotional spectrum. The final dataset, which comprises VA scores from user assessments and ChatGPT-4 evaluations, was utilized to generate weighted average VA values using different weight combinations (0.5, 0.9). Fig. 5 presents the VA distribution generated from these weighted averages, where the X and Y axes representing valence and arousal, respectively.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 3: 2987-2992, 2025
DOI: 10.55214/25768484.v9i3.5898

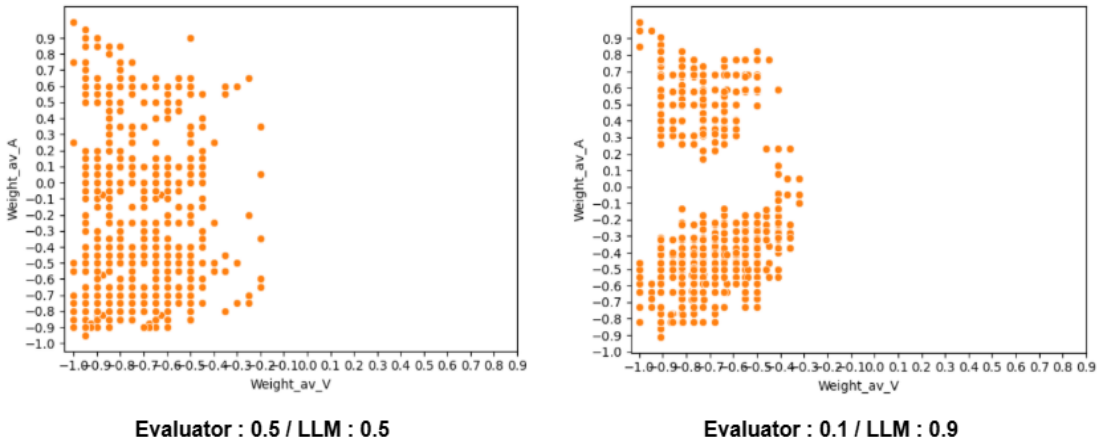

Figure 5.
Weighted evaluation metrics visualization.

Previous studies have highlighted the challenges of capturing the full emotional spectrum in VA evaluations [16]. They have also noted the emergence of distinct patterns, including V- and U-shaped distributions, in the evaluation results[11, 12]. In this study, equal weights (0.5) applied to both user and model evaluations resulted in a balanced integration of subjective user assessments and model predictions, leading to a richer emotional spectrum. By contrast, when user evaluations were heavily weighted (0.9) relative to model evaluations (0.1) the results closely resembled the V- or U-shaped distributions observed in previous research. These findings illustrate the adaptability of weighted evaluation methods in influencing the balance between subjective judgments and model predictions. By adjusting the weight assignments, this approach enables a more extensive exploration of emotional data dynamics and provides a robust framework for customizing evaluation methods to meet specific research objectives.

4. Conclusion

This study presents a VA-based data processing framework that integrates user evaluations with ChatGPT-4 using weighted methodologies to produce reliable and robust emotion datasets. The weighted evaluation approach reveals significant interactions between user and model assessments, which can lead to V- or U-shaped patterns when model evaluations are assigned higher weights. By effectively balancing subjective and automated assessments, this framework improves the reliability and diversity of emotion datasets, paving the way for advanced applications in sentiment analysis and emotional dialogue systems. Future research will focus on expanding the dataset to include high-valence regions and evaluating the performance of augmented datasets across various machine learning models to further enhance the analysis of complex emotional states.

Transparency:

The authors confirm that the manuscript is an honest, accurate, and transparent account of the study; that no vital features of the study have been omitted; and that any discrepancies from the study as planned have been explained. This study followed all ethical practices during writing.

Copyright:

References

[1] J.-K. Sung, Y. B. Kim, and Y.-G. Kim, "Deep learning-based Multilingual Sentimental Analysis using English Review Data," *The Journal of The Institute of Internet, Broadcasting and Communication*, vol. 19, no. 3, pp. 9-15, 2019.

[2] A. Biró, A. I. Cuesta-Vargas, and L. Szilágyi, "Precognition of mental health and neurogenerative disorders using AI-parsed text and sentiment analysis," *Acta Universitatis Sapientiae*, vol. 15, no. 2, pp. 359-403, 2023.

[3] T. Matsuzaki, H. Mizoguchi, and K. Yamada, "The use of text mining to obtain a historical overview of research on therapeutic drug monitoring," *Biological and Pharmaceutical Bulletin*, vol. 47, no. 11, pp. 1883-1892, 2024, doi: https://doi.org/10.1248/bpb.b24-00319.

[4] R. Gul and M. Bashir, "Feature selection for sentiment analysis using hybrid multiobjective evolutionary algorithm," *Journal of Intelligent & Fuzzy Systems*, no. Preprint, pp. 1-16, 2024, doi: https://doi.org/10.3233/jifs-234615.

[5] L. R. Krosuri and R. S. Aravapalli, "Novel heuristic-based hybrid ResNeXt with recurrent neural network to handle multi class classification of sentiment analysis," *Machine Learning: Science and Technology*, vol. 4, no. 1, p. 015033, 2023, doi: https://doi.org/10.1088/2632-2153/acc0d5.

[6] J. A. Russell, "A circumplex model of affect," *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980, doi: https://doi.org/10.1037/h0077714.

[7] K. K. Imbir, J. Duda-Goławska, A. Wielgopolan, A. Sobieszek, M. Pastwa, and J. Zygierewicz, "The role of subjective significance, valence and arousal in the explicit processing of emotion-laden words," *PeerJ*, vol. 11, p. e14583, 2023, doi: https://doi.org/10.7717/peerj.14583.

[8] E. T. Pereira and H. M. Gomes, "The role of data balancing for emotion classification using EEG signals," presented at the In 2016 IEEE International Conference on Digital Signal Processing (DSP) (pp. 555-559). IEEE, 2016.

[9] A. Dubois, C. L. Azevedo, S. Haustein, and B. Miranda, "Deep-seeded Clustering for Unsupervised Valence-Arousal Emotion Recognition from Physiological Signals," *arXiv preprint arXiv:2308.09013*, 2023.

[10] P. Lewis *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in neural information processing systems*, vol. 33, pp. 9459-9474, 2020.

[11] M. Yik *et al.*, "On the relationship between valence and arousal in samples across the globe," *Emotion*, vol. 23, no. 2, p. 332, 2023, doi: https://doi.org/10.1037/emo0001095.

[12] A. Toet *et al.*, "The relation between valence and arousal in subjective odor experience," *Chemosensory Perception*, vol. 13, pp. 141-151, 2020, doi: https://doi.org/10.1007/s12078-019-09275-7.

[13] B. Shickel, M. Heesacker, S. Benton, and P. Rashidi, "Automated emotional valence prediction in mental health text via deep transfer learning," presented at the In 2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE) (pp. 269-274). IEEE, 2020.

[14] H.-M. Park, C.-H. Kim, and J.-H. Kim, "Generating a Korean sentiment lexicon through sentiment score propagation," *KIPS Transactions on Software and Data Engineering*, vol. 9, no. 2, pp. 53-60, 2020.

[15] Z. Xie, C. Gadepalli, and B. M. Cheetham, "Measurement of rater consistency by chance-corrected agreement coefficients," presented at the In 2018 UKSim-AMSS 20th International Conference on Computer Modelling and Simulation (UKSim) (pp. 14-20). IEEE, 2018.

[16] A. Toisoul, J. Kossaifi, A. Bulat, G. Tzimiropoulos, and M. Pantic, "Estimation of continuous valence and arousal levels from faces in naturalistic conditions," *Nature Machine Intelligence*, vol. 3, no. 1, pp. 42-50, 2021, doi: https://doi.org/10.1038/s42256-020-00280-0.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 3: 2987-2992, 2025
DOI: 10.55214/25768484.v9i3.5898