Enhancing VADER Sentiment Analysis Through Lexicon Expansion with BERT Word Embeddings

Fritzie T. Nuñez
Department of Computer Science, College of Computing and Information Sciences, Caraga State University

Nerissa L. Baradillo
Department of Computer Science, College of Computing and Information Sciences, Caraga State University

Cristopher C. Abalorio
Department of Computer Science, College of Computing and Information Sciences, Caraga State University

https://doi.org/10.5109/7395650

出版情報：Proceedings of International Exchange and Innovation Conference on Engineering & Sciences (IEICES). 11, pp.1109-1115, 2025-10-30. International Exchange and Innovation Conference on Engineering & Sciences
バージョン：

Enhancing VADER Sentiment Analysis Through Lexicon Expansion with BERT Word Embeddings

Fritzie T. Nuñez[1] , Nerissa L. Baradillo[2], Cristopher C. Abalorio[3]
[1,2,3]Department of Computer Science, College of Computing and Information Sciences,
Caraga State University-Main Campus, Butuan City, Philippines
ccabalorio@carsu.edu.ph

Abstract: *This paper focuses on enhancing the widely used Valence Aware Dictionary and sEntiment Reasoner (VADER), which was introduced by Hutto and Gilbert in 2014. VADER works well for short texts like product reviews and social media. Nevertheless, its static lexicon restricts its capacity to interpret new slang as well as emerging language trends, especially prevalent among online consumers. To overcome this limitation, we enlarged the lexicon of VADER based on BERT word embeddings with cosine similarity. This enables the model to leverage the contextual sense of contemporary expressions and deduce sentiment from out-of-vocabulary (OOV) words that are usually ignored by rule-based systems. The addition of BERT enables the improved model to learn more effectively about subtle and new language, enhancing its performance in sentiment analysis tasks without needing to retrain a deep learning model from the beginning.*

Keyword: VADER, BERT word-embeddings, Cosine Similarity, Sentiment Analysis

1. INTRODUCTION

The rapid growth of e-commerce platforms has resulted in an enormous volume of user reviews, making sentiment analysis a valuable tool to interpret consumer opinions and feedback [1][2] . One widely used tool for sentiment analysis is VADER (Valence Aware Dictionary and sEntiment Reasoner), developed by Hutto and Gilbert (2014), which is particularly effective for analyzing shorter texts such as product reviews and social media posts. Although efficient, VADER is limited by its static, hand-crafted lexicon, which hinders its ability to process contemporary slang, abbreviations, and evolving language patterns often used by digital-native consumers like Gen Z [3][4][5].

The objective of this study is to address that limitation by incorporating BERT (Bidirectional Encoder Representations from Transformers) word embeddings into VADER’s pipeline [6][7][8]. By expanding VADER’s vocabulary using cosine similarity between BERT-generated embeddings and existing sentiment words, the improved model gains the ability to assign sentiment to out-of-vocabulary (OOV) terms and interpret subtle expressions [9]. This hybrid approach retains VADER’s rule-based transparency and computational speed while introducing the contextual understanding of transformer-based models like BERT. In doing so, it enhances sentiment classification accuracy on modern, user-generated content without sacrificing interpretability.

As far as modern e-commerce websites and online feedback spaces are concerned, the precision of sentiment analysis systems matters significantly in informing business strategy, improving product development, and maximizing customer satisfaction [10]. Effective interpretation of customer feedback allows companies to identify market trends, build a clearer picture of pain areas, and tailor services or products to meet evolving consumer demands [11]. However, as language usage continues to evolve particularly among younger generations such as Gen Z, who tend to use internet slang, emojis, neologisms, and culturally changing expressions traditional rule-based sentiment analyzers such as VADER become progressively limited. These models rely on a static lexicon that cannot adapt to newly coined words, domain-specific vocabulary, or phonetically creative misspellings [12].

To achieve this crucial gap, our contribution proposes a sentiment analysis hybrid that combines the interpretive power of VADER and contextual sensitivity of BERT. Specifically, we enrich VADER's fixed lexicon with dynamic sentiment scores derived from dynamically calculated sentiment scores based on BERT embeddings. Using cosine similarity, we map unseen or out-of-vocabulary words to their semantically identical counterparts present in VADER's dictionary, and thus they have their rightful sentiment scores. Our approach retains VADER's defining strengths of transparency, interpretability, and efficiency in computation but adds a layer of semantic flexibility that accommodates linguistic inventiveness and user diversity.

To do so, we reconcile rule-based accuracy with the context adaptability of deep learning. Our system answers not just more accurately to free or creative text commonly found in user-generated content but also improves overall sentiment classification performance on traditional and dynamic test sets. Most significantly, our work contributes to constructing stronger and forward-looking sentiment analysis solutions for practical use in the real world.

2. RELATED WORK

2.1 Sentiment Analysis

Sentiment analysis has become a basic tool for gauging public opinion and user sentiment, especially in areas like product reviews, social media, and customer comments. Initial sentiment analysis models used machine learning models trained on tagged datasets [13][14]. Recent lexicon-based methods like VADER (Valence Aware Dictionary and sEntiment Reasoner), presented by Hutto and Gilbert in 2014, became more prevalent because they were easier to use, faster, and more effective at processing short, casual text [12].

2.2 Valence Aware Dictionary and sEntiment Reasoner (VADER)

VADER employs a rule-based approach with a fixed vocabulary of English words, each of which is tagged with a sentiment polarity score. It also employs heuristics to deal with negations, punctuation, and intensifiers [15]. VADER is good on general and informal text but is weaker in identifying sentiment in new slang, new expressions, and context-dependent language [16]. This weakness shows more clearly in dynamic web-based environments like social media and online reviews.

To enhance contextual comprehension, word embeddings have been investigated in recent studies using transformer-based models such as BERT (Bidirectional Encoder Representations from Transformers) [17][18]. BERT produces contextualized word representations that reflect the semantic meaning of words given their context. These embeddings have demonstrated robust performance in a range of NLP tasks, including sentiment classification.

Some have employed BERT for end-to-end sentiment classification, while others have investigated embedding-based lexicon expansion. Few, though, have tried to incorporate BERT into rule-based systems such as VADER directly. This work fills that gap by introducing a hybrid approach: it detects out-of-vocabulary words in VADER, calculates their BERT embeddings, and applies cosine similarity to give them sentiment scores based on proximity to known lexicon entries. This method enables the system to learn to accommodate new language patterns without having to retrain a deep learning system.

2.3 Cosine Similarity

Cosine similarity is widely applied in NLP to estimate semantic proximity among word vectors. Cosine similarity returns a value from –1 (completely opposite) to +1 (exactly the same) by calculating the cosine of the angle between two vectors in high-dimensional space. In this research, cosine similarity facilitates matching between VADER lexicon entries and newly added slang or OOV words embedded with BERT. For instance, if "fireee" and "amazing" are highly similar in cosine similarity, "fireee" copies the sentiment polarity of "amazing."

Cosine Similarity Formula:

$$cosine_{similarity(x,y)} = \frac{x * y}{||x||\,||y||}$$

3. METHODOLOGY

The paper suggests a hybrid model for sentiment analysis that improves the base VADER lexicon with BERT word embeddings and cosine similarity. The methodology is broken down into various phases: data gathering, preprocessing, sentiment tagging using various methods, lexicon extension, and performance assessment.

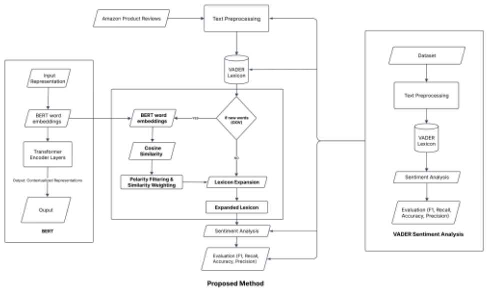

Fig 1. Conceptual Framework.

Figure 1 illustrates the conceptual framework of the suggested method. It starts with collecting Amazon product reviews, text preprocessing, and baseline sentiment analysis with the VADER model. Concurrently, the hybrid model detects out-of-vocabulary (OOV) words, produces BERT embeddings, calculates their cosine similarity with known sentiment-carrying words, and enriches the VADER lexicon with induced sentiment scores. The enriched vocabulary is subsequently employed for rescoring sentiment, which is measured by the usual metrics: accuracy, precision, recall, and macro-averaged F1-score.

VADER, in its native state, works with a static dictionary and rule-based grammar, serving well for short and casual texts but weak for working with changing or colloquial text. To counter this weakness, the lexicon is dynamically extended through BERT's contextualized embeddings so that a better semantic comprehension of OOV words can be achieved through cosine similarity. The resulting hybrid model combines the rule-based accuracy of VADER with the contextual understanding of BERT, making it better for use with informal text like product reviews.

The study integrates both qualitative and quantitative methods. Qualitative analysis studies the problem presented by slang and colloquialisms, whereas quantitative assessment quantifies model performance using the above-mentioned metrics. Implementation of the code is done in Python with the help of Google Colab and GPU acceleration for swift computation of BERT embeddings.

3.1 Data Collection

The data collection came from product reviews on Amazon and Kaggle and was in the form of English-language user reviews. They followed a train-test split with 70/30. There was automatic annotation for the test set via pre-trained BERT sentiment classification in Hugging Face. A script using Python balanced the sentiment class distribution.

3.2 Text Preprocessing

Preprocessing consisted of converting text to lowercase, deleting special characters, normalizing whitespace, and removing stop words. Although there was no direct slang normalization in preprocessing, informal words were subsequently handled via lexicon expansion via BERT.

3.3 Sentiment Analysis Techniques

- **Original VADER:** Syntactic rule-based lexicon model.
- **Expanded VADER:** The model proposed in this work, expanding sentiment scores on out-of-vocabulary words based on cosine similarity between BERT embeddings.

3.4 Hybrid Approach and Justification

This work introduces a hybrid sentiment model that enriches the baseline VADER lexicon with semantic content from BERT embeddings. In contrast to relying on BERT as an end classifier, this approach utilizes solely the pre-trained BERT word embeddings to infer sentiment for OOV words. The procedure follows:

Initial VADER Scoring: The VADER is first utilized to distribute the sentiment scores according to its pre-defined lexicon.

1. OOV Word Detection: Non-existent words in the VADER dictionary are marked for closer examination.
2. BERT Embedding Generation: Applying a pre-trained Hugging Face Transformers BERT model (no fine-tuning), contextualized word embeddings for OOV words as well as known entries of the lexicon are created.
3. Cosine Similarity Matching: Every OOV word embedding is matched against VADER lexicon word embeddings using cosine similarity to find semantically similar words.
4. Polarity Transfer and Weighting: Sentiment polarity is inferred for OOV words based on their most similar matches, applying a weighted average according to similarity scores.
5. Lexicon Expansion: The inferred OOV word sentiments are added to the VADER lexicon, forming an expanded lexicon tailored to the dataset.
6. Final Sentiment Scoring: Sentiment analysis is performed again using the expanded lexicon to improve classification accuracy.

By integrating BERT embeddings into the lexicon expansion process, the model is able to dynamically interpret modern, informal, or emerging language—especially slang and abbreviations that are common in user-generated content.

This study's hybrid approach, which combines the benefits of contextual and lexicon-based models.

1. Efficiency: It maintains VADER's compact rule-based architecture.
2. Contextual Power: It introduces the semantic adaptability of BERT without necessitating complete retraining.
3. Domain Adaptability: It enables recognition of new slang and domain-specific expressions.

3.5 Evaluation Metrics

The models were tested on accuracy, precision, recall, and macro-averaged F1-score. These were calculated across all three sentiment classes (positive, neutral, negative) to provide a balanced and unbiased comparison.

4. RESULTS AND DISCUSSION

The performance of the suggested expanded VADER sentiment analysis model was compared with the basic VADER using a balanced test set of Amazon product reviews. The two models were compared on how well they can classify the sentiments into positive, neutral, and negative categories.

4.1 Data Collection

The dataset used in this study was sourced from Kaggle and consisted of 568,454 raw Amazon product reviews. Each entry included the review text but lacked predefined sentiment labels (positive, negative, or neutral). To enable model evaluation without introducing human bias, a subset of 30% of the dataset was annotated automatically using a pre-trained sentiment classification model from Hugging Face. This labeled subset was then used to evaluate the original VADER tool, the BERT-based sentiment predictions, and the proposed hybrid model that combines VADER with lexicon expansion through BERT word embeddings and cosine similarity.

4.2 Performance Comparison

The table below presents the evaluation metrics for the original and expanded VADER models.

Table 1. the evaluation metrics for the original and expanded VADER models.

Model	Accuracy	Macro F1	Positive F1	Neutral F1	Negative F1
Original VADER	0.4991	0.4248	0.5941	0.1023	0.5781
Expanded VADER	0.4922	0.4352	0.6097	0.1690	0.5270

The original VADER model obtains an overall accuracy of 49.91% and a macro f1 score of 0.4248. The model performs decently well at detecting positive (F1 = 0.5941) and negative (F1 = 0.5781) sentiments, but really struggles with correctly classified Neutral sentiments (F1 = 0.1023) This emphasizes where VADER is weak; all of its words are built on a relatively simple dictionary and corpus, and in more complex entities like product review text the subtleties of language will often skew its performance.

These results highlight the necessity for advanced sentiment analysis methodologies that can account for OOV words and embrace context-based meaning which motivates the modifications detailed in the next section.

The improved VADER model achieves a higher macro-averaged F1 score with 0.4352 compared to the original baseline with 0.4248, showing a more balanced performance across sentiment classes. While the overall accuracy had a slight decrease on the result, having 0.4922 (Improved VADER) and 0.4991 (Original VADER), this kind of balance is common in tasks with more than two categories, where it's often more important to treat all categories fairly than to only focus on getting the highest accuracy.

The biggest improvement was in the Neutral F1-score, which went up from 0.1023 (Original VADER) to 0.1690 (Improved VADER). This helped fix one of the main identified limitations with the original VADER. The Positive F1-score also increased slightly from 0.5941 (Original VADER) to 0.6097 (Improved VADER), because the Improved lexicon now can understand modern slang and casual positive words like “aweeesome”, “slaps”, “fireee” and “goated”.

The F1-score for Negative sentiment dropped slightly (from 0.5781 to 0.5270), but that’s likely because the model became better at identifying Neutral sentiment. That’s a good sign, since it means the model is doing a better job at treating all classes more fairly.

Although overall accuracy was not greatly affected, the expanded VADER model had a better macro-averaged F1-score, especially in identifying neutral sentiments. This matters because neutral reviews tend to be subtle and hard to classify with rule-based models. Using BERT embeddings and cosine similarity, the expanded model could deduce sentiment for words and phrases that it hadn't previously been trained on, enhancing its semantic awareness.

4.3 Discussion

The findings show that incorporating contextual embeddings into a rule-based system can improve its capacity without compromising interpretability. The enhanced detection of neutral sentiment indicates that BERT effectively filled lexical gaps within the baseline VADER model. Though the improvements are incremental, they point to the extent to which hybrid models can respond to changing language patterns, especially in informal, user-generated sources such as product reviews.

The improved neutral F1-score of the larger model exhibits proper handling of ambiguous or subtle expressions. For example, BERT embeddings helped to correctly interpret reviews like "it's alright, nothing special" that early VADER labeled positive due to the occurrence of "alright." Also, informal terms like "mid" (negative) or "fireee" (extremely positive) were properly translated according to embedding proximity.

4.4 Analysis of Confusion Matrix

To improve the understanding of the classification ability of both models, Figure X is a side-by-side confusion matrix comparison of the baseline VADER and enhanced VADER using BERT lexicon expansion.

The original VADER's confusion matrix indicates that:

- It identified 79,572 positive examples correctly but identified a large portion of neutral (6,353) and negative (4,748) ones as positive incorrectly.
- It fared particularly poorly on neutral sentiments, accurately classifying only 635 of 25,000 neutral examples.

By contrast, the enhanced VADER:

- Cut the number of neutral samples incorrectly labeled as positive (to 2,499 from 6,353).
- Accurately classified 6,832 of 10,005 neutral examples a big improvement.
- Also demonstrated improved balance across all three classes, with a truer neutral and negative prediction profile.

These results corroborate the hypothesis that the expanded vocabulary strengthened by BERT embeddings and cosine similarity helped to identify fine as well as modern phrases more effectively and, more importantly, improved in recognizing neutral sentiment.

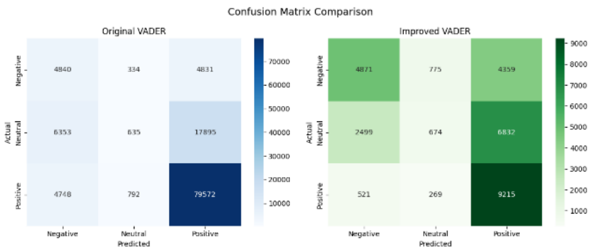

Fig 2. Confusion matrix comparison of original VADER and improved VADER.

5. CONCLUSION

This study shows that by expanding VADER’s original list of sentiment words using BERT word embeddings and cosine similarity, we were able to make it better at analyzing product reviews. Basically, we taught VADER to understand more words, especially modern or less common ones by linking them to similar words it already understands, based on how they’re used in real-life language.

The improved model, which combines both rule-based and context-aware techniques, helped fix some of VADER’s main weaknesses. It became more capable of handling new or evolving vocabulary, and it got better at recognizing when a sentence is neutral, something the original version often struggled with.

What makes this approach especially useful is that it strikes a good balance: it keeps VADER's simplicity and transparency, while adding some of the deeper understanding that comes from modern language models like BERT. And importantly, it does all this without needing to train a big neural network from scratch making it a practical and efficient upgrade.

In the end, this study shows that blending smart, context-aware tools with traditional methods can lead to meaningful improvements in sentiment analysis, especially when resources are limited, or interpretability is important.

6. REFERENCES

[1] Judith A Chevalier and Dina Mayzlin, "The Effect of Word of Mouth on Sales: Online Book Reviews," *J. Mark. Res.*, vol. 43, no. 3, pp. 345–354, Aug. 2006, doi: 10.1509/jmkr.43.3.345.

[2] R. Lakshmi and S. Baskar, "Novel term weighting schemes for document representation based on ranking of terms and Fuzzy logic with semantic relationship of terms," *Expert Syst. Appl.*, vol. 137, pp. 493–503, 2019, doi: 10.1016/j.eswa.2019.07.022.

[3] C. D. Suizo, "Gen Z's Language Identity and Contemporary Philosophies in the Digital Era," *J. Nat. Lang. Linguist.*, vol. 2, no. 1, pp. 13–17, May 2024, doi: 10.54536/jnll.v2i1.2396.

[4] M. Isnan, G. N. Elwirehardja, and B. Pardamean, "Sentiment Analysis for TikTok Review Using VADER Sentiment and SVM Model," *Procedia Comput. Sci.*, vol. 227, pp. 168–175, 2023, doi: https://doi.org/10.1016/j.procs.2023.10.514.

[5] N. P. Xuan and H. Le Quang, "A New Improved Term Weighting Scheme for Text Categorization," vol. 1, pp. 261–270, 2014, doi: 10.1007/978-3-319-02741-8.

[6] J. B. B. Escol, J. L. Guibone, W. K. E. Duenas, C. C. Abalorio, and J. C. M. Bustillo, "Optimizing Local Repository Search with Bidirectional Encoder Representations from Transformer Models," no. September, 2023, doi: 10.46338/ijetae0923.

[7] N. Al Hafidh and A. Al-Karawi, "Advanced Sentiment Analysis of Amazon Electronics Reviews Leveraging BERT: Model Optimization and Evaluation," *Procedia Comput. Sci.*, vol. 258, pp. 3608–3618, 2025, doi: https://doi.org/10.1016/j.procs.2025.04.616.

[8] N. Darraz, I. Karabila, A. El-Ansari, N. Alami, and M. El Mallahi, "Integrated sentiment analysis with BERT for enhanced hybrid recommendation systems," *Expert Syst. Appl.*, vol. 261, p. 125533, 2025, doi: https://doi.org/10.1016/j.eswa.2024.125533.

[9] T. P. Rinjeni, A. Indriawan, and N. A. Rakhmawati, "Matching Scientific Article Titles using Cosine Similarity and Jaccard Similarity Algorithm," *Procedia Comput. Sci.*, vol. 234, pp. 553–560, 2024, doi: https://doi.org/10.1016/j.procs.2024.03.039.

[10] M. Rui, A. Sparacino, V. M. Merlino, F. Brun, S. Massaglia, and S. Blanc, "Exploring consumer sentiments and opinions in wine E-commerce: A cross-country comparative study," *J. Retail. Consum. Serv.*, vol. 82, p. 104097, 2025, doi: https://doi.org/10.1016/j.jretconser.2024.104097.

[11] A. Daza, N. D. González Rueda, M. S. Aguilar Sánchez, W. F. Robles Espíritu, and M. E. Chauca Quiñones, "Sentiment Analysis on E-Commerce Product Reviews Using Machine Learning and Deep Learning Algorithms: A Bibliometric Analysis, Systematic Literature Review, Challenges and Future Works," *Int. J. Inf. Manag. Data Insights*, vol. 4, no. 2, p. 100267, 2024, doi: https://doi.org/10.1016/j.jjimei.2024.100267.

[12] M. Panagiotou and K. Gkatzionis, "Lexicon development to measure emotions evoked by foods: A review," *Meas. Food*, vol. 7, p. 100054, 2022, doi: https://doi.org/10.1016/j.meafoo.2022.100054.

[13] Y. Sakai, K. Takeuchi, and E. Ito, "A Study of Flaming Comment Detection using Text based machine learning," *Inf. Process. Soc. Japan*, vol. 203, no. 9, pp. 1–5, 2021.

[14] E. Ishita, S. Fukuda, Y. Tomiura, and D. W. Oard, "Using text classification to improve annotation quality by improving annotator consistency," *Proc. Assoc. Inf. Sci. Technol.*, vol. 57, no. 1, 2020, doi: 10.1002/pra2.301.

[15] J. de Jesús Estrella Ramírez and J. C. Gomez, "Evolutionary learning of selection hyper-heuristics for text classification," *Appl. Soft Comput.*, vol. 147, p. 110721, 2023, doi: https://doi.org/10.1016/j.asoc.2023.110721.

[16] G. Felici, "Talking to Sibilla: An approach to context dependent natural language comprehension," *Eur. J. Oper. Res.*, vol. 85, no. 2, pp. 263–281, 1995, doi: https://doi.org/10.1016/0377-2217(94)00045-E.

[17] M. Ciapparelli and A. B. T.-R. M. in S. S. Gregor de Varda, “The BERT Model Family and Encoder-Only Transformers,” Elsevier, 2025. doi: https://doi.org/10.1016/B978-0-323-95504-1.00507-X.

[18] S. L. Mirtaheri, A. Pugliese, N. Movahed, and R. Shahbazian, “A comparative analysis on using GPT and BERT for automated vulnerability scoring,” *Intell. Syst. with Appl.*, vol. 26, p. 200515, 2025, doi: https://doi.org/10.1016/j.iswa.2025.200515.