A Comprehensive Pipeline for Sentiment Analysis of Amazon Reviews Using Python and VADER

Vi Loi Truong[1], Thang Ha Viet[1], Tri Minh Duong[1], Vinh Phong Vo[1], Toan Tran Song[1], Thai Duy Pham[1]

[1] Dept. of Aritificial Inelligence, FPT University, Vietnam
ViTLCE180322@fpt.edu.vn
ThangHV.CE190509@gmail.com
TriDMCE181107@fpt.edu.vn
VinhVPCE180238@fpt.edu.vn
ToanTSCE180161@fpt.edu.vn
ThaiPDCE180857@fpt.edu.vn

Abstract. In recent years, sentiment analysis has emerged as a critical tool for interpreting customer opinions in the e-commerce landscape. This study investigates the application of the VADER (Valence Aware Dictionary and sEntiment Reasoner) model to analyze customer reviews on Amazon. VADER, a lexicon and rule-based sentiment analysis tool, is especially effective for short, informal texts such as product reviews. The methodology involves collecting review data, performing text preprocessing, scoring sentiments using VADER, and visualizing the sentiment distribution across product ratings. The results demonstrate that VADER can efficiently classify large-scale review datasets and provide meaningful insights into customer satisfaction and product perception. This paper contributes a practical framework for businesses seeking to leverage sentiment analysis for customer feedback interpretation and decision-making in the e-commerce environment.

Keywords: Sentiment Analysis, VADER, Amazon Reviews, Opinion Mining, Natural Language Processing (NLP), E-commerce

1 Introduction

In the era of e-commerce, online customer reviews have become a crucial factor influencing consumer purchasing decisions and business strategies. Among the various e-commerce platforms, Amazon hosts millions of customer reviews that provide rich textual data for understanding user sentiments [1]. Analyzing such reviews offers valuable insights into customer satisfaction, product quality, and service effectiveness.

Sentiment analysis, also known as opinion mining, is a natural language processing (NLP) technique used to identify and extract subjective information from text. It plays a vital role in interpreting the emotional tone behind textual content, especially in customer reviews [2]. Numerous approaches have been proposed for sentiment classification, ranging from machine learning models to deep learning techniques [3].

However, lexicon-based models remain popular due to their simplicity, efficiency, and ease of implementation on short informal texts [4].

VADER (Valence Aware Dictionary and sEntiment Reasoner) is a rule-based sentiment analysis tool specifically designed for social media and short-text content. Developed by Hutto and Gilbert [5], VADER uses a sentiment lexicon and a set of rules to account for nuances such as capitalization, punctuation, degree modifiers, and emoticons. It outputs polarity scores—positive, negative, neutral, and compound—that can be used to classify the sentiment of a given text snippet.

This paper explores the application of the VADER sentiment dictionary model in classifying customer reviews on the Amazon platform. The study aims to demonstrate the effectiveness of VADER in analyzing large-scale product review datasets by transforming raw textual content into meaningful sentiment scores. The methodology encompasses data collection, preprocessing, sentiment scoring, and visualization to uncover sentiment trends across product ratings. By leveraging the VADER model, this research contributes to the growing field of sentiment analysis in e-commerce and offers a practical framework for businesses seeking to monitor and interpret customer feedback efficiently.

2 Methodology and Result

2.1 Library Importation and Environment Setup

The implementation begins with the importation of essential Python libraries. Specifically, pandas and NumPy are employed for data manipulation and numerical operations [6], while matplotlib.pyplot and seaborn are used for generating and customizing visualizations [7][8]. To ensure stylistic consistency, the plotting style is set to 'ggplot'. Additionally, the script utilizes the nltk (Natural Language Toolkit) library for natural language processing [9] and shutil for file system operations. This modular setup reflects common practices in reproducible computational research.

2.2 NLTK Data Management

To ensure a clean and consistent environment for natural language processing (NLP) tasks, the script initiates by performing a preliminary check and removal of any outdated or potentially conflicting NLTK data. Specifically, the existing NLTK data directory (/root/nltk_data) is deleted using the shutil.rmtree function with the ignore_errors=True parameter, thereby preventing execution interruptions if the directory is not found. Following this cleanup step, the script proceeds to download the essential NLTK components required for subsequent analysis. These include: punkt, a tokenizer used for sentence segmentation; averaged_perceptron_tagger, a part-of-speech (POS) tagger; and vader_lexicon, the core lexicon necessary for executing sentiment analysis using the VADER model [5].

2.3 Data Ingestion and Preprocessing

The dataset, assumed to contain customer reviews from Amazon, is loaded using `pd.read_csv`. To ensure robustness, the parameter `on_bad_lines='skip'` is included to bypass corrupt or improperly formatted rows. The dataset's initial dimensions are printed for validation. For demonstrative purposes, the dataset is then truncated to the first 500 entries, and the revised shape is displayed (table 1).

Table 1. Example data

ProfileName	HelpfulnessNumerator	HelpfulnessDenominator	Score	Summary	Text
delmartian	1	1	5	Good Quanlity Dog Food	I have bought several of the Vitality canned d...
dll pa	0	0	1	Not as Advertised	Product arrived labeled as Jumbo Salted Peanut...
Natalia Corres "Natalia Corres"	1	1	4	"Delight" says it all	This is a confection that has been around a fe...
Karl	3	3	2	Cough Medicine	If you are looking for the secret ingredient i...
Michael D. Bigham "M. Wassir"	0	0	5	Great taffy	Great taffy at a great price. There was a wid...

2.4 Exploratory Data Analysis (EDA)

An initial exploratory data analysis (EDA) is conducted to examine the distribution of review scores. Using the `value_counts()` method, the script calculates the frequency of each star rating. A bar chart is generated to visually represent the distribution, with star ratings on the x-axis and corresponding frequencies on the y-axis. This analysis serves to assess the balance of classes in the dataset, which is crucial for downstream sentiment classification.

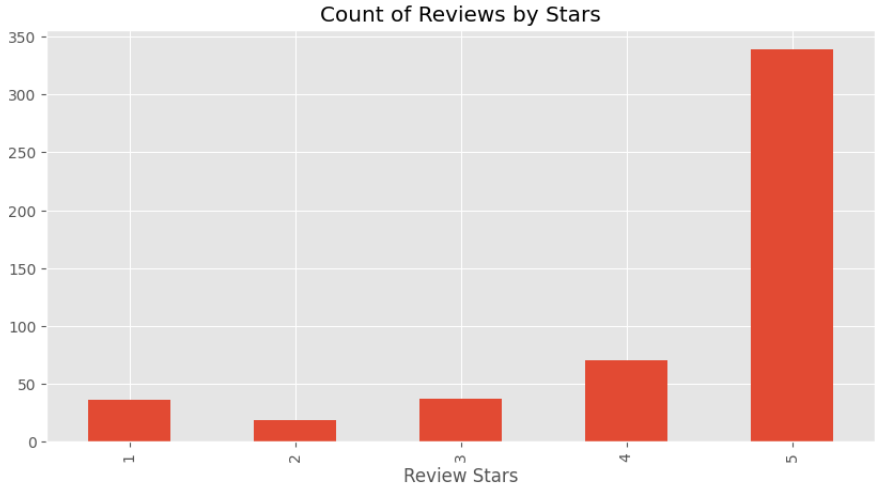

Fig. 1. This graph shows the number of reviews by stars

2.5 Sentiment Analysis Using VADER

The VADER (Valence Aware Dictionary and sEntiment Reasoner) model employs a lexicon and rule-based approach to perform sentiment analysis on social media and other informal texts. The diagram below illustrates the sequential process through which VADER analyzes and interprets sentiment from raw input text.

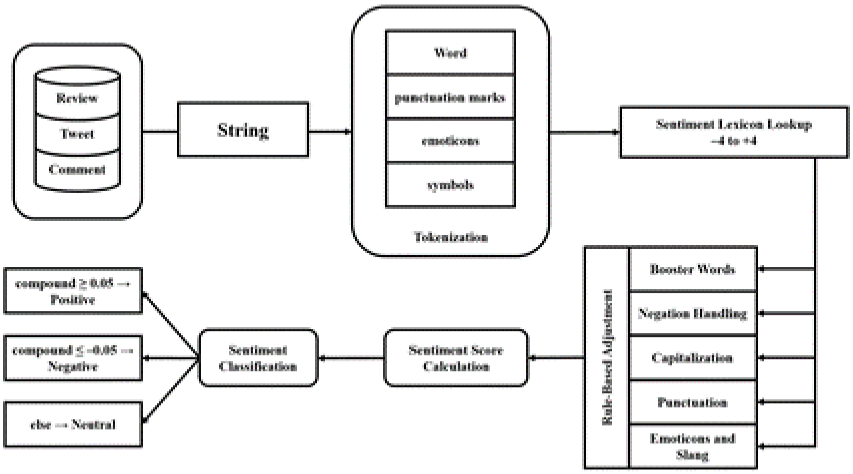

Fig. 2. Workflow of the VADER Sentiment Analysis Model.

The VADER (Valence Aware Dictionary and sEntiment Reasoner) model performs sentiment analysis through a lexicon- and rule-based approach, specifically optimized for social media and online text. The process begins with input text normalization, where raw textual data is analyzed directly without the need for traditional preprocessing techniques such as stemming, lemmatization, or stop-word removal. This characteristic enables VADER to preserve the contextual meaning embedded in complete sentences, reflecting its design to handle natural, informal language effectively.

The second phase is tokenization, in which the input text is segmented into individual elements including words, punctuation marks, emoticons, and symbols. Each token is treated independently, facilitating the subsequent sentiment lexicon lookup. This fine-grained decomposition allows VADER to capture sentiment nuances encoded not only in words but also in stylistic textual features.

Following tokenization, each token undergoes a sentiment lexicon lookup. VADER utilizes a predefined sentiment lexicon containing over 7,500 lexical features such as words, phrases, acronyms, and emojis, each associated with a valence score ranging from –4 (highly negative) to +4 (highly positive). For example, the word "excellent" carries a score of +3.1, while "bad" scores –2.5. This lexicon-driven approach enables the model to quantify the sentiment orientation of a given text accurately.

To refine sentiment detection, VADER incorporates a series of contextual rule-based adjustments. These heuristics account for the influence of linguistic and stylistic modifiers on sentiment polarity. First, booster words such as “very,” “extremely,” or “super” amplify the intensity of adjacent sentiment-bearing terms. Second, negation handling is implemented to invert the polarity of following words, as observed in constructions like “not good,” which is interpreted as negative despite the inherently positive valence of the word “good.” Third, capitalization is used to indicate emphasis;

for instance, “BAD” is scored more negatively than “bad.” Fourth, punctuation, particularly the use of exclamation marks, is factored into the sentiment calculation, enhancing the intensity of expressions such as “Great!!!.” Finally, VADER incorporates emoticons and internet slang, including elements like “:)”, “lol”, and “omg,” into its lexicon to support analysis of informal digital communication.

The final stage involves sentiment score computation, where the model outputs a dictionary comprising four sentiment metrics: pos (proportion of positive sentiment), neu (proportion of neutral sentiment), neg (proportion of negative sentiment), and compound. The compound score is a normalized, aggregated value ranging from –1 (most negative) to +1 (most positive), serving as the principal indicator for overall sentiment classification. Based on standard thresholds, a compound score ≥ 0.05 denotes a positive sentiment, ≤ –0.05 denotes a negative sentiment, and values in between are classified as neutral.

2.6 Batch Processing of Review Data

To scale the sentiment analysis across the entire dataset, the script employs an iterative process using `iterrows()` to traverse each review. A progress bar, facilitated by `tqdm`, provides real-time feedback on the analysis status. For each review (from the `Text` column), sentiment scores are computed and stored in a dictionary indexed by the review’s unique identifier (`Id`). This dictionary is then converted into a DataFrame, transposed, and merged with the original dataset using the `Id` column as the key. The resulting DataFrame integrates both review metadata and computed sentiment scores.

2.7 Visualization of Sentiment Analysis Results

Two primary visualizations are produced to interpret the sentiment analysis outcomes: **Aggregate Compound Score Visualization**: A bar plot illustrates the average compound sentiment score for each review star rating. This chart reveals the relationship between textual sentiment and numerical star ratings.

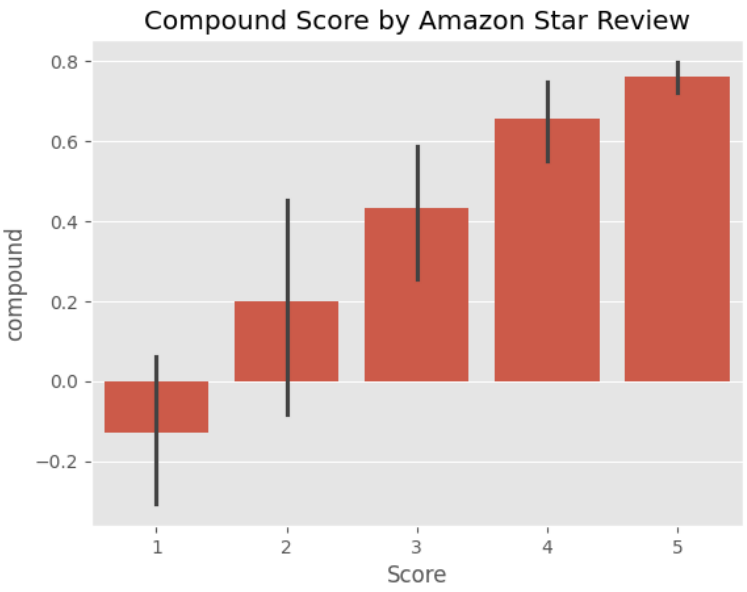

Fig. 3. Average VADER Sentiment Scores by Review Star

Table 2. Example of Average Sentiment Scores by Review Star

	compound	pos	neu	neg
Score				
1	-0.127569	0.073833	0.819583	0.106639
2	0.198178	0.116667	0.788889	0.094333
3	0.431411	0.128649	0.816486	0.054973
4	0.655927	0.188600	0.763829	0.047529
5	0.762773	0.218546	0.752587	0.028873

Figure 3 displays the average sentiment scores computed by the VADER model for Amazon product reviews, grouped by their corresponding star ratings (from 1 to 5). The metrics include the compound score (overall sentiment polarity), and the proportions of positive (pos), neutral (neu), and negative (neg) sentiment within each group.

In table 2, A clear ascending trend can be observed in the compound scores, rising from –0.1276 for 1-star reviews to +0.7628 for 5-star reviews. This steady increase strongly indicates that VADER’s sentiment scoring aligns closely with users’ explicit ratings, confirming its effectiveness in capturing overall sentiment polarity.

The positive sentiment proportion (pos) also increases progressively, from 0.0738 (1-star) to 0.2185 (5-star), while the negative sentiment (neg) shows a decreasing trend, from 0.1066 (1-star) to 0.0289 (5-star). Interestingly, the neutral sentiment proportion (neu) remains the dominant class across all ratings but slightly decreases as the review score increases, suggesting that highly positive reviews tend to include more emotionally expressive language.

These quantitative patterns reinforce the reliability of VADER in sentiment analysis tasks, especially when interpreting user-generated reviews, and validate its potential utility for automated sentiment-based review classification.

Component-wise Sentiment Analysis: A three-panel subplot depicts the distribution of positive, neutral, and negative sentiment components across varying review scores. This detailed visualization provides a nuanced understanding of sentiment composition, supplementing the aggregate compound analysis.

To explore the relationship between star ratings and the emotional tone of the reviews, sentiment scores were analyzed using the VADER (Valence Aware Dictionary and sEntiment Reasoner) model. The following visualizations illustrate the aggregate and component-wise sentiment scores across different review levels.

Figure 1 and Figure 2 present the sentiment distribution in both compound and component forms respectively.

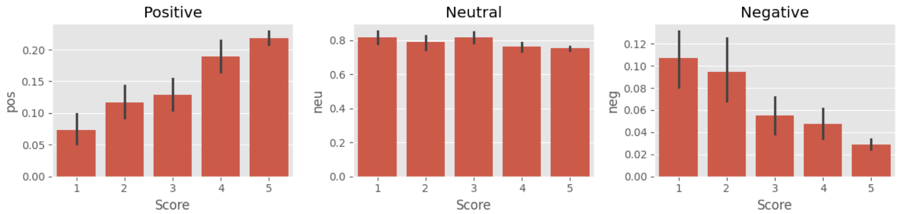

Fig. 4. Compound Sentiment Score by Amazon Star Review

The bar plot in Figure 4 illustrates the average compound sentiment score—a normalized metric ranging from -1 (most negative) to +1 (most positive)—across different star ratings from 1 to 5. A clear upward trend is observed: reviews with 1-star ratings have the lowest average compound score (-0.1276), while those with 5-star ratings reach the highest average score (0.7628). This consistent increase indicates a strong positive correlation between the numerical rating and the emotional tone expressed in the review content.

Table 3. Component-wise Sentiment Scores by Review Star

Score	pos	neu	neg
1	0.073833	0.819583	0.106639
2	0.116667	0.788889	0.094333
3	0.128649	0.816486	0.054973
4	0.188600	0.763829	0.047529
5	0.218546	0.752587	0.028873

To complement the compound sentiment score analysis, table 3 provides a detailed breakdown of the positive, neutral, and negative sentiment proportions by star rating. Positive Sentiment: Increases steadily from 0.0738 at 1 star to 0.2185 at 5 stars. This trend reflects that higher ratings are associated with a higher proportion of positively expressed words. Neutral Sentiment: Shows a mild decline from 0.8196 at 1 star to

0.7526 at 5 stars. Despite remaining the dominant tone across all reviews, neutral sentiment becomes less prominent in higher-rated reviews as positive expressions increase. Negative Sentiment: Decreases significantly from 0.1066 at 1 star to 0.0289 at 5 stars, indicating that negative word usage is more prevalent in low-star reviews and nearly absent in high-star ones. These findings suggest a coherent alignment between the numerical rating and the underlying emotional content of customer reviews, reinforcing the reliability of sentiment analysis as a complementary method to traditional rating-based evaluations.

3 Experimental Results and Discussion

3.1 Dataset Overview

The original dataset comprised 568,454 review entries, which was truncated to 500 samples for the purpose of analysis. The distribution of review scores showed a significant skew toward five-star ratings, indicating that the majority of customers provided highly positive feedback.

3.2 Sentiment Evaluation

Sentiment analysis was conducted using the VADER (Valence Aware Dictionary and sEntiment Reasoner) tool. The model effectively captured nuanced sentiment expressions in both individual reviews and the overall dataset. For example, the review "This oatmeal is not good…” yielded a compound score of -0.5448, clearly indicating a negative sentiment. This supports VADER’s capability to interpret even subtly expressed opinions [5].

3.3 Aggregated Sentiment Insights

The analysis of average sentiment scores across different star ratings revealed a positive linear trend. Higher review scores consistently corresponded with increased positive sentiment and decreased negative sentiment. This relationship confirms that sentiment polarity, as measured by VADER, aligns closely with user-assigned ratings.

3.4 Visual Interpretation

Graphical representations—specifically bar charts of compound, positive, neutral, and negative sentiment scores—further corroborate the quantitative findings. These visuals illustrate a clear trend: as the review rating increases, positive sentiment intensifies, while negative sentiment diminishes, and neutral sentiment remains relatively stable. The visualizations not only validate the numerical data but also provide an intuitive understanding of sentiment distribution across ratings [1], [10].

4 Conclusion

This study has introduced a comprehensive Python-based pipeline for conducting sentiment analysis on Amazon product reviews, utilizing the VADER sentiment analysis tool from the NLTK library. The pipeline encompasses all key stages—data ingestion, preprocessing, sentiment scoring, and the generation of both aggregate metrics and component-level visualizations—providing a cohesive and scalable framework for understanding consumer sentiment in e-commerce platforms.

The experimental findings clearly demonstrate that higher review scores correlate with stronger positive sentiment, a trend consistently observed in both numerical sentiment metrics and graphical visualizations. The compound sentiment scores progressively increase with higher star ratings, while negative sentiment components decline, indicating a strong alignment between the sentiment model outputs and user perceptions.

Furthermore, the breakdown of sentiment into positive, neutral, and negative components allows for a more granular interpretation of customer feedback. This level of detail is especially valuable for businesses seeking to extract actionable insights from customer reviews.

Looking ahead, future research may explore scaling the pipeline to process larger datasets, integrating machine learning models to enhance sentiment classification accuracy, and extending the framework to support multilingual sentiment analysis or domain-specific sentiment tuning.

Reference

1. Pang, B., & Lee, L. (2008). *Opinion Mining and Sentiment Analysis*. Foundations and Trends in Information Retrieval, 2(1–2), 1–135.
2. Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers.
3. Zhang, L., Wang, S., & Liu, B. (2018). Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4), e1253.
4. Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4), 1093–1113.
5. Hutto, C. J., & Gilbert, E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media (ICWSM-14)*.
6. McKinney, W. (2010). Data Structures for Statistical Computing in Python. *Proceedings of the 9th Python in Science Conference*, 51–56.
7. Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, 9(3), 90–95.
8. Waskom, M. (2021). Seaborn: Statistical Data Visualization. *Journal of Open Source Software*, 6(60), 3021.
9. Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media.
10. Feldman, R. (2013). *Techniques and applications for sentiment analysis*. Communications of the ACM, 56(4), 82–89. https://doi.org/10.1145/2436256.2436274