

**Benchmarking VADER and TextBlob for sentiment analysis : An
evaluation across diverse textual domains in retail product reviews**

Vijay Kumar [#]

*Department of Computer Science
Bharati Vidyapeeth's College of Engineering
Paschim Vihar 110063
New Delhi
India*

Nitu Sehrawat [†]

*Department of Applied Science
Bharati Vidyapeeth's College of Engineering
Paschim Vihar 110063
New Delhi
India*

Anu Yadav [§]

Rachna Narula [‡]

Dishant Kapoor [@]

*Department of Computer Science
Bharati Vidyapeeth's College of Engineering
Paschim Vihar 110063
New Delhi
India*

Anil Kumar ^{*}

*Department of Applied Science
Bharati Vidyapeeth's College of Engineering
Paschim Vihar 110063
New Delhi
India*

[#] E-mail: vjaikr@gmail.com

[†] E-mail: nitu.sehrawat@bharativedyapeeth.edu.in

[§] E-mail: annyr Rao2007@gmail.com

[‡] E-mail: rachuchaudhary04@gmail.com

[@] E-mail: dishantkapoor29@gmail.com

^{*} E-mail: dranilchhikara@gmail.com (Corresponding Author)

Devansh Tyagi [§]

Ruchi Sharma [^]

*Department of Computer Science
Bharati Vidyapeeth's College of Engineering
Paschim Vihar 110063
New Delhi
India*

Abstract

Sentiment analysis holds the potential to become an essential tool for businesses to understand how consumers feel about their products and services on digital platforms. For this study, we focused on two well-known sentiment analysis tools: VADER and TextBlob. Our results highlight the strengths and weaknesses of both tools and offer insights for researchers and practitioners looking to apply sentiment analysis in retail-related applications. We concluded that text analysis tools can be invaluable.

Subject Classification: 68T50.

Keywords: VADER, TextBlob, Sentiment analysis, Benchmarking, Product reviews, Textual domains, Performance evaluation, Sentiment nuances.

1. Introduction

Sentiment analysis is becoming increasingly important for businesses aiming to understand how customers feel about their products and services, particularly on digital platforms. With the sheer volume of customer reviews growing on sites like Amazon and Flipkart, analyzing this feedback is now essential for companies to address customer needs and increase satisfaction.

In this study, we focus on applying sentiment analysis to customer reviews of musical products. These include items such as CDs, DVDs, instructional videos, operatic pieces, and materials related to music theory, all collected from various retail websites. By leveraging natural language processing (NLP), sentiment analysis would enable us to automatically identify and classify sentiments in these reviews, uncovering trends and preferences that would otherwise be hard to detect manually, due to the quantity of data that would need to be searched. [1].

[§] E-mail: devanshtyagi0607@gmail.com

[^] E-mail: ruchi.sharma@bharatividyaapeeth.edu

For our analysis, we used two well-known tools: VADER (Valence Aware Dictionary and sEntiment Reasoner)[2] and TextBlob. VADER, a lexicon-based tool, is particularly effective for shorter, informal texts like social media posts, but it also works well with other types of content. On the other hand, TextBlob, a Python library, relies on a pre-trained classifier to provide sentiment scores, including polarity (positive or negative) and subjectivity. While VADER excels in capturing nuanced tones, TextBlob offers a straightforward approach to sentiment evaluation [3].

This research ultimately enhances our understanding of how sentiment analysis techniques can be applied to retail reviews, offering useful insights for businesses seeking to use customer feedback to improve their products and services [4].

2. Background and Related Work

Sentiment analysis, a key aspect of natural language processing (NLP), focuses on identifying and categorizing emotions within text. This technology helps businesses uncover meaningful insights from large volumes of data, such as online reviews. [5] The techniques used in sentiment analysis range from lexicon-based approaches to more sophisticated machine learning algorithms, each with distinct advantages and challenges.

Lexicon-based methods use predefined dictionaries of words associated with specific sentiment polarities (positive, negative, or neutral). However, these approaches often fall short when it comes to handling subtle language nuances, ambiguity, and context, making them less effective at detecting fine-grained sentiment variations[6]. On the other hand, machine learning methods use algorithms trained on labeled data to automatically learn patterns between text features and sentiment labels. Models like BERT (Bidirectional Encoder Representations from Transformers) leverage deep neural networks to capture contextual information and complex relationships in the text, leading to better performance in sentiment analysis tasks across various fields [7][8].

Two widely used sentiment analysis tools, VADER (Valence Aware Dictionary and sEntiment Reasoner) and TextBlob, illustrate these contrasting approaches. VADER employs a lexicon-based, rule-driven methodology that integrates domain-specific knowledge. It evaluates both individual words and grammatical structures to determine whether a statement is positive, negative, or neutral, along with a quantified

sentiment score. This makes VADER particularly well-suited for analyzing concise and informal texts, such as social media posts.

TextBlob, on the other hand, uses a pre-trained classifier to assess sentiment and provides two outputs: polarity (positive or negative sentiment) and subjectivity (whether the sentiment is based on facts or personal opinions). While the subjectivity score is not directly utilized in this study, it has potential future applications, such as distinguishing between general product quality and individual customer experiences. TextBlob's simplicity and versatility make it a popular choice for such applications [9].

3. Methodology

3.1 *Data Collection and Preprocessing*

The dataset utilized in this study was obtained from the University of California San Diego archives of Amazon reviews. Textual reviews spanning a diverse range of products, including CDs, DVDs, instructional videos, operatic works, and music theory materials, were collected to ensure representation across different musical genres and formats.

3.2 *Sentiment Analysis Techniques*

Each review was preprocessed and then analyzed using both VADER and TextBlob, generating sentiment scores for comparison [10].

3.3 *Evaluation Metrics*

To assess the performance of VADER and TextBlob, we used the following metrics:

Accuracy: The percentage of reviews where the sentiment (positive, negative, or neutral) was correctly classified, relative to the total number of reviews.

Precision: The percentage of correctly classified positive (or negative) sentiments out of all reviews that were classified as positive (or negative).

These metrics were graphed out for both VADER and TextBlob, allowing us to compare their effectiveness visually by seeing how the star ratings correlate to the average outputs (for positive, negative or overall) for both of our models. [11].

3.4 *Comparison with Alternative Approaches*

When considering alternative approaches to VADER in sentiment analysis, machine learning-based models and rule-based systems emerge as notable contenders.

A common point can be found by having semi-customized models. These machine learning models can leverage data-driven learning patterns in a way that humans may not be able to comprehend. These models can be very accurate but require an incredible amount of computational power and substantial training to be of any use.

Rule-based systems are simple in comparison offers balanced middle ground, combining decent accuracy with user-friendly implementation and low barrier to entry, computation wise. This makes them our preferred choice in this paper. [12] Analyzing the spreading patterns of innovations like ChatGPT across digital platforms, and comparing these dynamics with other significant events, is crucial for evaluating the social media platforms' impact in our society. [13]

3.5 *Limitations*

Potential limitations of the methodology include biases inherent in the dataset collection process and challenges associated with accurately interpreting sentiments expressed in textual data. [14] The reviews may be biased for any reason, maybe a previous bad experience with the brand might lead people to rate the product more poorly than they generally would.

4. **Results and Discussion**

4.1 *Visualizing VADER Scores*

Through the utilization of bar plots, we delineated the distribution of compound sentiment scores across different overall review ratings as shown in Fig 1.[15] [16]

4.2 *Visualizing TextBlob Scores*

The bar graph illustrating TextBlob's compound sentiment scores, which represent the mean of the positive and negative sentiment scores, reveals a clear trend across different rating categories.[17] This trend aligns with the expected sentiment polarities associated with different star ratings (in Fig 2).

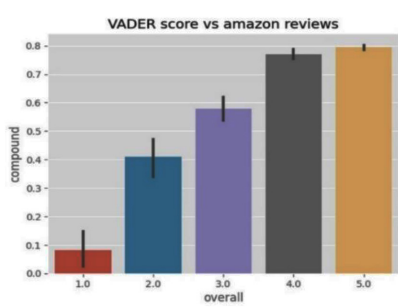


Figure 1

Vader scores benchmarked

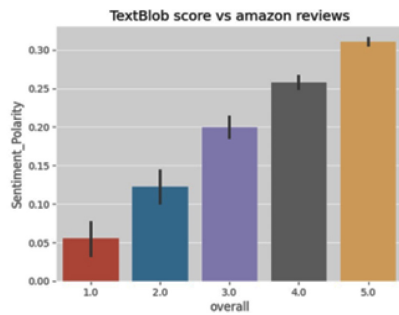


Figure 2

TextBlob scores benchmarked

4.3 Exploring Sentiment Trends

The breakdown of sentiment into positive, neutral, and negative categories provides a granular understanding of consumer sentiment dynamics, as shown in Fig 3, 4 and 5. [18]

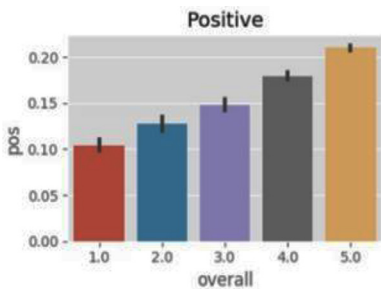


Figure 3

Positive scores vs. overall ratings

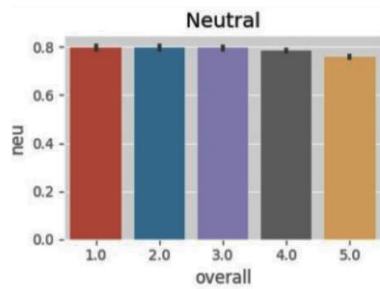


Figure 4

Neutral scores vs. overall ratings

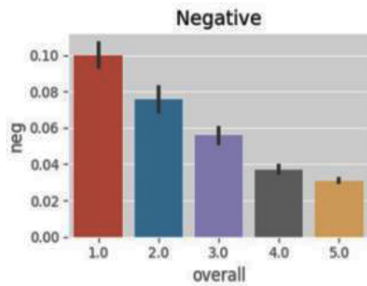


Figure 5

Negative scores vs. overall ratings

5. Discussion

Compound Reviews vs. Star Rating: The scatter graph in Fig. 6 illustrating compound reviews vs. star rating highlights distinct patterns in sentiment distribution. Notably, a clear positive trend is observed for 5-star reviews, where compound sentiment scores predominantly cluster in the positive range. [19]

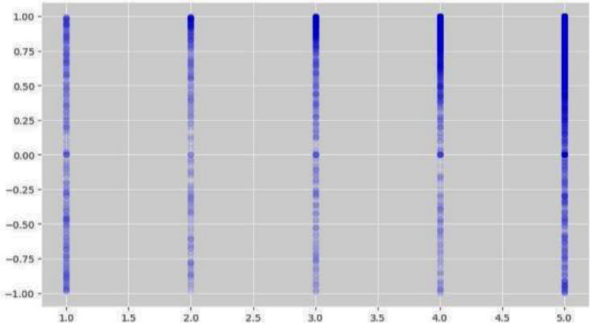


Figure 6
Compound scores

Positive Sentiment Only: In contrast, the scatter graph focusing solely on positive sentiment (Fig 7) reveals a differentiation between 5-star and 1-star reviews. Interestingly, the distinction between 1-star and 2-star reviews appears minimal, however, a significant shift in positive sentiment is evident between 3-star and 4-star reviews.

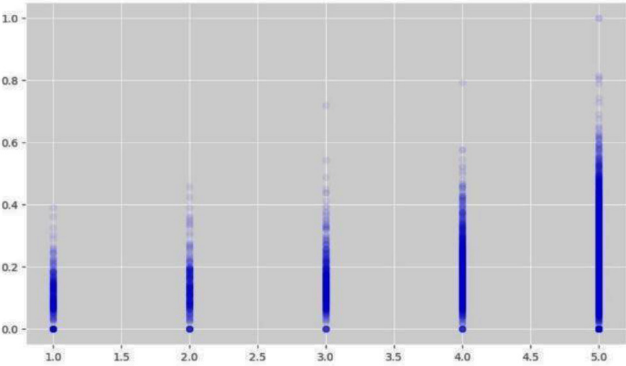


Figure 7
Positive sentiment polarity

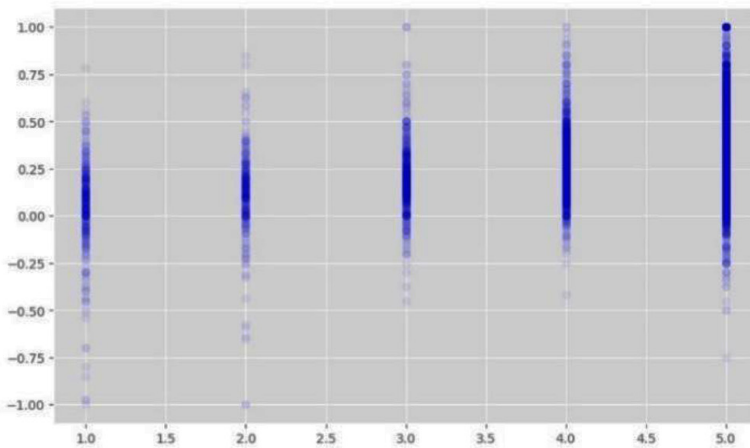


Figure 8
TextBlob polarity

TextBlob Analysis Polarity Analysis: The scatter graph depicting TextBlob's polarity scores against star ratings exhibits a distinct positive trend. [20] This differentiation between rating categories appears more pronounced compared to VADER's compound scores, indicating that TextBlob may be more effective in capturing sentiment nuances at the extreme ends of the rating spectrum (Fig 8).

Comparative Analysis: The comparative analysis of VADER and TextBlob reveals both similarities and differences in their performance. While both techniques exhibit similar positive trends in sentiment scores across star ratings, TextBlob appears to offer a more pronounced differentiation between rating categories, particularly for 5-star and 1-star reviews. This differentiation suggests that TextBlob may be more effective in capturing sentiment nuances at the extreme ends of the rating spectrum.

Limitations

While our scatter graph analysis offers valuable insights, it's important to recognize the limitations of automated sentiment analysis tools like VADER and TextBlob. These tools may struggle to accurately capture context-specific nuances, which can result in misinterpretations of sentiment. Moreover, since sentiment expression is inherently subjective, it introduces complexities and challenges in interpretation.

6. Conclusion

In conclusion, our benchmarking analysis of VADER and TextBlob offers a thorough understanding of their strengths and weaknesses when it comes to analyzing retail product reviews. Our findings highlight the importance of carefully assessing the suitability of sentiment analysis techniques based on the characteristics of the textual data, the level of interpretability required, and the balance between simplicity and accuracy.

References

- [1] P. Törnberg, "How to use Large Language Models for Text Analysis," arXiv:2307.13106v1 (2023).
- [2] C. P. W. Roma and E. B. P. Luque, "Comparative Analysis of Libraries for the Sentimental Analysis," arXiv:2307.14311v1 (2023).
- [3] I. Okpala, A. Johnson, B. Liu, and C. Kim, "A semantic approach to negation detection and word disambiguation with natural language processing," arXiv:2302.02291v3 (2023).
- [4] J.C. Young, M. Lee, H. Parker, and L. Thomas, "CIDER: Context-sensitive polarity measurement for short-form text," arXiv:2307.07864v2 (2023).
- [5] J. Juroš, T. Kovač, and R. Šojat, "LLMs for targeted sentiment in news headlines: Exploring different levels of prompt prescriptiveness," arXiv:2403.00418v1 (2024).
- [6] Y. Chen and M. Si, "Enhancing Sentiment Analysis Results through Outlier Detection Optimization," arXiv:2311.16185v1 (2023).
- [7] A. Buscemi and D. Proverbio, "ChatGPT vs Gemini vs LLaMA on Multilingual Sentiment Analysis," arXiv:2402.01715v1 (2024).
- [8] A. Kuila, S. Ghosh, P. Verma, and R. Mukherjee, "Analyzing sentiment polarity reduction in news presentation through contextual perturbation and large language models," arXiv:2402.02145v1 (2023).
- [9] A. K. M. Masum, M. S. Rahman, and N. Islam, "Hybrid quantum-classical machine learning for sentiment analysis," arXiv:2311.14282 (2023).
- [10] A. K. T. K. Ajay Krishnan, S. M. Arun, and R. V. Reddy, "ClimateNLP: Analyzing public sentiment towards climate change using NLP," arXiv:2111.12569 (2023).

- [11] M. Nkongolo, "Enhancing Search Engine Precision through Sentiment- based Polysemy Resolution," arXiv:2307.13106v1.
- [12] L. E. M. Äyräväinen, T. J. Koskinen, P. M. Rantala, and M. S. Vainio, "Sentiment analysis in digital spaces: An overview of reviews," (2023).
- [13] S. Alipour, A. Rezaei, M. K. R. Shokri, and H. Mohammadi, "Cross-platform social dynamics: ChatGPT vs. COVID-19 vaccine discussions," arXiv:2212.02168 (2023).
- [14] T. Zhang, J. Liu, Z. Wang, and H. Chen, "Revisiting sentiment analysis for software engineering in the era of large language models," arXiv:1911.12569 (2023).
- [15] Y. Tang, J. Li, and S. Zhang, "FinEntity: A dataset for entity-level sentiment classification in financial texts," arXiv:2309.16234 (2023).
- [16] M. Hämäläinen, A. Heikkinen, P. Järvinen, and L. Mattila, "Video games as a corpus: Sentiment analysis using Fallout New Vegas dialog," arXiv:2212.02168 (2023).
- [17] V. Capraro, S. Johnson, A. G. Smith, and M. T. Jones, "Language-based game theory in the age of artificial intelligence," J. R. Soc. Interface, vol. 21, no. 20230720 (2024).
- [18] N. Jain, "Customer Sentiment Analysis using Weak Supervision for Customer-Agent Chat," arXiv:2111.14282 (2023).
- [19] R. Chen, L. Zhang, Y. Li, and K. Huang, "Video-based cross-modal auxiliary network for multimodal sentiment analysis," arXiv:2208.13954 (2023).
- [20] A. Kumar, M. Singh, R. Verma, and S. Shukla, "Emotion helps sentiment: A multi-task model for sentiment and emotion analysis," arXiv:1911.12569 (2023).

Received August, 2024