Enhanced Customer Insight Through Sentiment Analysis of Online Reviews

***Abstract—* Sentiment analysis plays an important role in extracting meaningful insights from customer reviews, aiding businesses in enhancing their services and products. This study introduces a hybrid approach that combines the lexicon-based VADER method with machine learning techniques, specifically logistic regression, to improve the accuracy and efficiency of sentiment analysis. Utilizing a dataset of Amazon Alexa reviews, the proposed model leverages both sentiment scores and text features, achieving a accuracy of 93%. The results show the potential of incorporating analytical methods to refine sentiment assessments*.***

# Introduction

In today's digital world, online customer reviews are crucial to shaping business strategies and consumer perceptions. These reviews provide essential insights into customer satisfaction and product performance but analyzing them manually is challenging due to their volume and complexity.

Sentiment analysis, a specialized area within natural language processing (NLP), enables the automated detection of emotional nuances in text.. Lexicon-based methods, known for their simplicity, are limited in their ability to handle complex expressions and contextual nuances. Conversely, machine learning approaches offer more accuracy through adaptable models but require significant computational resources and data for training.

This research proposes a hybrid approach that combines the strengths of both lexicon-based techniques and machine learning. By integrating the Sentiment Intensity Analyzer (VADER) with a logistic regression model enhanced by TF-IDF vectorization, this study aims to improve the accuracy and efficiency of sentiment analysis. Using Amazon Alexa product reviews as a dataset, this paper will explore the methodology, implementations, and compare its performance against traditional methods, demonstrating its potential to provide faster, more accurate customer insights.

# Methodology

The study utilizes a dataset consisting of customer reviews for Amazon Alexa products, sourced directly from Kaggle. Each review includes textual feedback and a binary feedback indicator, which categorizes the review as positive or negative. Initial preprocessing involved cleaning the text by removing non-alphanumeric characters and converting all text to lowercase. Stopwords were eliminated to focus on meaningful words, and the text was lemmatized to reduce words to their base forms, using the NLTK library [2]. This normalization ensures that the analysis focuses on the essential content of the reviews.

The Sentiment Intensity Analyzer (VADER) from the NLTK library was first applied to obtain sentiment scores for each review, then provided a sentiment classification based on the threshold [1]. To enhance the model, the reviews were also vectorized using the TF-IDF technique, which transforms text into a numerical values that emphasize the most unique words in each review.

A hybrid model was developed by combining the sentiment scores with TF-IDF features. The dataset was split into training (80%) and testing (20%) data using the train\_test\_split function to ensure a balanced distribution of positive and negative feedback in both sets. Logistic regression was chosen for its efficacy in binary classification tasks. The model was trained on the training set, where both the sentiment scores and TF-IDF vectorized text served as input features.

The performance of the logistic regression model was evaluated based on its accuracy in predicting the sentiment of the reviews in the test set. Comparisons were made with the results from standalone VADER and machine learning model to assess the value of the hybrid approach in terms of accuracy and efficiency.

# Results and discussion

Upon evaluation, the hybrid model achieved an accuracy of 93%, surpassing the standalone VADER model at 91.2% and the TF-IDF only model at 91.6%. This improvement demonstrates that the hybrid approach benefits from both the VADER and the text analysis provided by TF-IDF vectorization. Such integration effectively captures nuanced sentiments that may be missed by using either method alone.

The enhanced performance can be attributed to the model’s focus on the sentiment intensity of words and their significance within the text context. This comprehensive analysis offers deeper insights into customer sentiments, crucial for businesses seeking actionable feedback. Additionally, the use of logistic regression, which efficiently handles both numerical scores and textual data, highlights the practical applicability of this approach in real-world applications.

This section highlights the value of merging different analytical techniques to boost the accuracy of sentiment analysis models, showing potential advancements for enhancing business intelligence.

# future work

The encouraging outcomes of this study open up multiple promising paths for future research. First, to further enhance the performance of sentiment analysis models, experimenting with different machine learning algorithms such as Support Vector Machines (SVMs) or neural networks could be investigated. These models may offer improved capabilities for handling complex patterns in sentiment data , especially when combined with advanced NLP features.

Another potential area of research is extending the analysis to datasets in different languages. This would test the adaptability and scalability of the hybrid model across different cultural contexts, which is crucial for multinational applications. It would also allow the model to be refined and generalized, making it versatile against diverse datasets.

Real-time analysis of customer feedback represents a further extension of this work. Developing methods for streaming data analysis, where the model could offer real-time sentiment assessments, would be beneficial for businesses seeking immediate feedback from online platforms.

Finally, investigating strategies for deploying the hybrid sentiment analysis model within existing business intelligence systems could deliver actionable insights into decision-making processes. This exploration would require addressing technical aspects such as scalability, system integration, and user interface design to confirm the model’s practical usefulness in a business setting.

These future works not only aim to refine the methodologies established in this study but also seek to broaden the impact of sentiment analysis tools in practical applications, contributing to more responsive business strategies.

# Conclusion

This research presented a hybrid approach to sentiment analysis by integrating lexicon-based methods with machine learning techniques, specifically utilizing NLTK's VADER and logistic regression enhanced by TF-IDF vectorization. The study leveraged a dataset of Amazon Alexa product reviews, preprocessed to optimize the sentiment analysis. The results demonstrated that the hybrid model outperformed the standalone methods, achieving an accuracy of 93%. This highlights the effectiveness of combining analytical techniques to better interpret customer sentiments.

The findings of this study highlight the potential of hybrid sentiment analysis models to provide more nuanced and accurate insights into customer feedback, which is critical for businesses looking to improve their product offerings and customer service. The successful application of the model in this study suggests that similar approaches can be applied to other data-rich fields where understanding sentiment and opinion is crucial.

Looking ahead, the paths outlined for future work promise to not only enhance the capabilities of sentiment analysis models but also expand their applicability across different languages and real-time scenarios. This could ultimately lead to more responsive business intelligence tools, enabling companies to make well informed decisions efficiently. Incorporating advanced sentiment analysis methods into business practices represents a significant advancement in utilizing data to enhance customer relationships and drive business success.

# References

[1] S. Vinodini, "Analyzing Sentiments in Paulo Coelho’s Literary Works Using VADER Sentiment Analysis," 2023 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES), Chennai, India, 2023, pp. 1-6, doi: 10.1109/ICSES60034.2023.10465319.

[2] J. Praveen Gujjar, H. R. Prasanna Kumar and M. S. Guru Prasad, "Advanced NLP Framework for Text Processing," *2023 6th International Conference on Information Systems and Computer Networks (ISCON)*, Mathura, India, 2023, pp. 1-3, doi: 10.1109/ISCON57294.2023.10112058.