

Comparative Analysis of Sentiment Prediction Methods Using Aspects in Consumer Reviews

Harish Palakurthi, Oklahoma State University

ABSTRACT

The proliferation of online reviews has made consumer sentiment a critical data point for businesses of all sizes. However, discrepancies between overall star ratings and the textual content of reviews can complicate sentiment analysis tasks. This study investigates the effectiveness of incorporating aspect identification within reviews to improve sentiment prediction accuracy. We compare four distinct approaches: traditional machine learning models, advanced OpenAI large language models (LLMs), SAS Model Builder, and the Sentiment Intensity Analyzer (VADER). The evaluation employs a benchmark dataset of consumer reviews paired with corresponding star ratings and the sentiment extracted from the ratings. The sentiment of consumer reviews was analyzed using the four methods. The results demonstrate that OpenAI LLMs achieve superior sentiment prediction accuracy (88%) followed by aspect-based sentiment analysis using traditional machine learning models (82.5%), SAS Model Builder (75%), and the simple Sentiment Intensity Analyzer-VADER (66%). These findings highlight the value of aspect-based sentiment analysis for bridging the gap between ratings and reviews, keeping the computational requirements of methods used in mind. The aspects provide insights into pain points of consumers and strengths of products and services of businesses. The research offers valuable insights into the strengths and limitations of each method, informing potential strategies to mitigate discrepancies in sentiment perception.

INTRODUCTION

The digital landscape thrives on consumer reviews, acting as a powerful indicator of sentiment and influencing purchasing decisions across various industries. However, discrepancies between overall star ratings and the detailed content of reviews can complicate sentiment analysis tasks for businesses, particularly in the competitive realm of online retail. This study delves into the effectiveness of aspect-based sentiment analysis for improving the accuracy of sentiment prediction in consumer reviews.

This research focuses on the men's clothing segment, a sector heavily reliant on customer reviews for brand image and service evaluation. Platforms like Trustpilot provide a rich source of unstructured textual data in the form of reviews. The challenge lies in extracting meaningful insights from this data to understand customer satisfaction and preferences.

This study addresses this gap by employing a comparative analysis of sentiment prediction methods. We investigate the efficacy of aspect-based sentiment analysis traditional machine learning models, advanced OpenAI Large Language Models (LLMs), SAS Model Builder, and the simple Sentiment Intensity Analyzer-VADER in predicting sentiment based on aspects within reviews. Curated a dataset of 400 reviews from a pool of 6000 reviews on Trustpilot focusing on men's clothing. By incorporating aspect identification, the study aims to achieve a more nuanced understanding of consumer sentiment compared to simple traditional sentiment analysis approaches.

This research offers valuable insights into the strengths and limitations of each method. We evaluate the performance of each model based on its ability to accurately predict sentiment compared to the sentiment inferred from star ratings. The findings will shed light on potential strategies to bridge the gap between ratings and reviews, ultimately improving businesses' ability to glean actionable insights from customer feedback.

RELATED WORK

Farhadloo et al. conducted a seminal study on modeling customer satisfaction from unstructured data using a Bayesian approach. Their work effectively analyzed free-form text reviews from platforms like TripAdvisor, focusing on extracting and quantifying the impact of individual aspects on overall customer satisfaction.

This Bayesian framework provides a valuable foundation for this research. We aim to expand the scope of data analysis in customer satisfaction by incorporating a broader array of machine learning techniques.

Building upon this foundation, this study leverages advancements in machine learning. This study employs models like Logistic Regression, Random Forest, XGBoost, advanced OpenAI Large Language Models (LLMs), SAS Model Builder, and the simple Sentiment Intensity Analyzer-VADER. These methods offer enhanced capabilities for handling the larger datasets typical of online retail feedback, particularly within the men's clothing sector. By integrating natural language processing (NLP) to identify key aspects from textual reviews, this study extends the Bayesian analysis of customer sentiment.

This comparison study offers nuanced insights into how various service elements influence consumer perceptions and satisfaction levels and the strengths and weaknesses of each of the methods employed. This research acknowledges the significant contributions of Farhadloo et al. This comparative analysis provides a comprehensive understanding of both continuity and innovation in the field of customer satisfaction research. By bridging the gap between probabilistic models and machine learning, this study offers a robust framework for businesses to leverage data analysis and enhance customer experiences.

METHODOLOGY

This section outlines the data processing, text preprocessing, sentiment analysis, and modeling techniques employed in our research. Each of the methods used will be detailed here along with the steps involved.

DATA PROCESSING

The online reviews will generally contain information like 'DateofPurchase', 'DateofExperience' etc. To focus on the content of customer feedback, removal of this unnecessary information from each review is carried out as the first step. This ensured our analysis centered solely on the textual feedback provided.

SENTIMENT FROM USER RATINGS (BASE RATINGS FOR COMPARISON)

Reviews with ratings exceeding 3 were labeled positive, below 3 were negative, and exactly 3 were neutral.

LLM BASED SENTIMENT ANALYSIS

Large language models are trained on billions of parameters with humungous amounts of data. They can complete a variety of tasks like Translation, creative writing, summarization, problem solving, coding etc., This study uses the ChatGPT 3.5 turbo API to do the sentiment analysis. Each review is submitted to the LLM through a prompt to establish the sentiment. The prompt is designed to contain 5 examples of sentiment analysis i.e., review along with the associated sentiment. This engineered prompt helps LLM do the sentiment analysis of the review submitted.

The LLM achieved an accuracy of 88% in sentiment prediction when compared to sentiment based on rating. The detailed results of the sentiment analysis by LLM will be discussed in the results section.

SAS MODEL BUILDER (TEXT ANALYTICS)

The SAS Viya Model Builder can carry out Data Mining and Machine Learning, Forecasting, Text Analytics all under one umbrella feature. It has prebuilt templates for the routine use cases. For this study, Text Analytics Type is used with Text Analytics: Generate Concepts, Topics, and Categories template. The data in the form of csv is loaded into the memory and run the prebuilt pipeline with all the requisite nodes (Data, Concepts, Text Parsing, Sentiment, Topics, Categories). The results from the pipeline are accessible from each node.

The SAS Builder achieved an accuracy of 75% in sentiment prediction when compared to sentiment based on rating. The detailed results of the sentiment analysis by SAS Model Builder will be discussed in the results section.

SENTIMENT INTENSITY ANALYZER - VADER

Sentiment Intensity Analyzer (VADER) is a tool for sentiment analysis, specifically designed for social media texts. VADER provides a straightforward and efficient method for determining the sentiment polarity and

intensity of textual data, making it a valuable resource for quick and reliable sentiment scoring. To carry out the Sentiment Intensity Analyzer on reviews we need to preprocess the reviews.

Preprocessing Reviews

In this step, each review is first split into sentences using the ``sent_tokenize`` function from the NLTK library. Each sentence is then tokenized into words using the ``word_tokenize`` function.

Stop words are removed from the list of words, and each word is stemmed to its root form. Finally, the stemmed words are joined back together to reconstruct the reviews.

Sentiment Analysis

In this step, on each preprocessed review the Sentiment Intensity Analyzers `polarity_scores` function is applied to obtain the sentiment scores. The compound score from `polarity_scores` will be used to assess the sentiment of the review.

The Sentiment Intensity Analyzer (VADER) achieved an accuracy of 66% in sentiment prediction when compared to sentiment prediction when compared to sentiment based on rating. The detailed results of the sentiment analysis by the Sentiment Intensity Analyzer (VADER) will be discussed in the results section.

TRADITIONAL MACHINE LEARNING – ASPECT-BASED

Noun Extraction

In our study, we employed a structured approach to analyze consumer reviews and identify key aspects and their associated sentiments. We began by extracting nouns from each review, as nouns typically represent the main topics and aspects discussed. This was done using the `'extract_nouns'` function, which tokenizes each sentence using the NLTK library, tags parts of speech, and filters out nouns. Preprocessing steps included lowercasing, punctuation removal, and stop word removal to ensure clean input for noun extraction.

Vectorization

To represent the extracted nouns quantitatively, we utilized both `'CountVectorizer'` and `'TfidfVectorizer'`. The `'CountVectorizer'` provided term frequency counts, while the `'TfidfVectorizer'` calculated term frequency-inverse document frequency (TF-IDF) weights, highlighting the importance of terms relative to the entire dataset. By multiplying the count vectors by the TF-IDF weights, we obtained a weighted representation that emphasized significant terms.

Clustering-Aspect Identification

We then applied the KMeans clustering algorithm to group the reviews into clusters, each representing a distinct aspect. The number of clusters was set to four, determined using the elbow method to find an optimal balance between intra-cluster variance and the number of clusters. The cluster centers, representing the most significant terms within each cluster, were identified, and aspects were assigned to each review based on their proximity to these cluster centers using cosine distance metrics.

Sentiment-Aspect Mapping

For sentiment analysis, utilized Sentiment from user ratings (Base ratings for comparison) to label each review with a sentiment score (positive, neutral, or negative). A dictionary was used to store sentiment counts for each aspect. By iterating through the reviews and their associated sentiments, we assigned sentiment counts to aspects containing the extracted nouns. This provided a detailed sentiment breakdown for each aspect, allowing us to calculate overall ratings based on the formula: $\text{overall_rating} = (\text{Positive} - \text{Negative}) / \text{Total Sentiments}$.

Aspect-Interpretation

The results were then interpreted to draw conclusions about consumer perceptions. Aspects with higher overall ratings indicated positive consumer sentiment, while lower ratings highlighted areas of concern. This

comprehensive approach enabled us to bridge the gap between star ratings and textual reviews, providing valuable insights into consumer feedback.

Data Preparation

One of the main aims of this study is to predict the sentiment based on the presence of aspects in a review. To achieve this, binary variables are created to indicate the presence of aspects in the reviews. These binary variables will be used as the independent variables in the prediction algorithms. As detecting the sentiment from a review is a classification problem, Logistic Regression, Random Forest and XGBoost were used to predict the sentiment (Positive, Neutral, Negative).

Prediction- Traditional Machine Learning Models

In this step, the data is split into training and testing sets. The models are trained on the training data and the prediction is done using the test data. The best model (Logistic Regression) as per the data/results achieved an accuracy of 82.5% in sentiment prediction when compared to sentiment based on rating. The detailed results of the sentiment analysis by Traditional machine learning – aspect-based will be discussed in the results section.

RESULTS

This section discusses the results obtained from the four sentiment analysis methods. The chosen metric for comparison is accuracy, measured against the sentiment derived from user ratings. Additionally, confusion matrices are provided to offer a detailed breakdown of each method's performance.

Accuracy is the proportion of correctly classified instances among the total instances. It is calculated as:

$$\text{Accuracy} = \frac{\text{Number of True Predictions}}{\text{Total Number of Predictions}}$$

A confusion matrix provides a more detailed performance evaluation by showing the number of true positives, true negatives, false positives, and false negatives in each of the sentiment class.

LLM BASED SENTIMENT ANALYSIS

Using the ChatGPT 3.5 turbo API, sentiment analysis was conducted on each review through a prompt engineered with 5 examples of sentiment analysis. The results are populated in a confusion matrix/heatmap and are as follows:

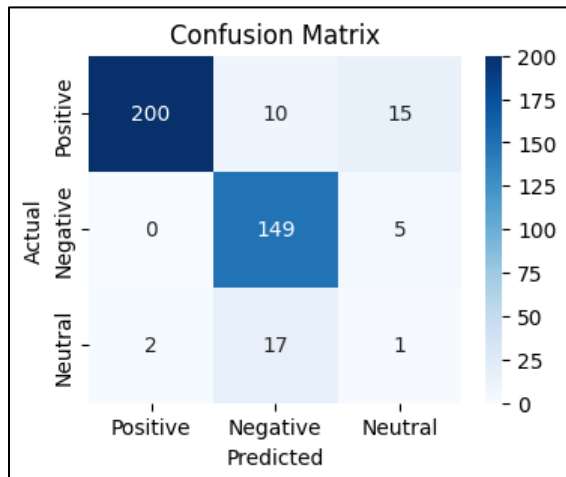


Figure 1. Confusion Matrix for LLM Based Sentiment

Calculation of Accuracy:

= Sum of True Predictions of Positive, Negative and Neutral Sentiments / Total number of predictions made

$$= \frac{200+149+1}{200+10+15+0+149+5+2+17+1} \times 100 = 88\%$$

From the accuracy results it is evident that the LLMs are superior in analyzing the sentiment of reviews. However, the Neutral sentiment prediction is not pleasing. This is due to the mismatch between rating provided by consumer is different from the written review. A few examples will be provided in the next section to demonstrate ambiguity of neutral sentiment and rating.

SAS MODEL BUILDER (TEXT ANALYTICS)

Using the SAS Model Builder, sentiment analysis was conducted using the predefined template provided by SAS. The results are populated in a confusion matrix/heatmap and are as follows:

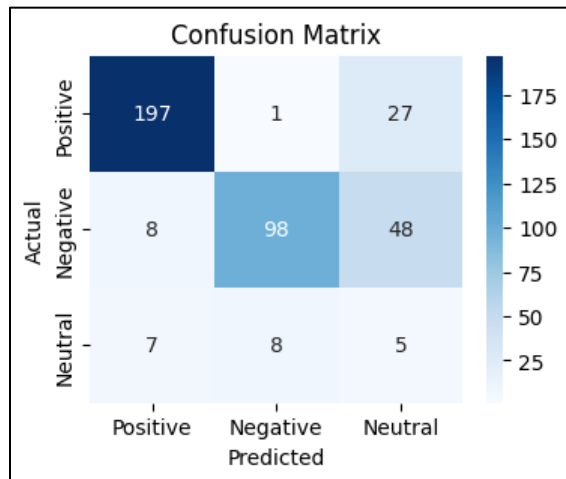


Figure 2. Confusion Matrix for SAS Based Sentiment

Calculation of Accuracy:

= Sum of True Predictions of Positive, Negative and Neutral Sentiments / Total number of predictions made

$$= \frac{197+98+5}{197+1+27+8+98+48+7+8+5} \times 100 = 75\%$$

The SAS Model Builder achieved an accuracy of 75% with predefined templates in analyzing the sentiment of reviews.

SENTIMENT INTENSITY ANALYZER - VADER

Using the traditional Sentiment Intensity Analyzer – Vader based sentiment analysis was conducted on the preprocessed reviews which are sentences of root words. The results are populated in a confusion matrix/heatmap and are as follows:

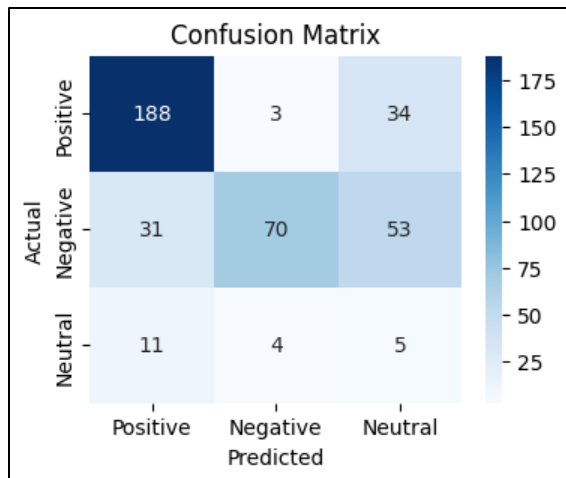


Figure 3. Confusion Matrix for VADER Based Sentiment

Calculation of Accuracy:

= Sum of True Predictions of Positive, Negative and Neutral Sentiments / Total number of predictions made

$$= \frac{188+70+5}{188+3+34+31+70+53+11+4+5} \times 100 = 66\%$$

The Sentiment Intensity Analyzer – Vader achieved an accuracy of 66% with preprocessed reviews in analyzing the sentiment of reviews.

TRADITIONAL MACHINE LEARNING – ASPECT-BASED

The aspect-based sentiment analysis with traditional models requires models to be trained with training data and testing on test data. This study employed 3 classification models namely Logistic regression, Random Forest and XGBoost. The data is divided into training and testing. All the models are trained and tested on the same data. The accuracies of the three models are as follows:

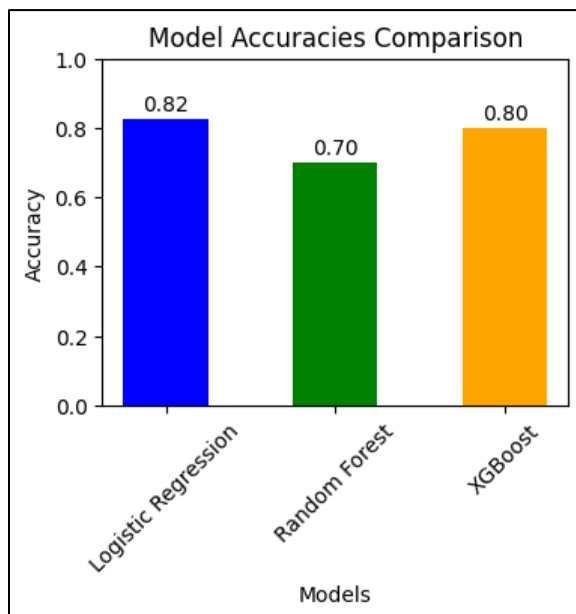


Figure 4. Comparison of Model Accuracies in Sentiment Prediction

From Figure 4 it is evident that Logistic Regression is the best model at predicting sentiment. The results are populated in a confusion matrix/heatmap and are as follows:

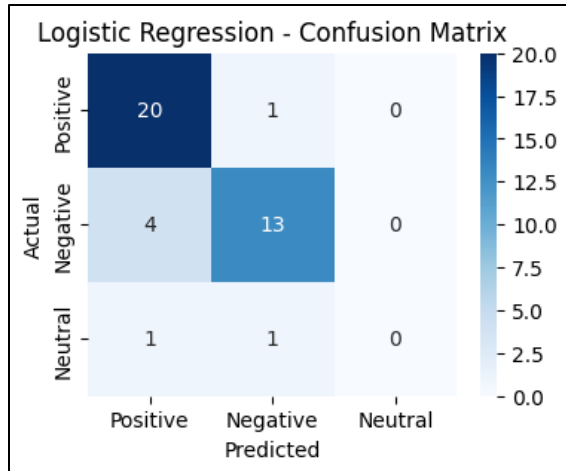


Figure 5. Confusion Matrix for Sentiment Based on Logistic Regression (Aspect Based)

Calculation of Accuracy:

= Sum of True Predictions of Positive, Negative and Neutral Sentiments / Total number of predictions made

$$= \frac{20+13+0}{20+1+0+4+13+0+1+1+0} \times 100 = 82.5\%$$

The Logistic Regression (Aspect Based) achieved an accuracy of 82% with binary variables (indicating presence of aspects in a review) in analyzing the sentiment of reviews.

DISCUSSION

The findings of this study underscore the critical role of aspect identification in sentiment analysis for consumer reviews even though there are some methods that provide better accuracy.

Three out of four methods which were explored in this study resulted in good accuracy in predicting sentiment in line with the star ratings. Out of all the methods LLM based sentiment analysis gave an accuracy of 88%. The reason could be due to its enormous training on very complex data. It can handle most of the intricacies with ease. The SAS Model builder with prebuilt template gave convincing results. However, it lacked in predicting neutral sentiments.

The aspect-based sentiment analysis with traditional machine learning models provided very convincing results. This is due to the multistep process in identifying aspects and classification using the presence of aspects in reviews. This is most effective considering the simplicity, computational complexity and cost when compared to LLM based sentiment analysis. To further discuss about aspects we need to touch upon:

Granularity: Reviews often encompass a range of emotions directed at various aspects of a product or service. Identifying these aspects allows us to move beyond a single sentiment label and understand the specific elements driving customer satisfaction or dissatisfaction.

Discrepancies: As highlighted in the abstract, star ratings and textual content can sometimes contradict each other. By analyzing aspects, we can pinpoint these discrepancies. For example, a customer might award a low star rating due to poor customer service but express overall satisfaction with the product itself.

Actionable Insights: Understanding specific aspects associated with positive or negative sentiment allows businesses to target their improvement efforts more effectively. They can focus on areas with consistently negative sentiment, like shipping times, or capitalize on praised aspects, like product quality.

Focusing solely on an overall rating might miss the complexities within a review. A customer might express frustration with a specific aspect (e.g., slow delivery) but still be happy with the product itself, leading to a seemingly contradictory sentiment.

However, it's important to acknowledge the potential limitations of relying solely on aspect-based sentiment analysis. While valuable, aspects may not capture the entirety of customer sentiment. There might be

underlying emotions or opinions that aren't directly tied to specific features of the product or service. Additionally, the accuracy of aspect identification itself can impact the overall sentiment analysis.

CONCLUSION

In this paper we studied effectiveness of incorporating aspect identification within reviews to improve sentiment prediction accuracy. We compare four distinct approaches: traditional machine learning models, advanced OpenAI large language models (LLMs), SAS Model Builder, and the Sentiment Intensity Analyzer (VADER). The evaluation employs a benchmark dataset of consumer reviews paired with corresponding star ratings and the sentiment extracted from the ratings. The sentiment of consumer reviews was analyzed using the four methods. The four methods resulted in accuracies of 82%, 88%, 75%, 67% respectively. Though the results of LLM based sentiment analysis is highest, the computational complexity and cost for doing the task is high as well. The aspect-based sentiment analysis and SAS model builder are not only providing a very good accuracy, but the analysis is also simple and can be done at less cost. The study helps in understanding importance of aspects in identifying the sentiment as well features of products or services which needs to be improved or maintained.

REFERENCES

1. M. Farhadloo, E. Rolland, (2013) Multi-class sentiment analysis with clustering and score representation, IEEE 13th International Conference on Data Mining, Workshops (ICDMW), IEEE, 2013. pp. 904–912.
2. M. Farhadloo, M., Patterson, R. A., & Rolland, E. (2016). Modeling customer satisfaction from unstructured data using a Bayesian approach. Decision Support Systems Volume 90 October 2016, Pages 1-11

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Harish Palakurthi
hpalaku@okstate.edu