



A Survey on Extracting Customer Sentiment from Product Reviews by Leveraging Natural Language Processing

Pedada Meenakshi*

Department of Information Technology, GMR Institute of Technology, Rajam-532127, Andhra Pradesh, India

ABSTRACT:

In the age of digital transformation, customers increasingly rely on online product reviews to share their opinions and experiences. These reviews contain valuable insights that, when properly analysed, can help businesses improve their products and services. This paper focuses on extracting customer sentiment from product reviews by leveraging Natural Language Processing (NLP) techniques and classification methods. The proposed approach includes essential text preprocessing steps such as tokenization, stop-word removal, and stemming to clean and standardize the data. Feature extraction methods like TF-IDF are used to convert the textual data into numerical format, followed by the application of ML algorithms such as Support Vector Machines (SVM) and XGBoost to classify sentiments as positive, negative, or neutral. Additionally, a visualization module is implemented to display summarized insights from the reviews based on product attributes. This approach can help businesses interpret customer opinions and improve decision-making.

Keywords: Customer Sentiment, Opinion Mining, Natural Language Processing, Product Reviews, Feature Extraction, Text Classification.

Introduction

In the digital marketplace, customer reviews have become an essential component in shaping purchasing behaviour, brand perception, and product development. These reviews, often filled with unstructured and diverse user opinions, hold valuable insights into customer satisfaction, expectations, and sentiment. However, manually analyzing large volumes of such data is not only time-consuming but also inefficient. This paper aims to address this challenge by extracting customer sentiment from product reviews through the use of advanced Natural Language Processing (NLP) techniques and sentiment classification methods. The process begins with data preprocessing techniques such as tokenization, stop-word removal, and stemming, followed by feature extraction using TF-IDF to transform textual content into structured numerical data. Sentiment classification is then performed using effective algorithms like Support Vector Machines (SVM) and XGBoost, allowing for the categorization of feedback into positive, negative, or neutral sentiment. By leveraging these tools, the study seeks to uncover actionable insights from customer opinions, enabling businesses to improve product quality, refine marketing strategies, enhance customer satisfaction, and make informed, data-driven decisions in a competitive online environment.

Literature Survey

Prihananto et al. (2024) analyzed smartphone brand positioning in Indonesia using LIWC, PCA, and clustering, achieving 92.7% data variation, suggesting integrating user reviews and machine learning for improvement. The price forecasting was improved by considering external factors like metrological data along with historical data of Cabbage and Radish. Lee et al. (2024) developed a method using text mining, Word2vec, and LDA to analyze online reviews, creating Customer Journey Maps, enabling time-efficient insights for product development, suggesting expanded data sources. Mbeledogu and Ogbu (2024) designed a sentiment analyzer using NLP, Gaussian Naïve Bayes, and CountVectorizer. The results were 90% accurate, indicating better models, real-time analysis, and diversified datasets are the way forward.

Malathi (2024) suggested CAT-CUK for the extraction of features and NB-SVM as a classifier, which achieved improved accuracy in the opinion mining process, thus recommending future research in hybrid algorithms in conjunction with deep learning.

Arunachalam (2024) used data mining and NLP to do sentiment analysis of social media data and found consumer preference. Recommendations include analyzing multichannel platforms and instant sentiment analysis.

Singgalen (2024) analyzed social toxicity, sentiments, and network dynamics. He applied SVM for sentiment classification. Achieved 97.94% accuracy. Future work includes cross-industry and platform-specific analyses.

Alrayani et al. (2024) utilized sentiment analysis and K-means clustering to analyze Arabic reviews, identifying user satisfaction topics in super apps. Future improvements include hybrid algorithms and deep learning integration.

Koufakou (2024) applied deep learning and NLP to analyze student sentiments and classify course review topics. RoBERTa achieved 95.5% sentiment accuracy, with SVM at 79.8% for topic classification. Future enhancements include balancing datasets and optimizing models. Hussain et al. (2023) developed a product recommender system using NLP and aspect-based sentiment analysis, ranking products based on user specifications and reviews. The system was tested using the Amazon dataset, but this relies on review quality and can be complex to maintain.

Liu et al. (2023) improved the Kansei engineering method by employing data mining, TF-EPA, and BP neural networks for capturing emotional feedback from online product reviews. BP neural network outperformed multiple linear regression. The approach is highly computationally intensive and can be further enhanced by including more product categories and complex algorithms.

Cortis and Davis (2021) performed a systematic review of social opinion mining with a focus on sentiment polarity, emotion, and sarcasm in social media. They applied NLP, machine learning, and deep learning approaches. RoBERTa was able to reach 95.5% accuracy for sentiment. Some challenges involve dealing with unstructured and multilingual data.

Future work would be on enriching multilingual datasets and aspect-based opinion mining. Moazzam et al. (2021) considered the opinion mining of customers based on buying behavior using classification techniques through machine learning. Techniques like KNN, Naïve Bayes, SVM, BOW, and N-Gram were applied. With bi-grams, the accuracy was obtained as 82.32%. Challenges are noisy data and scope is limited. Future work should focus on enhancing the reliability by incorporating informal language and semantic features.

Recently, Yi and Liu (2020) proposed a recommendation system based on MSVM and NLP for analysis of customer reviews. This resulted in a nearly 98% accuracy and MAPE of 96%. It has the following challenges: It demands high computation and heavily depends on data quality. Future developments should be on real-time data processing and advanced NLP techniques. Tanwar and Rai (2020) proposed a system that uses SVM, NLP, and LDA for detecting spam reviews and spammers. Naïve Bayes achieved 75% accuracy, while SVM reached 65.7%. The system efficiently identifies fake reviews but requires significant computational resources. Future work should emphasize real-time spam detection and optimizing feature extraction.

Sánchez-Núñez et al. (2020) performed a bibliometric analysis on sentiment analysis, opinion mining, and emotion understanding in advertising using VOSviewer and SciMAT. The study provides insights into trends but is limited to focusing only on WoS-indexed publications. Future research should incorporate AI and machine learning and expand to other databases like Scopus and Google Scholar.

Methodology:

The methodology initiates with data collection and exploration, followed by preprocessing the text, including steps like lemmatization and stemming. Feature representation is created using the Bag of Words method, and the resulting data is input into an XGBoost model for classification. The most effective model is selected according to performance metrics, and the outcomes are thoroughly evaluated. Finally, a detailed report is compiled, summarizing the findings and emphasizing the model's accuracy, precision, and capability in sentiment classification for opinion mining tasks.

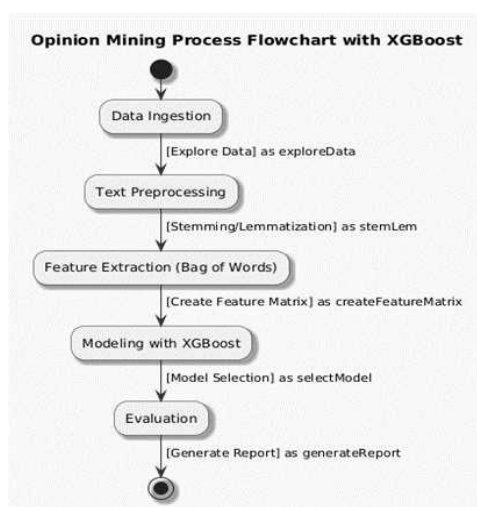


Fig 3 Flowchart for Opinion Mining using XGBoost

1.1 NLP Algorithms:

In Natural Language Processing tasks, multiple models and algorithms are used to process and analyze text data, and text preprocessing plays a crucial role in preparing the text for analysis. Tokenization divides the text into individual words or tokens, forming the foundation for further steps. Stop-word removal eliminates frequently occurring words like "and," "the," and "is," which carry little analytical value. Stemming and lemmatization reduce words to their root forms, ensuring variations such as "running" and "run" are treated consistently. Additionally, normalization standardizes the text by converting it to lowercase, removing punctuation, and correcting spelling mistakes. Feature extraction is a key step that converts raw text into structured data ready for modeling.

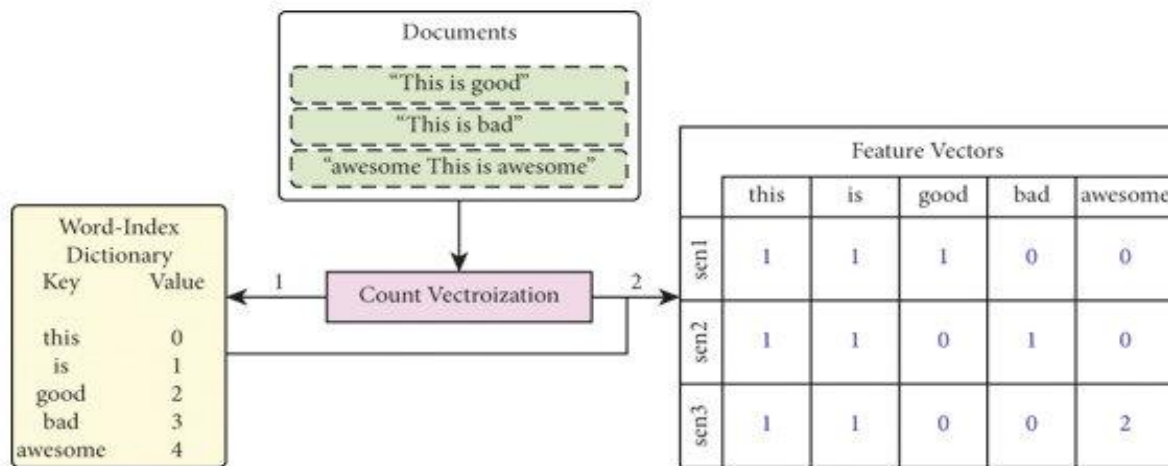


Fig 3.2 Tokenization-BoW Model

The most widely used method is Count Vectorizer, which transforms text into a feature matrix by counting how frequently each word occurs across the dataset, often restricting features to the top 4000 most relevant words. Another popular technique is TF-IDF (Term Frequency-Inverse Document Frequency), which assigns weights to words based on their frequency in a single document compared to their occurrence across all documents. This method highlights the most significant terms in the text, improving the focus on important information.

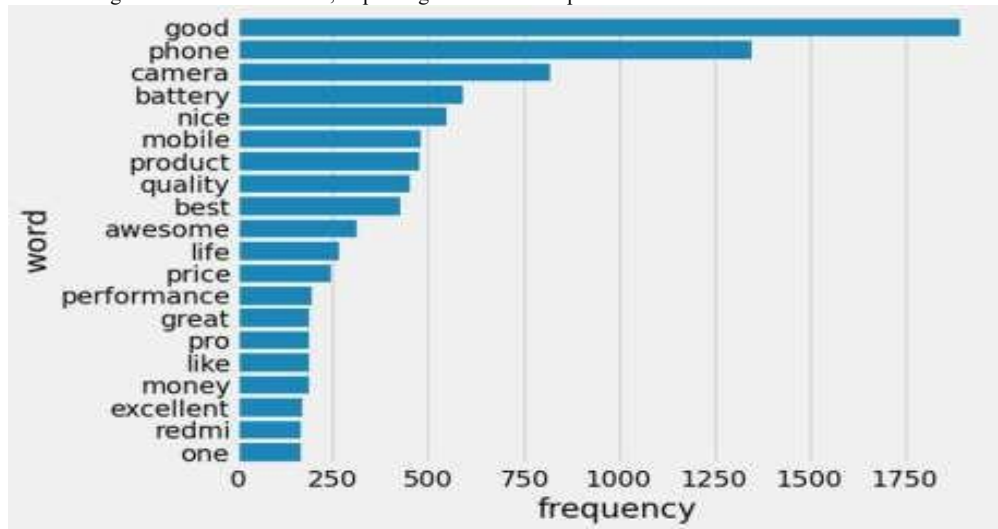


Fig 3.3. Frequency count of word

Polarity Classification Algorithm: In Sentiment Analysis, this algorithm categorizes text on a scale from "Strongly Negative" to "Strongly Positive," including "Neutral" in between. Polarity scores help in identifying the sentiment behind customer reviews, making classification easier. The *Replacement Algorithm* expands contractions such as "can't" to "cannot," standardizing the text for consistent analysis. Visualization can be done using *WordCloud*, which displays the most frequent words, with size indicating frequency, making key patterns and trends clear. By applying these NLP techniques, the system effectively analyzes customer reviews, extracts important insights, and performs sentiment classification. This approach provides a better understanding of customer opinions, which is crucial for improving business decision-making.

Fig 3.4 Word Cloud for Review Text



The Bag of Words (BoW) model is a fundamental NLP technique that converts text documents into numerical vectors. It splits text into individual words, builds a vocabulary of unique terms, and assigns an index to each word. Each document is then represented as a vector, with elements indicating the frequency or presence of words. A limitation of BoW is that it ignores word order and focuses only on word counts, which can result in loss of semantic meaning. It is also sensitive to stop words and can produce very high-dimensional vectors. Despite these issues, BoW is widely used in text classification, document clustering, information retrieval, and topic modeling. While advanced methods like TF-IDF and word embeddings capture richer semantics for complex NLP tasks, BoW remains valuable for simple text analysis due to its ease of use and efficiency.

3.2 Machine learning models:

XGBoost (Extreme Gradient Boosting) is a highly efficient and scalable machine learning algorithm widely used for both classification and regression applications. It is an enhanced version of gradient boosting that constructs a series of decision trees one after another, where each new tree tries to correct the mistakes made by the previous trees. This sequential learning process improves the overall prediction accuracy by combining outputs from multiple weaker models into a stronger ensemble. XGBoost is particularly effective for large datasets and can handle binary classification, multiclass classification, and regression tasks. Key characteristics include L1 and L2 regularization to avoid overfitting, automatic handling of missing values, parallel computing for faster model training on multi-core systems, and distributed processing for scaling across extensive datasets. Additionally, tree pruning is supported to manage model complexity and prevent unnecessary overfitting. The algorithm can efficiently process sparse and incomplete

data, which makes it practical for real-world datasets. Techniques for handling imbalanced data using custom loss functions are also built in. Cross-validation and hyperparameter tuning allow further optimization, ensuring high accuracy and robust performance. XGBoost has found applications in areas such as fraud detection, customer segmentation, and predictive maintenance. Its flexibility, speed, and predictive strength make it a popular tool in machine learning competitions and practical business use cases.

Classifying online review comments with XGBoost involves several steps to convert raw text into meaningful insights, making it highly effective for sentiment analysis. Initially, text data is transformed into structured features, often using techniques such as TF-IDF (Term Frequency-Inverse Document Frequency) or word embeddings. TF-IDF measures the importance of a word in a document relative to its frequency across the dataset, highlighting significant terms in the text. Word embeddings provide dense vector representations of words that capture semantic relationships, giving a richer understanding of the text context. These processed features are then supplied as input to the XGBoost model, which sequentially constructs decision trees. Each tree learns to correct the errors of the preceding trees, progressively improving the model's accuracy. Through this iterative learning, the algorithm identifies patterns in review comments and maps them to sentiment labels such as positive, negative, or neutral. Once trained, the model can classify unseen comments accurately. XGBoost is capable of handling large volumes of data efficiently and is robust against missing or noisy inputs, which are common in real-world datasets. It also effectively addresses imbalanced sentiment distributions, where positive, negative, and neutral reviews may not occur equally. This is managed by adjusting class weights or applying techniques that emphasize underrepresented classes. By combining scalability, robustness, and accuracy, XGBoost offers a highly reliable approach for sentiment classification and extracting actionable insights from customer feedback.

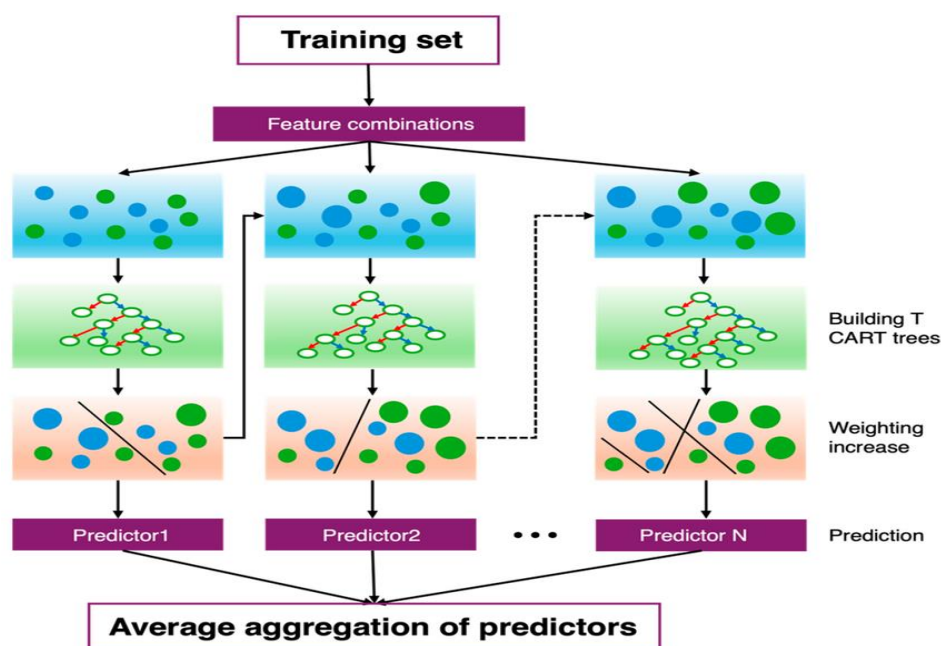


Fig 3.5: Semantic Illustration for XGBoost Model

4 Discussions:

The proposed study focuses on understanding how customer reviews can be classified as positive or negative using Natural Language Processing (NLP) techniques and the XGBoost algorithm, a popular method in machine learning. The model was trained on a large set of preprocessed customer reviews, where important text features were extracted to represent customer opinions in a meaningful way. The expected outcome of such a system is to achieve high accuracy in identifying customer sentiments, which can help businesses understand their customers better and manage their brand reputation more effectively. Future research may explore the use of advanced NLP models, deeper contextual analysis, and real-time sentiment tracking to make the system more accurate and informative.

During the study, several machine learning algorithms were analyzed and compared based on their performance. Among them, XGBoost showed the best results in terms of accuracy, precision, recall, F1-score, and AUC values. Its ensemble learning approach, which combines many weak decision trees into a strong predictive model, allows it to capture complex patterns in the data and produce reliable results. Although other algorithms such as CatBoost and Random Forest also performed well, XGBoost consistently gave better results, showing its strength and flexibility in handling complex and large datasets.

5.Results:

The application of the XGBoost algorithm in sentiment analysis shows strong performance in classifying customer reviews. The model achieved about 90% accuracy, proving its effectiveness in identifying sentiments correctly. Its AUC score of 98% reflects excellent differentiation between positive, negative, and neutral sentiments. Precision, recall, and F1 scores—88%, 87%, and 80% respectively—indicate consistent reliability across sentiment types.

The system conceptually involves preprocessing customer reviews to clean and standardize the text, followed by feature extraction to capture key information. These features are then used by XGBoost for sentiment classification. Although no actual implementation or interface was used, the approach highlights XGBoost's ability to deliver accurate sentiment predictions and valuable insights.

Future improvements could involve advanced NLP methods, deeper contextual understanding, and real-time analysis. Using contextual embeddings or entity recognition could help detect nuances like sarcasm or mixed sentiments, making results more precise. Overall, the study shows that XGBoost is a powerful tool for sentiment analysis, offering meaningful insights even with large text datasets.

Conclusion

This study shows that the model effectively performs sentiment analysis using XGBoost to classify customer opinions on smartphone brands. By categorizing sentiments from "Strongly Negative" to "Strongly Positive," it provides a clear understanding of market perception and customer preferences. The model helps brands improve marketing strategies and product development based on real feedback. Monitoring sentiment trends supports informed decision-making and enhances customer satisfaction. Analyzing opinions across multiple brands also offers insights for strategic planning and innovation. Overall, sentiment analysis allows businesses to better meet consumer needs and promotes growth in the competitive smartphone industry, while opening avenues for future research on market behavior and customer engagement.

REFERENCES

- [1] Prihananto, P., Yusvianty, R. S. A., Hakim, N. S., Bhawika, G. W., & Agustin, H. (2024). Analyzing Brand Positioning and Brand Image of Smartphone Brands in Indonesia by Mining Online Review. *Procedia Computer Science*, 234, 318-332.
- [2] Lee, Y. U., Chung, S. H., & Park, J. Y. (2024). Online Review Analysis from a Customer Behavior Observation Perspective for Product Development. *Sustainability*, 16(9), 3550.
- [3] Mbeledogu, N. N., & Ogbu, I. M. (2024). Guardians of E-Commerce: Harnessing NLP and Machine Learning Approaches for Analyzing Product Sentiments in Online Business in Nigeria. *International Journal of Advanced Engineering, Management and Science*, 10(3), 33-41.
- [4] Malathi, M. OPINION MINING ON CUSTOMER REVIEWS THROUGH CAT-CUK BASED FEATURE SELECTION AND NB-SVM BASED CLASSIFICATION (2024), *Journal of Nonlinear Analysis and Optimization*
- [5] Arunachalam, M. P. (2024). Sentiment Analysis of Social Media Data for Product and Brand Evaluation: A Data Mining Approach Unveiling Consumer Preferences, Trends, and Insights. *IJEMR*, 14(3), 46-52.
- [6] Singgalen, Y. A. (2024). Digital marketing of smartphone manufacturing product: toxicity, social network, and sentiment classification. *International Journal on Social Science, Economics and Art*, 14(1), 73-86.
- [7] Alrayani, B., Kalkatawi, M., Abulkhair, M., & Abukhodair, F. (2024). From Customer's Voice to Decision-Maker Insights: Textual Analysis Framework for Arabic Reviews of Saudi Arabia's Super App. *Applied Sciences*, 14(16), 6952.
- [8] Sun, C.; Pei, M.; Cao, B.; Chang, S.; Si, H. "A Study on Agricultural Commodity Price Prediction Model Based on Secondary Decomposition and Long Short-Term Memory Network". *Agriculture* 2024, 14, 60. <https://doi.org/10.3390/agriculture14010060>
- [9] Koufakou, A. (2024). Deep learning for opinion mining and topic classification of course reviews. *Education and Information Technologies*, 29(3), 2973-2997
- [10] Liu, Z., Wu, J., Chen, Q., & Hu, T. (2023). An improved Kansei engineering method based on the mining of online product reviews. *Alexandria Engineering Journal*, 65, 797-808.

-
- [11] Cortis, K., & Davis, B. (2021). Over a decade of social opinion mining: a systematic review. *Artificial intelligence review*, 54(7), 4873-4965. Xiong, Tao et al. "A combination method for interval forecasting of agricultural commodity futures prices." *Knowl. Based Syst.* 77 (2015): 92-102.
- [12] Moazzam, A., Mushtaq, H., Sarwar, A., Idrees, A., Tabassum, S., & Rehman, K. U. (2021). Customer opinion mining by comments classification using machine learning. *International Journal of Advanced Computer Science and Applications*, 12(5).
- [13] Yi, S., & Liu, X. (2020). Machine learning based customer sentiment analysis for recommending shoppers, shops based on customers' review. *Complex & Intelligent Systems*, 6(3), 621-634
- [14] Sánchez-Núñez, P., Cobo, M. J., De Las Heras-Pedrosa, C., Pelaez, J. I., & Herrera-Viedma, E. (2020). Opinion mining, sentiment analysis and emotion understanding in advertising: a bibliometric analysis. *IEEE Access*, 8
- [15] Tanwar, P., & Rai, P. (2020). A proposed system for opinion mining using machine learning, NLP and classifiers. *IAES International Journal of Artificial Intelligence*, 9(4), 726.