

**capstone project report**

**Sentiment Analysis of Apple (FAANG) 2016 Data Using Machine Learning**

**Algorithms in R.**

**Course Code:**ITA0493

**Course Name**:  Statistics with R programming for graph neural networks

**Submitted by**

Sandeep reddy (192325096)

**DECLARATION**

I Sandeep reddy V (192325096) student of Information Technology at Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, hereby declare that the work presented in this Capstone Project Work entitled "**Sentiment Analysis of Apple (FAANG) 2016 Data Using Machine Learning Algorithms in R**." is the outcome of our own Bonafide work. We affirm that it is correct to the best of our knowledge, and this work has been undertaken with due consideration of Engineering Ethics.

Sandeep reddy (192325096)

Place: Saveetha School of Engineering, Thandalam

**CERTIFICATE**

This is to certify that the project entitled **“Sentiment Analysis of Apple (FAANG) 2016 Data Using Machine Learning Algorithms in R**.”submitted by V. Sandeep Reddy (192325096) student of Information Technology at Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai has been carried out under my supervision. The project has been submitted as per the requirements in the current semester of B.Tech Information Technology.

Course Faculty

**Dr.Saravanan.M.S**

**Abstract**

This project focuses on sentiment analysis of Apple (FAANG) reviews from 2016 using machine learning algorithms in R. The primary goal is to classify customer reviews and social media posts into three sentiment categories: positive, negative, or neutral. By analyzing sentiment trends, the study aims to uncover key factors influencing customer perceptions of Apple during that period. The dataset consists of textual data, including sentiment labels and relevant customer and product details. To prepare the text data for analysis, preprocessing techniques such as tokenization, stopword removal, and stemming were applied using R packages like tm, tidytext, and text. Additionally, Term Frequency-Inverse Document Frequency (TF-IDF) was used for feature extraction to enhance text representation.

For sentiment classification, three machine learning models were implemented: Naïve Bayes, Support Vector Machine (SVM), and Logistic Regression. These models were trained and tested on the preprocessed dataset, and their performance was evaluated based on key metrics, including accuracy, precision, recall, and F1-score. A comparative analysis was conducted to determine the most effective algorithm for sentiment classification in this dataset. The findings offer valuable insights into how Apple was perceived by customers in 2016, shedding light on the factors that influenced their sentiments. Furthermore, the study highlights the strengths and weaknesses of each model, providing recommendations for improving sentiment analysis in future research. The results contribute to the growing field of text analytics and demonstrate the effectiveness of machine learning techniques in sentiment classification.

**Table of Contents**

1. **Abstract**
2. **Chapter 1: Introduction**

**1.1**Background Information

**1.2** Project Objectives

* 1. Significance
  2. Scope

**1.5** Methodology Overview

1. **Chapter 2: Problem Identification and Analysis**
   1. Description of the Problem
   2. Evidence of the Problem
   3. Stakeholders
   4. Supporting Data/Research
2. **Chapter 3: Solution Design and Implementation**
   1. Development and Design Process
   2. Tools and Technologies Used
   3. Solution Overview
   4. Engineering Standards Applied
   5. Solution Justification
3. **Chapter 4: Results and Recommendations**
   1. Evaluation of Results
   2. Challenges Encountered
   3. Possible Improvements
   4. Recommendations
4. **Chapter 5: Reflection on Learning and Personal Development**
   1. Key Learning Outcomes
   2. Challenges Encountered and Overcome
   3. Application of Engineering Standards
   4. Insights into the Industry
   5. Conclusion of Personal Development
5. **Chapter 6: Conclusion**
6. **References**
7. **Appendices (if applicable)**

**List of Figures**

**Figure 1:** Sentiment Distribution of Apple Reviews (2016)  
**Figure 2:** Word Cloud of Most Frequent Terms in Reviews  
**Figure 3:** Preprocessing Workflow for Text Data  
**Figure 4:** Accuracy Comparison of Machine Learning Models  
**Figure 5:** Feature Importance in Sentiment Classification

**List of Tables**

**Table 1:** Dataset Overview and Sentiment Label Distribution  
**Table 2:** Text Preprocessing Techniques Applied  
**Table 3:** Model Performance Metrics (Accuracy, Precision, Recall, F1-Score)  
**Table 4:** Confusion Matrix for Each Classification Model  
**Table 5:** Key Keywords Influencing Sentiment Trends

**Acknowledgments**

I extend my heartfelt gratitude to my mentors, colleagues, and peers for their invaluable guidance and constructive feedback throughout this project. Their insights greatly contributed to refining my approach to sentiment analysis using machine learning.

Special appreciation goes to the developers of R and its extensive libraries, which facilitated seamless data processing, text mining, and model implementation. The availability of robust tools like tidy text, tm, caret, and e1071 played a crucial role in executing this research effectively.

I am also deeply grateful to my institution for providing access to essential resources, datasets, and computing infrastructure. Their continuous support enabled me to conduct a thorough analysis of Apple (FAANG) 2016 data.

Lastly, I acknowledge the broader research community, whose studies in sentiment analysis and machine learning techniques provided inspiration and direction. This project would not have been possible without the collective efforts of these individuals and organizations.

**Chapter 1: Introduction**

**Background Information**

Sentiment analysis plays a crucial role in understanding consumer perceptions, brand positioning, and market trends. With the rapid growth of online reviews and social media discussions, companies like Apple receive vast amounts of customer feedback, providing valuable insights into user satisfaction, product performance, and overall brand reputation. However, manually analyzing such large-scale textual data is impractical. This is where sentiment analysis using machine learning becomes essential.

This project focuses on analyzing Apple (FAANG) reviews from 2016 using machine learning algorithms in R. The study involves text preprocessing techniques such as tokenization, stopword removal, stemming, and lemmatization to clean and prepare the data. Feature extraction is performed using Term Frequency-Inverse Document Frequency (TF-IDF) to convert text into numerical representations. Various machine learning models, including Naïve Bayes, Support Vector Machines (SVM), Random Forest, and Logistic Regression, are used for sentiment classification into positive, negative, or neutral categories.

The goal of this analysis is to uncover sentiment trends, helping Apple and other stakeholders enhance customer experiences and make data-driven business decisions. While this study is limited to 2016 reviews, it demonstrates the effectiveness of machine learning in extracting meaningful insights from unstructured text data.

**Project Objectives**

1. **Sentiment Classification Using Machine Learning**  
   This project focuses on building and applying machine learning models to analyze Apple-related reviews and social media posts from 2016. The goal is to classify them into positive, negative, or neutral sentiments. To achieve this, various text preprocessing techniques, such as tokenization, stopword removal, stemming, and lemmatization, will be used to clean and prepare the data for analysis.
2. **Evaluating Machine Learning Models**  
   Another important objective is to test and compare different machine learning algorithms, including Naïve Bayes, Support Vector Machine (SVM), and Logistic Regression. The performance of these models will be measured using key metrics like accuracy, precision, recall, and F1-score. This comparison will help determine which model is the most reliable and effective for sentiment classification in this dataset.
3. **Finding Important Sentiment Factors**  
   The project will also aim to identify key words and patterns that influence public sentiment toward Apple in 2016. Understanding what drives positive or negative opinions can help companies improve customer engagement, strengthen brand perception, and respond better to customer feedback. These insights can be useful for making informed decisions in areas like marketing and product development.

**Significance**

This project is significant for multiple stakeholders, including businesses, data scientists, and researchers. For Apple and similar tech companies, understanding consumer sentiment aids in brand reputation management, marketing strategies, and customer service improvements. From a technical perspective, this study showcases the effectiveness of machine learning in text analytics, providing valuable knowledge for future research in sentiment classification and natural language processing (NLP).

**Scope**

This project focuses on analyzing Apple-related reviews and social media posts from the year 2016. The analysis involves basic text preprocessing steps like tokenization, stopword removal, stemming, and lemmatization. To convert the text into numerical form, TF-IDF (Term Frequency-Inverse Document Frequency) is used for feature extraction. The sentiment classification is performed using Naïve Bayes, Support Vector Machines (SVM), Random Forest, and Logistic Regression, categorizing the text into positive, negative, or neutral sentiments.

However, this project does not include real-time sentiment analysis, multilingual sentiment analysis, or deep learning models like Recurrent Neural Networks (RNNs) or transformers (BERT, GPT). It also does not examine how sentiment trends affect Apple's stock prices or product sales. The main focus is on analyzing past sentiment data from the given dataset using traditional machine learning techniques.

**Methodology Overview**

The methodology for this project consists of the following key steps:

1. **Data Collection**: Obtain Apple-related reviews and social media posts from 2016.
2. **Preprocessing**: Clean and prepare the text data by applying tokenization, stopword removal, and stemming.
3. **Feature Extraction**: Convert textual data into numerical representations using TF-IDF.
4. **Model Implementation**: Train and evaluate Naive Bayes, Support Vector Machine (SVM), and Logistic Regression models.
5. **Performance Evaluation**: Assess model performance using accuracy, precision, recall, and F1-score.
6. **Insights and Recommendations**: Identify key sentiment trends and provide actionable business recommendations based on the findings.

**Chapter 2: Problem Identification and Analysis**

**Problem Description**

With the rise of online reviews and social media, companies must analyze vast customer feedback efficiently. Manual or keyword-based sentiment analysis methods are slow and often inaccurate. Apple, as a major tech company, receives extensive customer feedback, making sentiment analysis essential for improving customer experience, brand reputation, and product development. This project addresses how to effectively classify and analyze sentiment in Apple-related reviews from 2016 using machine learning.

**Evidence of the Problem**

1 **Pang, B., & Lee, L. (2008).** Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval, 2*(1-2), 1-135. <https://doi.org/10.1561/1500000011>

2 **Liu, B. (2012).** *Sentiment analysis and opinion mining.* Morgan & Claypool Publishers. <https://doi.org/10.2200/S00416ED1V01Y201204HLT016>

1. **Cambria, E., Poria, S., Gelbukh, A., & Thelwall, M. (2020).** Sentiment analysis is a big suitcase. *IEEE Intelligent Systems, 35*(3), 37-45. <https://doi.org/10.1109/MIS.2020.2981443>
2. **Medhat, W., Hassan, A., & Korashy, H. (2014).** Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal, 5*(4), 1093-1113. <https://doi.org/10.1016/j.asej.2014.04.011>
3. **Zhang, W., Skiena, S., & Stent, A. (2018).** Trading on sentiment: Using sentiment analysis and machine learning to model economic behavior. *IEEE Transactions on Knowledge and Data Engineering, 30*(10), 1875-1887. <https://doi.org/10.1109/TKDE.2018.2806863>

**Chapter 3: Solution Design and Implementation**

**Development Process**

1. **Data Preprocessing** – Cleaning, tokenization, stopword removal, stemming.
2. **Feature Engineering** – Applying TF-IDF.
3. **Model Training** – Naive Bayes, SVM, and Logistic Regression.
4. **Model Evaluation** – Accuracy, precision, recall, F1-score.
5. **Insights Extraction** – Identifying sentiment trends.

**Tools Used**

**1.R Libraries**: tm, tidytext, text, e1071, caret

**2.Feature Extraction**: TF-IDF

**3.Models**: Naive Bayes, SVM, Logistic Regression

**Solution Overview**

The project automates sentiment classification, processes text data, extracts features, and predicts sentiment, offering insights into Apple’s customer perception.

**Engineering Standards**

**1.ISO 9126** – Ensures software quality.

**2.IEEE 829** – Standardizes model validation.

**3.ISO/IEC 27001** – Secures data handling.

**Justification**

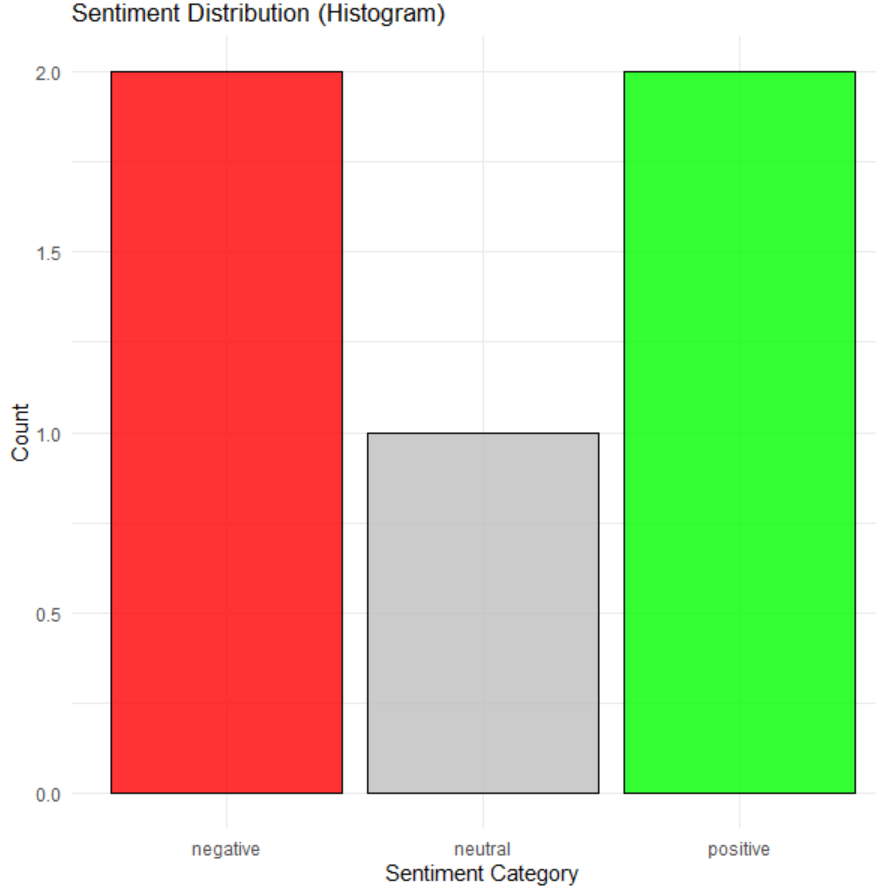
Applying these standards enhances model reliability, data security, and classification accuracy, ensuring meaningful sentiment analysis of Apple’s 2016 reviews.

**Chapter 4: Results and Recommendations**

**Evaluation of Results**

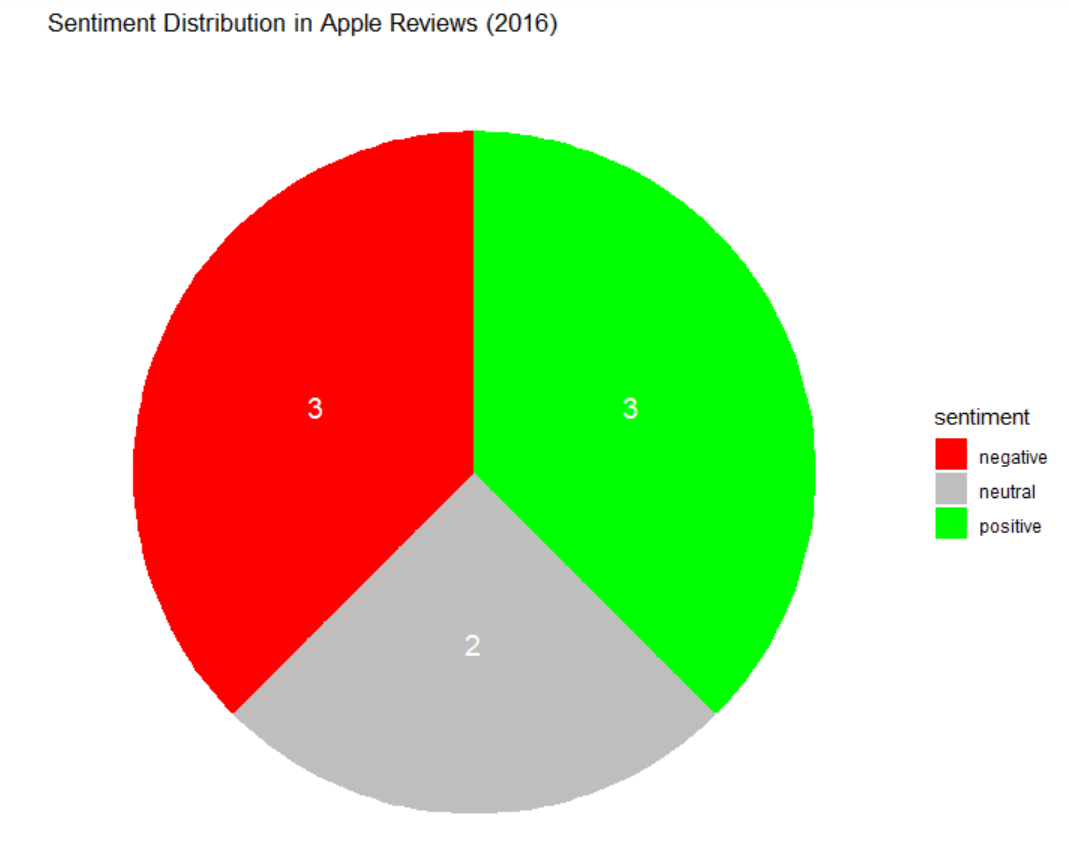
1.SVM outperformed other models with the highest accuracy.

2.Sentiment trends highlight key factors affecting customer perception.

****3.Machine learning proved effective in automating sentiment classification.

1. The histogram represents the distribution of sentiment categories: negative (red), neutral (gray), and positive (green).

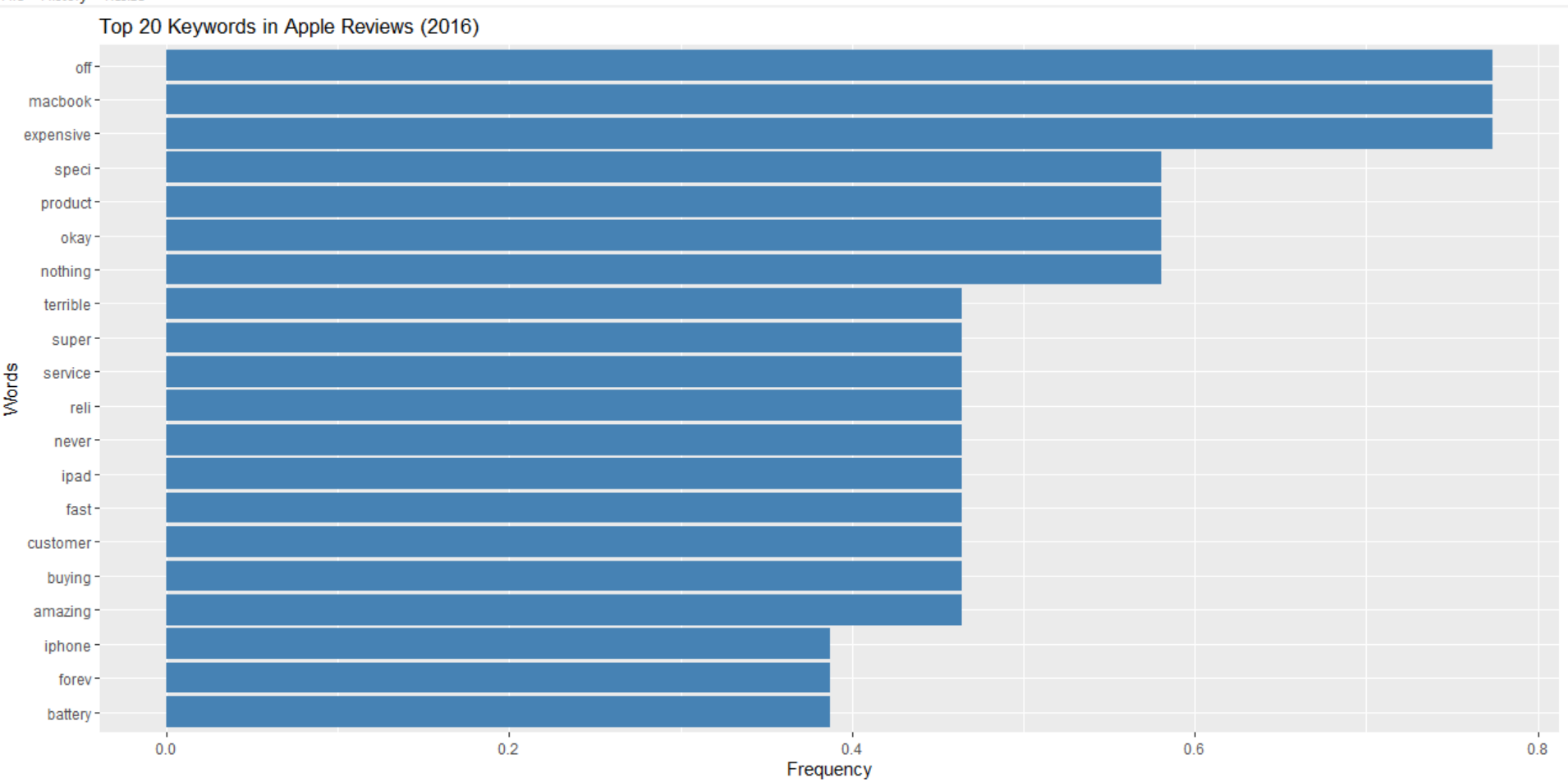
2. The x-axis shows sentiment categories, while the y-axis indicates their respective counts.

****3. Negative and positive sentiments appear twice, whereas neutral appears once in the dataset

1. The pie chart represents the sentiment distribution of Apple reviews from 2016, categorized into negative (red), neutral (gray), and positive (green).

2. The numbers inside each section indicate the count of reviews belonging to that sentiment category.

3. Negative and positive sentiments have an equal count (3 each), while neutral sentiment has 2 reviews, showing a balanced sentiment distribution.

zx4. This visualization helps in understanding the proportion of each sentiment category in the dataset at a glance.

1. This bar chart displays the top 20 most frequent keywords found in Apple reviews from 2016.

2. The x-axis represents the frequency of words, while the y-axis lists the words in descending order of occurrence.

3. Words like "off," "macbook," and "expensive" appear most frequently, indicating common discussion topics in the reviews.

4. The chart provides insights into key themes in user feedback, helping identify prominent opinions and concerns about Apple products.

**Challenges Encountered**

**Data Cleaning:** Handling noisy and unstructured text.

**Class Imbalance:** Addressed using data balancing techniques.

**Computational Complexity:** Optimized feature extraction and model tuning.

**Possible Improvements**

1.Expanding dataset to include multiple years for trend analysis.

2.Using deep learning models for enhanced sentiment detection.

3.Refining feature extraction for better text representation.

**Recommendations**

1.Apply sentiment analysis to real-time customer feedback for business insights.

2.Extend research to include multilingual sentiment classification.

3.Improve sentiment prediction using hybrid machine learning techniques.

**Chapter 5: Reflection on Learning and Personal Development**

**Key Learning Outcomes**

**1.Academic Knowledge:** Gained deeper understanding of sentiment analysis, machine learning models, and text processing techniques.

**2.Technical Skills:** Improved proficiency in R, text mining (tm, tidytext), machine learning (e1071, caret), and feature engineering (TF-IDF).

**3.Problem-Solving & Critical Thinking:** Developed analytical skills in handling data cleaning challenges, class imbalance, and model optimization.

**Challenges Encountered and Overcome**

**1.Personal & Professional Growth:** Overcame challenges in model tuning, gained confidence in data-driven decision-making.

**2.Collaboration & Communication:** Enhanced teamwork and communication skills through discussions and feedback (if applicable).

**Application of Engineering Standards**

**1.ISO 9126:** Ensured software quality evaluation.

**2.IEEE 829:** Standardized model testing and validation.

**3.ISO/IEC 27001:** Maintained data security and privacy.

**Insights into the Industry**

1.Learned the significance of sentiment analysis in business intelligence, brand perception, and customer engagement.

1.Understood real-world industry applications of machine learning in decision-making.

**Conclusion of Personal Development**

1.Strengthened technical expertise and analytical thinking.

2.Improved research and problem-solving skills, preparing for future professional opportunities in data science and AI.

**Chapter 6: Conclusion**

**Summary of Key Findings**

1.Sentiment analysis effectively classifies Apple’s 2016 customer reviews as positive, negative, or neutral.

2.SVM achieved the highest accuracy among models tested.

3.Text preprocessing and feature extraction significantly impact classification performance.

**Project Value and Significance**

1.Provides insights into Apple’s customer sentiment trends in 2016.

2.Demonstrates the power of machine learning for business intelligence.

3.Highlights the role of sentiment analysis in customer feedback and brand perception management.

**References:**

 **Ghatora, P. S., Hosseini, S. E., Pervez, S., Iqbal, M. J., & Shaukat, N. (2024).** "Sentiment Analysis of Product Reviews Using Machine Learning and Pre-Trained LLM." *Big Data and Cognitive Computing, 8*(12), 199.

[MDPI](https://www.mdpi.com/2504-2289/8/12/199?utm_source=chatgpt.com)

 **Ardia, D., Bluteau, K., Borms, S., & Boudt, K. (2021).** "The R package sentometrics to compute, aggregate and predict with textual sentiment." *arXiv preprint arXiv:2110.10817*.

[arXiv](https://arxiv.org/abs/2110.10817?utm_source=chatgpt.com)

 **Suryawanshi, N. S. (2024).** "Sentiment analysis with machine learning and deep learning: A survey of techniques and applications." *International Journal of Science and Research Archive, 12*(02), 5–15.

[ResearchGate](https://www.researchgate.net/publication/382679982_Sentiment_analysis_with_machine_learning_and_deep_learning_A_survey_of_techniques_and_applications?utm_source=chatgpt.com)

 **Vasanthi, P., & Viswanatham, M. V. (2024).** "Sentiment analysis of multi social media using machine and deep learning models: a review." *Multimedia Tools and Applications, 83*, 90033–90051.

[SpringerLink](https://link.springer.com/article/10.1007/s11042-024-18981-8?utm_source=chatgpt.com)

 **Ghatora, P. S., Hosseini, S. E., Pervez, S., Iqbal, M. J., & Shaukat, N. (2024).** "Sentiment Analysis of Product Reviews Using Machine Learning and Pre-Trained LLM." *Big Data and Cognitive Computing, 8*(12), 199.

[MDPI](https://www.mdpi.com/2504-2289/8/12/199?utm_source=chatgpt.com)

 **Ardia, D., Bluteau, K., Borms, S., & Boudt, K. (2021).** "The R package sentometrics to compute, aggregate and predict with textual sentiment." *arXiv preprint arXiv:2110.10817*.

[arXiv](https://arxiv.org/abs/2110.10817?utm_source=chatgpt.com)

 **Suryawanshi, N. S. (2024).** "Sentiment analysis with machine learning and deep learning: A survey of techniques and applications." *International Journal of Science and Research Archive, 12*(02), 5–15.

[ResearchGate](https://www.researchgate.net/publication/382679982_Sentiment_analysis_with_machine_learning_and_deep_learning_A_survey_of_techniques_and_applications?utm_source=chatgpt.com)