Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

Sentiment Analysis on Customer Reviews: A Comprehensive Case Study

Tirtha Ray, Dr. Shilpi Bose and Dr. Chandra Das
Department of Computer Science and Engineering, Netaji Subhash Engineering College, Kolkata, India

Abstract - The rapid growth of Internet-based platforms such as social media and e-commerce sites has produced vast volumes of text expressing user opinions. Sentiment analysis automatically identifies and classifies such opinions as positive, negative, or neutral, supporting decision-making for businesses, governments, and researchers. This paper presents a systematic case study of sentiment analysis using four heterogeneous review datasets — movie, stock market, employee, and pharmaceutical reviews — to evaluate a unified classical machine-learning pipeline across domains. We propose a pipeline encompassing data preprocessing, exploratory data analysis, feature engineering (Bag-of-Words, TF-IDF, n-grams), feature selection (Chi-squared test), and classification using Logistic Regression, Decision Tree, Random Forest, and AdaBoost. On the IMDB movie-review dataset, Logistic Regression combined with TF-IDF was the best-performing model, achieving an F1-score of 0.889 in baseline evaluation and 90.8% accuracy (F1 = 0.908, ROC-AUC = 0.968) after refinement; the remaining three datasets were used for cross-domain preprocessing, EDA, and feature-analysis validation. Model behaviour was interpreted using Explainable AI techniques (LIME and SHAP), and a probability-thresholding mechanism was introduced to extend the binary classifier into a pseudo-trinary (positive/neutral/negative) predictor without retraining. Finally, an interactive Streamlit-based application (SACR) was developed to make the entire pipeline accessible without coding.

Keywords - Sentiment Analysis; Natural Language Processing; TF-IDF; Logistic Regression; Explainable AI; LIME; SHAP

I. INTRODUCTION

Vast amounts of textual data are generated daily across social media, e-commerce websites, and news outlets. Sentiment analysis, a key component of natural language processing (NLP), enables the automated identification and classification of emotions, opinions, and attitudes expressed in text. Early approaches relied on rule-based and lexicon-based techniques, which struggled with ambiguity and context; the advent of machine learning, and later deep learning models such as RNNs, LSTMs, and Transformers (BERT, GPT), significantly improved accuracy and contextual understanding.

Sentiment analysis has wide-ranging applications: businesses gauge customer satisfaction through reviews, governments analyze public opinion, financial institutions predict market trends, and healthcare providers assess patient feedback. Despite these advances, the field faces persistent challenges — language ambiguity, sarcasm, multi-polarity, and negation frequently mislead simple sentiment models. This paper proposes a sentiment analyzer that is robust to such challenges while remaining interpretable and computationally efficient.

Sentiment analysis can be performed at different levels of granularity: document-level analysis summarizes an entire document's sentiment as positive, negative, or objective; sentence-level analysis classifies individual sentiment-bearing sentences; and phrase-level analysis identifies polarity of specific phrases within a sentence. Real-world applications increasingly demand this finer-grained insight — for instance, product analysis requires discovering which specific attributes or qualities of a product appeal to customers, rather than a single overall label. Motivated by this, our objectives in this study were threefold: (i) map numeric ratings into meaningful sentiment classes, (ii) address real-world challenges such as sarcasm, multi-polarity, and negation, and (iii) develop an NLP/ML pipeline capable of classifying English text reviews as positive, negative, or neutral.

Unlike prior single-domain studies, this work systematically evaluates a unified classical machine-learning pipeline across four heterogeneous review domains — movies, stock-market commentary, employee reviews, and pharmaceutical feedback — to test its generality beyond a single benchmark corpus. Building on this, we introduce a probability-thresholding mechanism that derives a neutral sentiment class from a binary classifier without retraining, and we validate every modeling decision using LIME and SHAP to ensure the resulting system is interpretable rather than a black box. The complete pipeline is further packaged into SACR, a no-code interactive tool, to make it directly usable by non-programmers. These four elements — cross-domain evaluation, threshold-based neutral-class extension, systematic explainability analysis, and no-code deployment — constitute the primary contributions of this paper.

II. LITERATURE REVIEW

Sentiment analysis and opinion mining are often used interchangeably in the literature, though some researchers distinguish opinions as concrete thoughts and sentiments as feelings [7]. Cambria et al. [5] proposed a three-layer structure comprising 15 NLP sub-problems spanning syntactic, semantic, and pragmatic layers — including microtext normalization, word-sense disambiguation, sarcasm detection, and polarity detection — and argued that a holistic approach, rather than simple classification, is necessary for robust sentiment analysis. Complementing this, Cambria et al. [4] categorize sentiment analysis and affective-computing approaches into three broad families: knowledge-based techniques that rely on sentiment lexicons

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

and rule sets, statistical approaches built on machine learning and deep learning, and hybrid techniques that combine both.

Several surveys have traced the evolution of feature representation and classification techniques in this space. Hemmatian and M. K. [8] and Ravi and Ravi [14] both review the tasks, approaches, and applications of opinion mining, tracing the shift from static lexicon-based scoring to adaptive, learning-based classifiers. Subhashini [16] presents a review of contemporary opinion-mining literature, covering how noisy or uncertain text features are extracted and represented. Mowlaei et al. [15] propose adaptive aspect-based lexicons using statistical and genetic-algorithm-based strategies, allowing lexicons to be dynamically updated for more precise aspect-level sentiment grading, while Uma and K. E. [17] similarly build context-aware dynamic lexicons by aggregating multiple dictionary sources.

Domain-specific studies further motivate the multi-dataset approach adopted in this work. Sentiment analysis has been applied to hotel reviews to understand customer likes and dislikes [19], to stock and cryptocurrency markets to predict price movement from market sentiment [6], and to Twitter data across varied domains including COVID-19 discourse [2], [9], [10]. The healthcare domain, in particular, has seen a surge of sentiment-analysis applications for patient-opinion mining, satisfaction analysis, and clinical-feedback review [3], [11], [12], [18]. Collectively, these works confirm that sentiment analysis techniques generalize across very different textual domains, but that each domain introduces its own vocabulary and labeling conventions — which is precisely why this study evaluates a single unified pipeline across four structurally different review datasets (movies, stock market commentary, employee reviews, and pharmaceutical reviews) rather than a single-domain corpus.

III. PROPOSED METHODOLOGY

This study uses four publicly available review datasets to build and evaluate a sentiment analyzer capable of classifying text as positive, negative, or neutral, while addressing real-world challenges such as sarcasm, multi-polarity, and negation. Fig. 1 summarizes the overall pipeline, from data ingestion through preprocessing, feature engineering, model training, and evaluation.

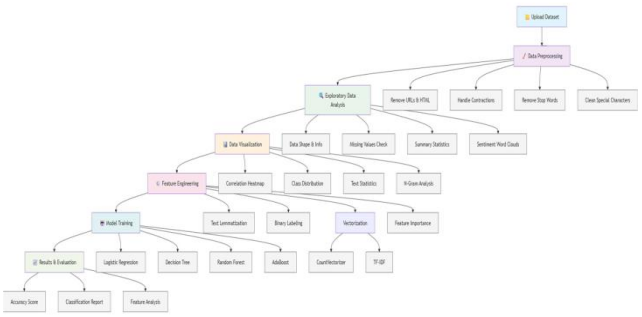

Fig. 1. End-to-end sentiment analysis pipeline.

A. Datasets

Dataset	Recs.	Key Attributes
IMDB Movie Reviews	40,000	Summary, review, movie name; binary rating
Stock Market	5,780	Summary; rating (+1 / −1)
Employee Reviews	67,529	Summary, pros/cons, company, job title; 1–5 rating
Pharmaceutical Reviews	112,197	Drug name, review, rating (1–10), date

Ratings on a 1–10 scale were mapped to three sentiment classes: 1–4 as Negative, 5–6 as Neutral, and 7–10 as Positive, converting the task into a well-defined text classification problem.

B. Data Preprocessing and EDA

Text was lowercased, cleaned of URLs, HTML tags, special characters, and numerals, then tokenized, stripped of stop words using NLTK's predefined list, and lemmatized (e.g., "running" → "run") to minimize vocabulary variation. Records with missing or non-informative attributes were identified and excluded — for instance, the pharmaceutical dataset's "condition" column had 899 missing values and was dropped, while the core review and rating columns were confirmed complete.

Exploratory data analysis examined class imbalance, missing values, and word-frequency patterns across all four datasets using word clouds and n-gram (unigram to 5-gram) analysis. Fig. 2 shows word clouds generated for the employee-review dataset: notably, high-frequency terms such as "good", "great", and "people" appear prominently in both positive and negative reviews, confirming that individual keywords alone are insufficient signals and that positive-sounding words frequently occur within negative reviews and vice versa.

Fig. 2. Word clouds of positive vs. negative employee reviews.

To resolve this ambiguity, n-gram distributions (bigrams and trigrams) were analyzed to capture short contextual phrases rather than isolated words, which proved effective at surfacing sarcastic or context-dependent statements. For example, the review "Good pay and benefits but spending more time sending emails than doing actual work" uses predominantly positive vocabulary, yet its bigram/trigram structure around "but spending more time" reveals an overall negative sentiment — a pattern invisible to unigram-only models. Fig. 3 shows a representative bigram frequency distribution used to identify such patterns.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

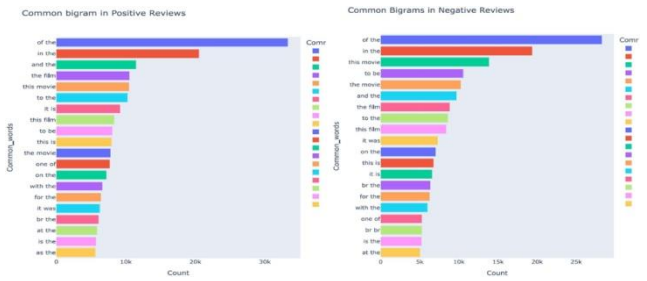

Fig. 3. Bigram frequency distribution used for context-aware analysis.

C. Real-World Challenges

Three recurring challenges were identified during EDA that any practical sentiment analyzer must address:

1) Sarcasm: users sometimes express negative sentiment using positive words (e.g., "My car has an awesome mileage of 4 km per liter"), which can easily fool a naive keyword-based model.

2) Multi-polarity: a single review may contain mixed sentiment across different aspects (e.g., "The screenplay, cuts, and sound deserve a standing ovation, but the story is not good"), where the polarity flips after a contrasting conjunction such as "but".

3) Negation: words such as "not", "never", "cannot", and "less" reverse the polarity of nearby terms (e.g., "I don't recommend this medicine"), requiring the model to capture local word dependencies rather than treat each token independently.

These challenges directly informed the choice of n-gram-based features (Section III-D) over pure unigram bag-of-words representations.

D. Feature Engineering and Selection

Text was vectorized using both CountVectorizer and TF-IDF, with unigrams, bigrams, and trigrams retained; 4-grams and 5-grams were excluded due to sparsity, higher overfitting risk, and diminishing accuracy returns beyond trigrams. Table II summarizes the key differences between the two vectorization strategies used.

Aspect	CountVectorizer	TF-IDF
Definition	Raw word-frequency counts	Frequency weighted by rarity across documents
Emphasis	Treats all terms equally	Downweights common terms, boosts rare ones
Best use	When raw frequency is informative	When document relevance/distinction matters

A Chi-squared test was then applied to evaluate the statistical dependence between each n-gram feature and the sentiment label, retaining the top-k most discriminative features. This both reduced dimensionality and improved generalization, since rare, highly document-specific n-grams (typical of 4-grams/5-grams) tend to overfit the training data without improving test performance.

E. Model Selection and Training

Four classifiers were trained and compared: Logistic Regression, Decision Tree (default and pruned, max depth = 11), Random Forest, and AdaBoost. Logistic Regression was chosen as a baseline for its efficiency on sparse, high-dimensional text data and its directly interpretable coefficients. Decision Trees capture non-linear splits but easily overfit high-dimensional text features unless depth-constrained. Random Forest, an ensemble of many trees trained on bootstrap samples, improves robustness to noise, while AdaBoost sequentially re-weights misclassified examples using shallow decision stumps as weak learners.

Data was split 80/20 for train/test. To tune hyperparameters — such as the Logistic Regression regularization strength (C) and solver, and Random Forest's number of estimators, maximum depth, and split criteria — we used grid search with 5-fold cross-validation on the training set, ensuring model selection was based on averaged validation performance rather than a single train/test split.

IV. RESULTS AND DISCUSSION

A. Model Comparison

On the IMDB movie-review dataset, Logistic Regression achieved the highest F1-score (0.889) among all baseline models, followed closely by its fine-tuned variant (0.886). Random Forest (0.844) and AdaBoost (0.829) performed moderately well, while single Decision Trees (0.70–0.72) lagged behind and showed clear signs of overfitting, with training accuracy near 100% but substantially lower test performance. Fig. 4 shows the comparative F1-scores across all evaluated models. The remaining three datasets (stock market, employee, and pharmaceutical reviews) were used to validate the preprocessing, EDA, and feature-engineering stages of the pipeline (Section III); a full quantitative model comparison on these datasets is left for future work, as noted in Section V.

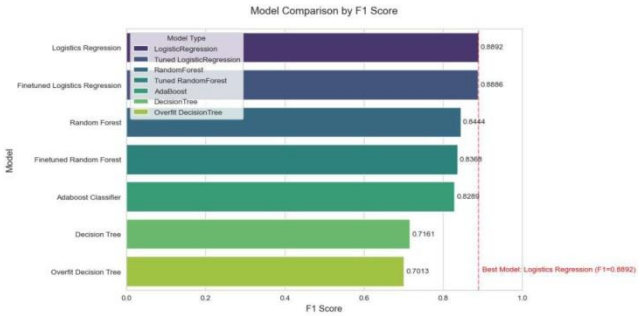

Fig. 4. Model comparison by F1-score.

B. Confusion Matrix and Metrics

In extended evaluation on the IMDB test set (12,000 reviews), the Logistic Regression + TF-IDF pipeline reached approximately 90.8% accuracy, with precision, recall, and F1-score all near 0.908, and ROC-AUC of 0.968. Fig. 5 presents the confusion matrix, showing 5,410 true negatives, 5,485 true positives, 554 false positives, and 551 false negatives — indicating balanced error rates across classes with no strong directional bias.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

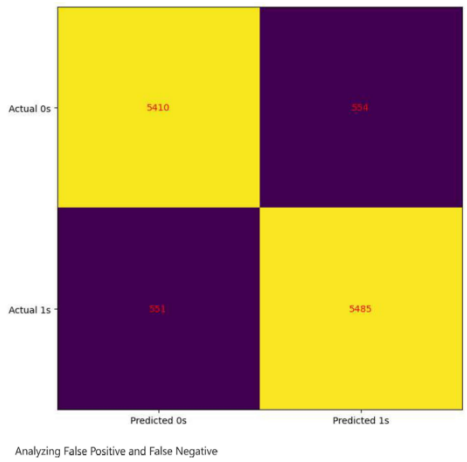

Fig. 5. Confusion matrix of the Logistic Regression model.

C. Explainable AI: LIME and SHAP

To interpret model decisions, LIME was applied to individual predictions to reveal which words drove a specific classification, while SHAP was used on the full test set to obtain global feature importance. Fig. 6 shows the SHAP summary plot: words such as "excellent" and "great" strongly push predictions toward positive sentiment, while "bad", "worst", and "terrible" push predictions toward negative sentiment — confirming that the model relies on intuitive, human-interpretable cues. Misclassifications were traced mainly to negation and sarcasm, which remain open challenges.

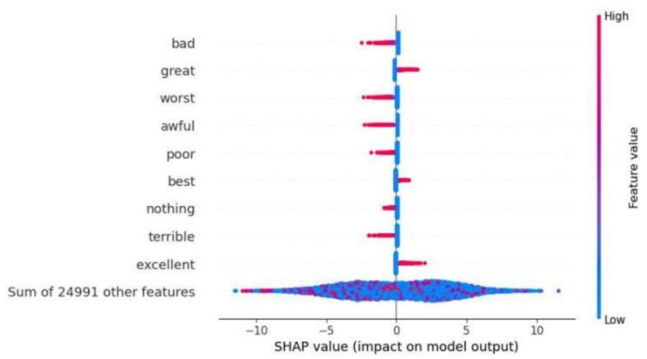

Fig. 6. SHAP summary plot of feature importance.

D. Neutral-Class Extension and Application

Although the base classifier was trained for binary classification, a threshold rule was added to the prediction logic: if the absolute difference between the positive and negative class probabilities is ≤ 0.30, the prediction is labeled "Neutral"; otherwise the class with higher average probability is assigned. This extends the binary classifier into a pseudo-trinary classifier without any retraining, and improves the model's expressiveness for ambiguous or mixed-sentiment text.

This complete pipeline was deployed as an interactive Streamlit-based web application, SACR (Sentiment Analysis for Comprehensive Review). Fig. 7 shows a sample output where a positive review is correctly classified along with its associated confidence scores. The application enables dataset upload (.csv, .xlsx, .txt, .json), automated text cleaning, exploratory data analysis (word clouds, n-gram distributions, class balance), feature engineering (CountVectorizer/TF-IDF with configurable n-gram range), and training/evaluation of all four classifiers — entirely through a no-code interface. Each stage's output persists across the session, so users can preprocess data once and test multiple models without repeating earlier steps.

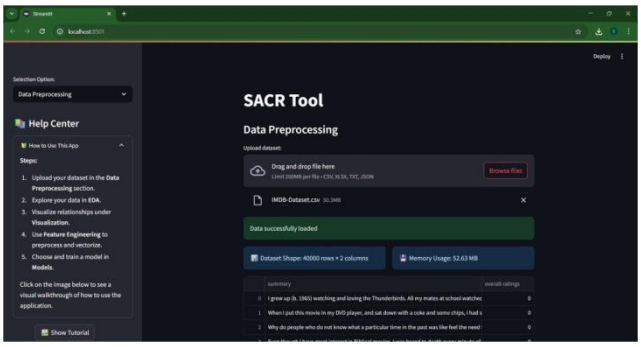

Fig. 7. SACR web application interface for end-to-end sentiment analysis.

This tool is intended to serve three audiences: students and ML beginners learning sentiment analysis concepts, researchers and developers who need to rapidly prototype models on new review data, and analysts who want to visualize and classify customer feedback without writing code.

V. CONCLUSION

This study proposed a robust, interpretable, and high-performing sentiment analyzer through a systematic four-phase pipeline — data preprocessing, feature engineering, model selection, and explainability analysis — validated on the IMDB movie-review dataset and cross-checked for generality on three additional review domains. Logistic Regression with TF-IDF emerged as the most effective and interpretable model, achieving up to 90.8% accuracy and an F1-score of 0.889–0.908 depending on the evaluation setting, while remaining lightweight and transparent through LIME and SHAP explanations. Ensemble methods (Random Forest, AdaBoost) offered competitive but slightly lower performance at higher computational cost, and unconstrained Decision Trees clearly overfit the high-dimensional text features.

We deliberately selected classical machine learning (Logistic Regression with TF-IDF) over transformer-based deep learning models such as BERT, prioritizing interpretability, low inference latency, and low computational cost — properties that are essential for transparent, resource-constrained, and regulation-sensitive deployment settings. This design choice is validated by our LIME/SHAP analysis, which confirms the model's decisions align with human-interpretable sentiment cues. BERT and similar contextual models remain the preferred choice when deep contextual, multilingual, or highly nuanced language understanding is the primary requirement, at the cost of higher computational and interpretive complexity.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

REFERENCES

[1] A. Kumar and A. Jaiswal, "Systematic literature review of sentiment analysis on Twitter using soft computing techniques," Concurrency and Computation: Practice and Experience, vol. 32, no. 1, art. e5107, Jan. 2020, doi: 10.1002/cpe.5107.

[2] A. Arora and P. Chakraborty, "Role of emotion in excessive use of Twitter during COVID-19 imposed lockdown in India," in Advances in Data Science and Management, Lecture Notes on Data Engineering and Communications Technologies, Springer, Singapore, 2021, pp. 370–377, doi: 10.1007/978-981-15-9774-9_34.

[3] K. C. Davis, "Over a decade of social opinion mining: a systematic review," Artificial Intelligence Review, vol. 54, pp. 4873–4965, Jun. 2021, doi: 10.1007/s10462-020-09955-1.

[4] E. Cambria and D. Das, "Affective computing and sentiment analysis," in A Practical Guide to Sentiment Analysis, E. Cambria, D. Das, S. Bandyopadhyay, and A. Feraco, Eds. Cham, Switzerland: Springer, 2017, pp. 1–10, doi: 10.1007/978-3-319-55394-8_1.

[5] E. Cambria, S. Poria, A. Gelbukh, and M. Thelwall, "Sentiment analysis is a big suitcase," IEEE Intelligent Systems, vol. 32, no. 6, pp. 74–80, Nov. 2017, doi: 10.1109/MIS.2017.4531228.

[6] F. Valencia, A. Gómez-Espinosa, and B. Valdés-Aguirre, "Price movement prediction of cryptocurrencies using sentiment analysis and machine learning," Entropy, vol. 21, no. 6, art. 589, Jun. 2019, doi: 10.3390/e21060589.

[7] F. A. Pozzi, E. Fersini, E. Messina, and B. Liu, "Challenges of sentiment analysis in social networks: an overview," in Sentiment Analysis in Social Networks, ch. 1, Morgan Kaufmann, 2017, pp. 1–11, doi: 10.1016/B978-0-12-804412-4.00001-2.

[8] F. Hemmatian and M. K. Sohrabi, "A survey on classification techniques for opinion mining and sentiment analysis," Artificial Intelligence Review, vol. 52, no. 3, pp. 1495–1545, 2019, doi: 10.1007/s10462-017-9599-6.

[9] F. Abid, M. Alam, M. Yasir, and C. Li, "Sentiment analysis through recurrent variants latterly on convolutional neural network of Twitter," Future Generation Computer Systems, vol. 95, pp. 292–308, Jun. 2019, doi: 10.1016/j.future.2018.12.018.

[10] H. W. Park, S. Park, and M. Chong, "Conversations and medical news frames on Twitter: infodemiological study on COVID-19 in South Korea," Journal of Medical Internet Research, vol. 22, no. 5, art. e18897, 2020, doi: 10.2196/18897.

[11] N. Ruffer, J. Knitza, S. Krusche, and F. Muehlensiepen, "Covid4Rheum: an analytical Twitter study in the time of the COVID-19 pandemic," Rheumatology International, vol. 40, pp. 2031–2037, 2020, doi: 10.1007/s00296-020-04710-5.

[12] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: review, opportunities and challenges," Briefings in Bioinformatics, vol. 19, no. 6, pp. 1236–1246, 2018, doi: 10.1093/bib/bbx044.

[13] K. Mite-Baidal, C. Delgado-Vera, E. Solís-Avilés, A. H. Espinoza, J. Ortiz-Zambrano, and E. Varela-Tapia, "Sentiment analysis in education domain: a systematic literature review," in Communications in Computer and Information Science, vol. 883, Springer, 2018, pp. 285–297, doi: 10.1007/978-3-030-00940-3_21.

[14] K. Ravi and V. Ravi, "A survey on opinion mining and sentiment analysis: tasks, approaches and applications," Knowledge-Based Systems, vol. 89, pp. 14–46, Nov. 2015, doi: 10.1016/j.knosys.2015.06.015.

[15] M. E. Mowlaei, M. S. Abadeh, and H. Keshavarz, "Aspect-based sentiment analysis using adaptive aspect-based lexicons," Expert Systems with Applications, vol. 148, art. 113234, Jun. 2020, doi: 10.1016/j.eswa.2020.113234.

[16] S. Subhashini, "Sentiment and context-aware recurrent convolutional neural network for sentiment analysis," in Proc. IEEE Int. Conf. on Recent Trends in Computing and Communication Technologies, Aug. 2023, pp. 1–6.

[17] V. Uma and K. E., "Intelligent sentinet-based lexicon for context-aware sentiment analysis," in Proc. Int. Conf. on Artificial Intelligence and Smart Systems (ICAIS), 2021, pp. 1–6.

[18] Y. Baashar, H. Alhussian, A. Patel, G. Kiru, N. Neamah, R. Anwar, and A. Alammary, "Customer relationship management systems (CRMS) in the healthcare environment: a systematic literature review," Computer Standards & Interfaces, vol. 71, art. 103442, 2020, doi: 10.1016/j.csi.2020.103442.

[19] K. Zvarevashe and O. O. Olugbara, "A framework for sentiment analysis with opinion mining of hotel reviews," in Proc. 2018 Conf. Information Communications Technology and Society (ICTAS), Durban, South Africa, Mar. 2018, pp. 1–4, doi: 10.1109/ICTAS.2018.8368746.