Decoding Consumer Sentiment through Machine Learning: Analysing Social Media Trends and Behaviours

C.K. Kotravel Bharathi, K. Elakkiyan

***Abstract:** Social media platforms have become indispensable channels for public opinion, customer feedback, and brand perception. Analysing this vast repository of user-generated content enables businesses to gain deep insights into consumer sentiment and behaviour. This paper presents the "Social Media Sentiment Analyzer," an interdisciplinary initiative that combines marketing and Information Technology to develop a machine learning-based tool for sentiment analysis. The tool processes social media posts to classify them as positive, negative, or neutral, offering organizations actionable insights for strategic decision-making [1].*

***Keywords:** Consumer Sentiment, Machine Learning, NLP, Marketing, Behaviour Insights*

Abbreviations:
BERT: Bidirectional Encoder Representations from Transformers
NLP: Natural Language Processing
SVM: Support Vector Machines
LSTM: Long Short-Term Memory
TF-IDF: Term Frequency-Inverse Document Frequency
NLTK: Natural Language Toolkit
AWS: Amazon Web Services

I. INTRODUCTION

The rapid proliferation of social media has transformed the landscape of consumer interaction and engagement. Platforms like Twitter, Facebook, and Instagram are not just communication channels but have evolved into rich data sources reflecting consumer opinions, preferences, and behaviours [2]. Understanding these sentiments is crucial for businesses aiming to stay competitive in a dynamic market environment.

From a marketing perspective, leveraging consumer behaviour concepts such as perception, motivation, and attitude formation is pivotal in interpreting social media data. Sentiment analysis empowers companies to decode the underlying emotions and opinions expressed online, allowing them to refine marketing strategies, enhance customer satisfaction, and build stronger brand loyalty [3].

Manuscript received on 08 April 2025 | First Revised Manuscript received on 25 April 2025 | Second Revised Manuscript received on 04 May 2025 | Manuscript Accepted on 15 June 2025 | Manuscript published on 30 June 2025.

*Correspondence Author(s)
Dr. C.K. Kotravel Bharathi*, Former Dean, SRM University – Tiruchirappalli Campus, Tamil Nadu, India. Email ID: doctorkotravel@gmail.com, ORCID ID: 0000-0003-0219-5545
K. Elakkiyan, Final year B.Tech. (IT) Student, Vellalar College of Engineering and Technology, Erode, Tamil Nadu, India. Email ID: kelakkiyanbc@gmail.com

II. OBJECTIVE

To develop a robust Natural Language Processing (NLP) based sentiment analysis tool capable of real-time classification of social media posts while supporting multiple languages, providing businesses with actionable insights into consumer sentiment and behaviour.

III. LITERATURE REVIEW

Sentiment analysis, also known as opinion mining, has garnered significant attention in academia and industry. Previous research has employed machine learning algorithms, including logistic regression, Support Vector Machines (SVM), Naive Bayes classifiers, and deep learning models such as Long Short-Term Memory (LSTM) networks and Bidirectional Encoder Representations from Transformers (BERT).

These models have demonstrated varying degrees of success in handling the nuances of natural language, including slang, sarcasm, and context dependency. Challenges persist in accurately interpreting sentiment due to the complexity of human language and the informal nature of social media communication [2].

A. Data Sources

i. *Datasets were sourced from Kaggle [5], including:*

- Flipkart Reviews Dataset [6]
- Twitter Sentiment Dataset [7]
- Sample Shopping Mall Dataset [8]

These diverse datasets provide a comprehensive foundation for analysis, allowing for a well-rounded examination of various e-commerce platforms and products. Each dataset was meticulously cleaned and integrated to ensure consistency and accuracy. Advanced data pre-processing techniques were employed to handle missing values and normalise the data, thereby facilitating robust model training and analysis.

IV. DETAILED PRODUCT RANGE

A. Dataset 1: Supermarket

i. *Total Products: 205,053*

- Cleaning Products: 61,516
- Furniture & Home Decor: 51,263
- Kitchenware & Appliances: 41,011
- Bedding & Linens: 20,505
- Home Electronics: 20,505

- Storage & Organization: 10,253
- Health & Personal Care: 10,253

B. Dataset 2: Flipkart

i. *Total Products: Similar distribution as Dataset 1*

- Cleaning Products: 60,000
- Furniture & Home Decor: 50,000
- Kitchenware & Appliances: 40,000
- Bedding & Linens: 20,000
- Home Electronics: 20,000
- Storage & Organization: 10,000
- Health & Personal Care: 10,000
- Dataset 3: Mobile Phones

V. METHODOLOGY

A. Program on NLP for Sentiment Analysis

An NLP program for sentiment analysis involves using natural language processing techniques to analyze and classify text into positive, negative, or neutral. The program begins by pre-processing the text, which involves steps such as tokenisation (splitting the text into words or sentences), removing stop words, and stemming or lemmatisation. Text is then converted into numerical representations using techniques like Term Frequency-Inverse Document Frequency (TF-IDF) [1]. Machine learning models such as Logistic Regression, SVM, or deep learning models like LSTM are trained on labelled data to predict sentiment. The program outputs a sentiment classification for the given text, aiding businesses and social media platforms in gauging public opinion and customer feedback [4].

B. Features

i. **Input**: Social media posts in English.

ii. **Output**: Sentiment labels – positive, negative, or neutral.

iii. **Language Support**: Initial focus on English with provisions for future expansion.

iv. **Training Data**: Labelled datasets of social media posts.

C. Data Collection and Pre-processing

i. **Data Integration**: Datasets were sourced from Kaggle [5] and integrated to create a comprehensive dataset for analysis.

ii. **Data Cleaning**: Handled missing values, removed duplicates, and normalised data.

iii. **Pre-processing Steps**:

D. Tokenization:

i. Word Tokenization: Splitting text into individual words.

ii. Sentence Tokenization: Breaking text into sentences for context analysis.

iii. Subword Tokenization: Handling complex words using methods like Byte Pair Encoding.

iv. Character Tokenization: Splitting text into individual characters to manage misspellings and typos.

v. Whitespace Tokenization: Splitting text based on spaces.

- **Stop Words Removal**: Eliminated commonly used words that do not contribute significantly to the sentiment analysis (e.g., "is," "the," "at").
- **Stemming and Lemmatization**: Reduced words to their root form to simplify the text without losing context.
- **Handling Emojis and Special Characters**: Converted emojis to textual sentiments; removed or encoded special characters.
- **Numerical Representation:**
- **TF-IDF Vectorization:** Converted textual data into numerical form based on the importance of words in the document corpus.
- **Model Development and Training**

VI. MODEL SELECTION

A. Machine Learning Models:

i. *Logistic Regression*: Used for its efficiency in binary and multi-class classification problems.

ii. *Support Vector Machines (SVM)*: Implemented for its effectiveness in high-dimensional spaces.

iii. *Naive Bayes Classifier*: Chosen for its simplicity and baseline performance.

B. Deep Learning Models:

i. *Long Short-Term Memory (LSTM)*: Utilized for capturing long-term dependencies in text sequences.

ii. *Bidirectional Encoder Representations from Transformers (BERT)*: Fine-tuned for advanced contextual understanding and managing complex language nuances [3].

C. Training Methodology:

i. **Supervised Learning**: Employed labelled datasets for training the models.

ii. **Cross-Entropy Loss Function**: Applied to measure the performance of classification models.

D. Cross-Validation:

i. *K-Fold Cross-Validation (k=5)*: The dataset was divided into five subsets to ensure model generalisation and prevent overfitting.

ii. *Data Leakage Prevention*: Ensured that pre-processing steps were conducted separately within each fold.

E. Optimization:

i. Used the Adam optimizer for efficient and effective convergence.

ii. Hyperparameter tuning was performed to optimize model performance.

F. Deployment:

i. Technology Stack

- Programming Language**: Python**
- Libraries**:**
- *Natural Language Toolkit (NLTK)*: For text pre-processing.
- TensorFlow *and Keras*: For

building and training deep learning models.

- *Scikit-learn*: For traditional machine learning algorithms.
- Transformers *(Hugging Face)*: For implementing BERT.
- Web Framework: Flask for developing the web interface.
- Cloud Hosting**:**
- Amazon Web Services (AWS)**:**
- *AWS Elastic Beanstalk* and *EC2 Instances*: For scalable and reliable deployment.
- *S3 Buckets*: For data storage and retrieval.

G. Challenges and Solutions:

i. *Multiple Meanings and Context Dependency:*

- **Challenge**: Words can have different meanings depending on context (e.g., "Great" in "I'm feeling great" vs. "Great, another delay!").
- **Solution**: Leveraged BERT for its ability to understand word meanings based on context, enhancing the model's semantic understanding [3].

ii. *Sarcasm Detection:*

- **Challenge**: Sarcasm often conveys the opposite sentiment of the literal words used.
- **Solution**: While detecting sarcasm remains complex, future iterations will integrate specialized modules

iii. *Slang and Informal Language:*

- **Challenge**: Rapidly evolving slang and informal expressions in social media.
- **Solution**: Incorporated continuous updates to the language model and expanded the vocabulary with the latest slang terms using dynamic word embedding [2].

iv. *Negations Handling:*

- **Challenge**: Negations can invert the sentiment of a sentence ("I don't like this" vs "I don't dislike this").
- **Solution**: Implemented algorithms to detect negation words and adjusted the sentiment scoring accordingly [1].

v. *Domain-Specific Sentiment Variations:*

- **Challenge**: Sentiment meanings can vary across different domains (e.g., "sick" meaning "cool" in slang).
- **Solution**: Fine-tuned models on domain-specific datasets and incorporated domain adaptation techniques to understand industry-specific jargon [2].

VII. RESULTS

A. Model Performance Metrics

i. *Logistic Regression:*

- Accuracy: 80%
- Precision: 78%
- Recall: 76%
- F1-Score: 77%

ii. *Support Vector Machine (SVM):*

- Accuracy: 82%
- Precision: 80%
- Recall: 79%
- F1-Score: 79.5%

iii. *Naive Bayes Classifier:*

- Accuracy: 78%
- Precision: 75%
- Recall: 73%
- F1-Score: 74%

iv. *LSTM Model:*

- Accuracy: 86%
- Precision: 84%
- Recall: 83%
- F1-Score: 83.5%

v. *Fine-Tuned BERT Model:*

- Accuracy: 93%
- Precision: 92%
- Recall: 91%
- F1-Score: 91.5%

vi. *Sample Outputs*

- Positive Sentiment: "Worth every penny," "Awesome," "Wonderful."
- Negative Sentiment: "Hated it!" "Utterly disappointed."
- Neutral Sentiment: "Does the job," "Fair."

VIII. DISCUSSION

The implementation of the Social Media Sentiment Analyzer demonstrates substantial potential in extracting meaningful insights from large-scale data [1]. By integrating advanced machine learning models and sophisticated NLP techniques, the tool effectively addresses challenges such as slang, sarcasm, and context dependency [2].

A. Marketing and Consumer Behaviour Insights

Understanding consumer sentiment allows businesses to tailor their marketing strategies more effectively. Key concepts from consumer behaviour applied include:

i. **Attitude Formation**: How consumers develop feelings toward a brand based on accumulated experiences and information.

ii. **Perception**: The process by which consumers select, organize, and interpret information to form a meaningful picture of the world.

iii. **Motivation**: Identifying the underlying drives that prompt consumers to take action.

By leveraging these insights, organizations can enhance customer satisfaction, foster brand loyalty, and influence purchasing decisions. The sentiment Analyzer aids in identifying not just what consumers are saying, but the underlying emotions and motivations driving those conversations.

IX. BENEFITS

A. **Public Opinion Analysis**: Provides real-time insights into consumer preferences, trends, and emerging needs.

B. **Brand Monitoring**: This enables proactive tracking of brand perception, early detection of potential PR crises, and management of reputational risks.

C. **Customer Feedback Analysis**: Facilitates a deeper understanding of customer satisfaction levels, informing product or service enhancements.

D. **Market Research**: Offers data-driven insights for strategic decision-making, enabling businesses to remain competitive and responsive.

i. *Implementation Plan*

- Technology Stack
- Programming Language: **Python**
- Libraries**:**
- *NLTK*: For text pre-processing.
- *Scikit-learn*: For machine learning algorithms.
- *TensorFlow and Keras*: For building deep learning models.
- *Transformers*: For implementing BERT.
- *Web Framework:* Flask
- *Deployment:* AWS for scalable cloud deployment

ii. *Steps to Implementation*

- **Data Collection**: Sourcing and integrating datasets from Kaggle and other platforms [5].
- **Data Pre-processing**:
- Cleaning text data.
- Tokenization, stemming and lemmatization.
- Removing stop words and handling negations.
- Converting text to numerical representations using TF-IDF.

iii. *Model Training:*

- Training traditional models: Logistic Regression, SVM, Naive Bayes.
- Training deep learning models: LSTM and fine-tuning BERT.
- Utilizing cross-validation techniques.

iv. *Evaluation:*

- Computing performance metrics: Accuracy, Precision, Recall, F1-Score.
- Hyperparameter tuning for optimization.

v. *Deployment:*

- Developing the web interface using Flask.
- Deploying the application on AWS for scalability and reliability.

vi. *Continuous Improvement:*

- Updating models with new data to handle evolving language patterns.
- Incorporating user feedback and improving UI/UX design.
- Monitoring performance and making iterative enhancements.

X. CONCLUSION

The Social Media Sentiment Analyzer effectively bridges Marketing and Information Technology to address contemporary challenges in sentiment analysis. By integrating advanced NLP techniques with marketing insights, the tool enables businesses to monitor and respond effectively to public sentiment.

The tool's ability to process and analyse real-time data enables organizations to stay ahead of market trends, understand consumer behaviour more deeply, and make informed strategic decisions. The integration of consumer behaviour concepts enhances the interpretability of the results, providing contextually rich insights.

A. Future Work

i. *Multilingual Support:* Expanding analysis capabilities to include multiple languages, catering to a global audience.

ii. *Enhanced Sarcasm and Irony Detection:* Implementing advanced NLP techniques to understand and interpret sarcastic or ironic statements.

iii. *Integration with Business Intelligence Tools:* Allowing seamless incorporation of sentiment analysis into existing business workflows and dashboards.

iv. *Ethical Considerations and Privacy Compliance:* Ensuring adherence to data protection regulations and ethics in data handling and analysis.

v. *Real-Time Analytics:* Improving the tool's ability to handle high-volume data streams with low latency.

ACKNOWLEDGMENT

We thank Kaggle for providing the datasets essential for this research. We also thank the developers of open-source tools and libraries such as Python, NLTK, Scikit-learn, TensorFlow, Keras, and Transformers, who were instrumental in this project.

DECLARATION STATEMENT

After aggregating input from all authors, I must verify the accuracy of the following information as the article's author.

- **Conflicts of Interest/ Competing Interests:** Based on my understanding, this article has no conflicts of interest.
- **Funding Support:** No organisation or agency has funded this article. This independence ensures that the research is conducted with objectivity and without any external influence.
- **Ethical Approval and Consent to Participate:** The content of this article does not necessitate ethical approval or consent to participate with supporting documentation.
- **Data Access Statement and Material Availability:** The adequate resources of this article are publicly accessible.
- **Author's Contributions:** The authorship of this article is contributed equally to all

participating individuals.

REFERENCES

1. Liu, B. (2012). Sentiment Analysis and Opinion Mining. *Synthesis Lectures on Human Language Technologies* (Vol. 5, Issue 1, pp. 1–167).
   DOI: https://doi.org/10.2200/S00416ED1V01Y201204HLT016
2. Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment Analysis Algorithms and Applications: A Survey. *Ain Shams Engineering Journal* (Vol. 5, Issue 4, pp. 1093–1113).
   DOI: https://doi.org/10.1016/j.asej.2014.04.011
3. Zhang, Y., & Wang, D. (2015). Deep Learning Based Sentiment Analysis on Twitter Data. *Proceedings of the IEEE International Conference on Big Data* (Vol. 1, Issue 1).
   DOI: https://doi.org/10.1109/BigData.2015.7363783
4. Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation* (Vol. 9, Issue 8, pp. 1735–1780).
   DOI: https://doi.org/10.1162/neco.1997.9.8.1735
5. Amazon Web Services (AWS). (n.d.). *Retrieved from* https://aws.amazon.com/

Kaggle.com/datasets

6. Flipkart Reviews Dataset. *Retrieved from* https://www.kaggle.com/datasets
7. Apple Twitter Sentiment Dataset. *Retrieved from* https://www.kaggle.com/datasets
8. Sample Shopping Mall Dataset. *Retrieved from* https://www.kaggle.com/datasets

AUTHOR'S PROFILE

Dr. C.K. Kotravel Bharathi has dedicated over 27 years to management education, making significant contributions to academia and research. He has mentored 25 MPhil scholars and six PhD candidates in Management Science, with two more under his guidance. His commitment to education is evident in the hundreds of MBA students he has taught. Dr. Bharathi has published over 30 research articles in esteemed national and international journals, covering key areas of Business Administration. He has also organised and participated in numerous seminars, conferences, faculty development programs (FDPs), and workshops. As a distinguished resource person, he has addressed international conferences, sharing his expertise. His professional affiliations include life memberships in prestigious organisations such as the All-India Management Association (AIMA), the National Institute of Personnel Management (NIPM), the Indian Society for Technical Education (ISTE), the National HRD Network (NHRD), the Indian Commerce Association, the Coimbatore Management Association (CMA), and the Case Research Society of India (CRSI). Dr. Bharathi has played an active role in academic governance, serving as a member of the Academic Planning Board of Bharathiar University and as a Senate Member of Periyar, Bharathiar, and Bharathidasan Universities. His contributions extend to curriculum design, notably the creation of the *Functional English for Executives* course for Bharathiar University in 2007, which Periyar University later adopted. An alumnus of IIM-Ahmedabad, he completed his full-time FDPM there. Beyond academia, he has a deep passion for stage oration, vocal music, poetry, lyric writing, tune composing, literary debates, and empirical research. Dr. Bharathi has held several leadership positions, including Principal/Director of GRD Academy of Management, Principal of SRM Trichy Arts and Science College, Founder Dean of the College of Science and Humanities at SRM University's Trichy Campus, and Dean for Data Analytics & Quality Management at SRM Group of Institutions, Tiruchirappalli Campus.

K. Elakkiyan is a motivated and detail-oriented B.Tech. (Information Technology) Student at Vellalar College of Engineering and Technology, expected to graduate in May 2025 with a higher CGPA. With a strong foundation in programming languages such as Java, Python, C, C++, and SQL, he is proficient in web development technologies, including HTML, CSS, JavaScript, and ReactJS. His technical expertise encompasses database management with MySQL, as well as tools such as Git and Docker. Elakkiyan has undertaken multiple projects, including an expense tracker written in Python, a to-do list application developed in Java, and a house price prediction system utilising linear regression. His commitment to continuous learning is reflected in his certifications from Spoken Tutorial and NPTEL in various domains, including Python, Java, and Software Testing. Beyond academics, he has held leadership roles in the English Literary Association, serving as Vice President, Outreach Head, and Executive Member. His interests lie in web development, data visualization, and data warehousing. Driven by a passion for technology and innovation, Elakkiyan seeks opportunities to apply his skills in a dynamic IT environment, contributing effectively to innovative projects and professional growth.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of the Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)/ journal and/or the editor(s). The Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP) and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.