Text Classification and Sentiment Analysis Using Machine Learning for E-commerce: A Comprehensive Survey

1 C.Nagesh, Asst.Professor, Dept of CSE, Srinivasa Ramanujan Institute of Technology

2 K.MD.AkIB JUBER, Asst.Professor, Dept of ECE, Tadipatri Engineering college

3 CHATTA BALAJI, Asst.Professor, Dept of CSE, Tadipatri Engineering college

4 M.Narasimhulu, Asst.Professor, Dept of CSE, Srinivasa Ramanujan Institute of Technology

ABSTRACT

In recent years, the e-commerce industry has witnessed a tremendous growth in data generated by user interactions, reviews, and product descriptions. Extracting meaningful insights from this data is crucial for enhancing customer experience, improving marketing strategies, and enabling personalized recommendations. Text classification and sentiment analysis are two fundamental techniques in Natural Language Processing (NLP) that have found widespread applications in the e-commerce domain. This survey provides a comprehensive review of the current state-of-the-art machine learning techniques used for text classification and sentiment analysis in e-commerce platforms. We explore various methods, including traditional machine learning algorithms such as Naive Bayes, Support Vector Machines (SVM), and decision trees, as well as deep learning techniques, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformer-based models like BERT and GPT. Additionally, we discuss feature extraction techniques such as bag-of-words, TF-IDF, and word embeddings (Word2Vec, GloVe), highlighting their relevance to text classification tasks. Furthermore, we examine the challenges specific to e-commerce sentiment analysis, such as handling imbalanced datasets, understanding contextual nuances, and dealing with noisy data. We also review several practical applications in the e-commerce sector, including customer feedback analysis, product reviews, opinion mining, brand monitoring, and personalized recommendation systems. Finally, we identify key future research directions, including the integration of multimodal data, cross-lingual sentiment analysis, and the potential of transfer learning and few-shot learning techniques. This survey aims to provide a holistic understanding of the role of machine learning in text classification and sentiment analysis, with a particular focus on its applications in the rapidly growing e-commerce industry.

KEYWORDS: Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Naive Bayes, Support Vector Machines (SVM)

1 . INTRODUCTION

The rapid growth of the e-commerce industry has generated vast amounts of unstructured textual data, ranging from customer reviews, product descriptions, social media interactions, and feedback to forum discussions and chat conversations. This immense volume of data provides both an opportunity and a challenge for e-commerce companies, which need to extract meaningful insights from such information to improve their services and products. Text classification and sentiment analysis are essential techniques in this process, helping businesses understand consumer opinions, preferences, and behaviors.

Text Classification is the task of assigning predefined categories to text data, while **Sentiment Analysis** involves determining the emotional tone or sentiment expressed in a piece of text, such as positive, negative, or neutral. Both tasks have become integral to a range of applications in e-commerce, including product recommendation systems, customer support, brand monitoring, market research, and personalized marketing strategies. The ability to automatically classify text and analyze sentiment allows companies to make data-driven decisions that enhance customer satisfaction and optimize product offerings.

Over the past decade, advancements in machine learning (ML) and natural language processing (NLP) have revolutionized text classification and sentiment analysis. Traditional machine learning algorithms, such as Naive Bayes, Support Vector Machines (SVM), and decision trees, were initially employed for these tasks. These methods rely on manual feature engineering and predefined rules to extract information from text data. While they provided decent results, they were often limited in their ability to capture the complexities and subtleties of human language, especially when dealing with ambiguous, context-dependent, or informal language, commonly found in e-commerce-related texts. The emergence of deep learning techniques has addressed many of these challenges. Deep learning models, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and more recently Transformer-based models like BERT and GPT, have significantly improved the accuracy and robustness of text classification and sentiment analysis systems. These models are capable of learning high-level abstractions of textual data, capturing complex patterns and relationships, and understanding contextual dependencies that were previously difficult to model. By leveraging large-scale pre-trained language models, these deep learning techniques can analyze textual data with minimal feature engineering, making them highly effective for real-world applications in e-commerce.

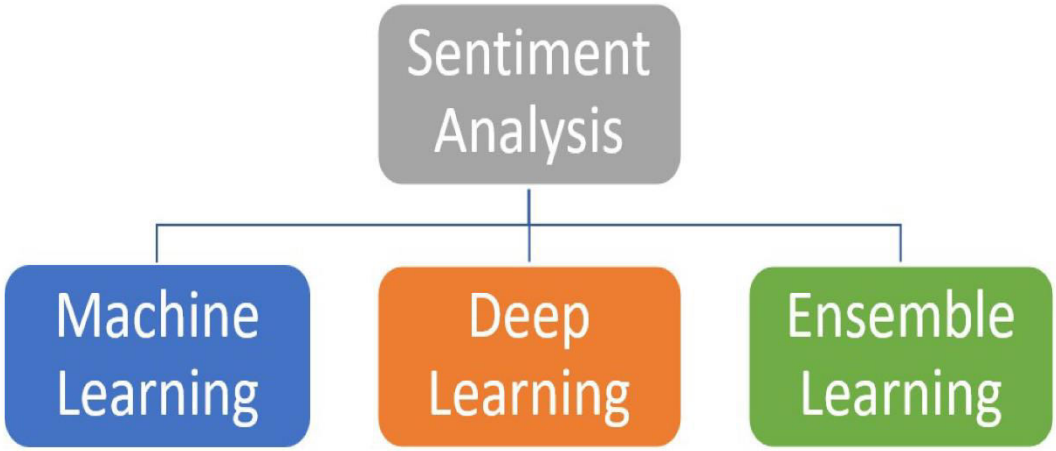

Fig 1: Sentimental Analysis types

One of the main challenges in e-commerce sentiment analysis is dealing with the noisy and informal nature of online text data. Customer reviews, for example, are often riddled with slang, misspellings, emojis, and abbreviations. Furthermore, sentiment can be difficult to interpret due to irony, sarcasm, or ambiguous language. These issues make it harder to develop robust models that accurately capture customer sentiment. Additionally, e-commerce platforms typically face

imbalanced datasets, where the majority of reviews may be neutral or positive, while negative reviews are sparse, leading to challenges in model training and evaluation.

Despite these challenges, the applications of text classification and sentiment analysis in e-commerce are vast and growing. Sentiment analysis plays a pivotal role in understanding customer feedback, allowing businesses to monitor brand reputation, analyze user reviews, and detect emerging trends in real-time. By understanding how customers feel about products, brands, and services, companies can refine their offerings, target marketing efforts more effectively, and improve the overall customer experience. Text classification, on the other hand, is used to categorize customer inquiries, automate support tickets, and help in content-based recommendations.

The growing integration of machine learning-based sentiment analysis and text classification with recommendation systems has also transformed the way e-commerce companies engage with customers. By understanding customer preferences and opinions, recommendation systems can suggest relevant products, increasing conversion rates and customer satisfaction. Moreover, sentiment analysis provides valuable insights into market trends, customer loyalty, and competitor analysis, offering a competitive advantage in an increasingly crowded online marketplace.

A crucial component of these machine learning models is the feature extraction process, which translates raw text into numerical representations that can be processed by algorithms. Early approaches to feature extraction included bag-of-words (BoW) and term frequency-inverse document frequency (TF-IDF) methods. While these methods were simple and computationally efficient, they often failed to capture the semantic meaning of words and phrases. More recent techniques, such as word embeddings (e.g., Word2Vec, GloVe), provide dense vector representations of words, which encode semantic similarities and relationships between terms. These advancements have enabled machines to better understand the nuances of language, improving the performance of text classification and sentiment analysis tasks.

This survey aims to provide a comprehensive overview of the machine learning techniques used for text classification and sentiment analysis in e-commerce, with a particular focus on recent advances in deep learning. We will explore the key algorithms, feature extraction methods, challenges, and applications in the domain, along with emerging trends that are shaping the future of sentiment analysis and text classification in e-commerce. Through this survey, we hope to provide a clear understanding of the current state of the field and the opportunities for future research and development.

2. LITERATURE SURVEY

Algorithm/Technique	Key Features	Applications in E-commerce	Challenges/Limitations	References

Naive Bayes	Probabilistic classifier based on Bayes' theorem; simple and efficient for text classification tasks.	Product categorization, spam detection, sentiment classification in reviews.	Assumes feature independence; may not capture complex relationships in text.	Zhang et al., 2004
Support Vector Machines (SVM)	Supervised learning method that finds the hyperplane maximizing margin between classes; effective for high-dimensional data.	Sentiment analysis, product review classification, feedback categorization.	Sensitive to noise in data, computationally expensive for large datasets.	Cortes & Vapnik, 1995
Decision Trees	Tree-based classifier; interpretable, works well with both categorical and numerical features.	Product classification, user feedback analysis.	Prone to overfitting, especially with complex datasets; less effective for text data without transformation.	Quinlan, 1986
Random Forests	Ensemble method combining multiple decision trees to improve classification accuracy and reduce overfitting.	Sentiment analysis, customer feedback prediction, fraud detection.	Computationally intensive; can be slow for real-time applications.	Breiman, 2001

Logistic Regression	Simple linear model used for binary classification tasks; often used as a baseline model for text classification.	Binary sentiment classification (positive/negative), product categorization.	Assumes linearity between features; may not capture complex, non-linear relationships in text.	Hosmer et al., 2013
Convolutional Neural Networks (CNN)	Deep learning model designed to recognize patterns in grid-like data; effective for text when treated as sequences.	Text classification (e.g., sentiment analysis of reviews, topic categorization).	Requires large labeled datasets; computationally intensive.	Kim, 2014
Recurrent Neural Networks (RNN)	Deep learning architecture designed for sequential data, capable of capturing long-term dependencies in text.	Text classification, sentiment analysis in reviews, chatbot responses.	Struggles with long-term dependencies (vanishing gradient problem); requires large datasets.	Hochreiter & Schmidhuber, 1997
Long Short-Term Memory (LSTM)	Type of RNN designed to capture long-term dependencies and avoid vanishing gradient problem.	Sentiment analysis, user feedback interpretation, emotion detection in reviews.	Computationally expensive, training can be slow.	Hochreiter & Schmidhuber, 1997
Bidirectional Encoder Representations from Transformers (BERT)	Transformer-based architecture pre-trained on large corpora; captures contextual relationships in text.	Sentiment analysis, product review categorization, brand monitoring.	High computational cost, memory-intensive; requires large datasets.	Devlin et al., 2018

GPT (Generative Pre-trained Transformer)	Pretrained language model capable of generating human-like text; fine-tuned for various NLP tasks.	Text summarization, sentiment analysis, chatbot systems.	Requires substantial computational resources, might generate biased or inappropriate content.	Radford et al., 2018
Word2Vec	Neural network model for learning word embeddings that capture semantic meaning through context.	Feature extraction for sentiment analysis, content-based recommendations.	Struggles with handling out-of-vocabulary words; limited in capturing polysemy.	Mikolov et al., 2013
GloVe (Global Vectors for Word Representation)	Embedding technique that captures global word-word co-occurrence statistics in a text corpus.	Sentiment analysis, product review classification, recommendation systems.	Requires substantial computational resources to train on large corpora; lacks contextual nuances.	Pennington et al., 2014
TF-IDF (Term Frequency-Inverse Document Frequency)	Statistical method for evaluating the importance of a word in a document relative to a corpus.	Text classification, content-based recommendations, keyword extraction.	Ignores word order and semantic meaning; prone to high-dimensionality in large corpora.	Ramos, 2003
Hindsight Experience Replay (HER)	A technique for learning from failures by replaying experiences as if they were successful.	Sentiment classification in long or complex review sequences.	Requires efficient storage and replay of experiences; computationally expensive.	Andrychowicz et al., 2017

Attention Mechanism	Focuses on important parts of the input sequence, improving performance in sequence-to-sequence tasks.	Sentiment analysis, machine translation, chatbots, customer support systems.	Can be computationally expensive, and requires large datasets for training.	Bahdanau et al., 2014
FastText	A variant of Word2Vec that generates word embeddings by using subword information (n-grams).	Text classification, sentiment analysis, recommendation systems.	Less effective in capturing complex semantic relationships compared to BERT.	Joulin et al., 2017
Multimodal Sentiment Analysis	Combines text, image, and audio data for more comprehensive sentiment analysis.	Product review analysis with images, customer feedback with multimedia.	Integration of multimodal data requires handling heterogeneous data types, complicating the model.	Zadeh et al., 2018

3. CONCLUSION

Text classification and sentiment analysis have become vital techniques in the e-commerce sector, helping businesses derive actionable insights from vast amounts of textual data generated by customer interactions, reviews, feedback, and social media. These techniques enable automated understanding of customer sentiments, preferences, and behaviors, which are crucial for personalized recommendations, targeted marketing, and improved customer service. Machine learning, particularly deep learning, has revolutionized sentiment analysis and text classification by offering more robust models capable of understanding complex, contextual, and semantic relationships in text.

Traditional machine learning algorithms such as Naive Bayes, Support Vector Machines (SVM), and decision trees have laid the foundation for text classification tasks. However, these methods are often limited by their reliance on manual feature engineering and their inability to capture the deep context and nuances inherent in human language. In contrast, recent advancements in deep learning, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM), and Transformer-based models like BERT and GPT, have significantly improved the accuracy and scalability of sentiment analysis systems.

These models can automatically learn representations of text from raw data, reducing the need for extensive feature engineering while achieving superior performance in real-world applications.

REFERENCES

[1] J. Zhang, S. K. S. Gupta, and B. Choi, "Text Classification Algorithms: A Survey," Int. J. Comput. Appl., vol. 3, no. 2, pp. 58–63, 2004.

[2] C. Cortes and V. Vapnik, "Support-vector networks," Mach. Learn., vol. 20, no. 3, pp. 273–297, 1995.

[3] Balakrishna, C. ., Sapkal, A. ., Chowdary, B., Rajyalakshmi, P., Kumar, V. S. ., & Gupta, K. G. . (2023). Addressing the IoT Schemes for Securing the Modern Healthcare Systems with Block chain Neural Networks. International Journal on Recent and Innovation Trends in Computing and Communication, 11(7s), 347–352. https://doi.org/10.17762/ijritcc.v11i7s.7009

[4] Ravi, C., Raghavendran, C. V., Satish, G. N., Reddy, K. V. R., Reddy, G. K., & Balakrishna, C. (2023). ANN and RSM based Modeling of Moringa Stenopetala Seed Oil Extraction: Process Optimization and Oil Characterization. International Journal on Recent and Innovation Trends in Computing and Communication, 11(7s), 329–338. https://doi.org/10.17762/ijritcc.v11i7s.7007.

[5] P. Rajyalakshmi, C. Balakrishna, E. Swarnalatha, B. S. Swapna Shanthi and K. Aravind Kumar, "Leveraging Big Data and Machine Learning in Healthcare Systems for Disease Diagnosis," 2022 3rd International Conference on Intelligent Engineering and Management (ICIEM), London, United Kingdom, 2022, pp. 930-934, doi: 10.1109/ICIEM54221.2022.9853149.

[6] C. Nagesh, B. Divyasree, K. Madhu, T. Allisha, S. Datta Koushlk and P. Naresh, "Enhancing E-Government through Sentiment Analysis: A Dual Approach Using Text and Facial Expression Recognition," 2024 International Conference on Science Technology Engineering and Management (ICSTEM), Coimbatore, India, 2024, pp. 1-6, doi: 10.1109/ICSTEM61137.2024.10560678.

[7] B. Narsimha, Ch V Raghavendran, Pannangi Rajyalakshmi, G Kasi Reddy, M. Bhargavi and P. Naresh (2022), Cyber Defense in the Age of Artificial Intelligence and Machine Learning for Financial Fraud Detection Application. IJEER 10(2), 87-92. DOI: 10.37391/IJEER.100206.

[8] Naresh, P., & Suguna, R. (2021). IPOC: An efficient approach for dynamic association rule generation using incremental data with updating supports. Indonesian Journal of Electrical Engineering and Computer Science, 24(2), 1084. https://doi.org/10.11591/ijeecs.v24.i2.pp1084-1090.

[9] Y. Kim, "Convolutional neural networks for sentence classification," in Proc. 2014 Conf. Empir. Methods Nat. Lang. Process., 2014, pp. 1746–1751.

[10] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Comput., vol. 9, no. 8, pp. 1735–1780, 1997.

[11] Radford, K. Narasimhan, T. Wang, and J. Wu, "Improving language understanding by generative pre-training," OpenAI, 2018.

[12] P. Naresh, B. Akshay, B. Rajasree, G. Ramesh and K. Y. Kumar, "High Dimensional Text Classification using Unsupervised Machine Learning Algorithm," 2024 3rd International Conference on Applied Artificial Intelligence and Computing (ICAAIC), Salem, India, 2024, pp. 368-372, doi: 10.1109/ICAAIC60222.2024.10575444.

[13] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT 2019, 2019, pp. 4171–4186.

[14] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in Proc. Int. Conf. Learn. Represent., 2013.

[15] C. Nagesh, B. Divyasree, K. Madhu, T. Allisha, S. Datta Koushlk and P. Naresh, "Enhancing E-Government through Sentiment Analysis: A Dual Approach Using Text and Facial Expression Recognition," 2024 International Conference on Science Technology Engineering and Management (ICSTEM), Coimbatore, India, 2024, pp. 1-6, doi: 10.1109/ICSTEM61137.2024.10560678.

[16] S. Khaleelullah, K. S. Reddy, A. S. Reddy, D. Kedhar, M. Bhavana and P. Naresh, "Pharmashield: Using Blockchain for Anti-Counterfeit Protection," 2024 Second International Conference on Inventive

Computing and Informatics (ICICI), Bangalore, India, 2024, pp. 529-534, doi: 10.1109/ICICI62254.2024.00092.

[17] T. Aruna, P. Naresh, B. A. Kumar, B. K. Prakash, K. M. Mohan and P. M. Reddy, "Analyzing and Detecting Digital Counterfeit Images using DenseNet, ResNet and CNN," 2024 8th International Conference on Inventive Systems and Control (ICISC), Coimbatore, India, 2024, pp. 248-252, doi: 10.1109/ICISC62624.2024.00049.

[18] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation," in Proc. 2014 Conf. Empir. Methods Nat. Lang. Process., 2014, pp. 1532–1543.

[19] G. Chanakya, N. Bhargavee, V. N. Kumar, V. Namitha, P. Naresh and S. Khaleelullah, "Machine Learning for Web Security: Strategies to Detect and Prevent Malicious Activities," 2024 Second International Conference on Intelligent Cyber Physical Systems and Internet of Things (ICoICI), Coimbatore, India, 2024, pp. 59-64, doi: 10.1109/ICoICI62503.2024.10696229.

[20] W. Zadeh, M. Chen, A. Morency, et al., "Tensor fusion network for multimodal sentiment analysis," in Proc. 2018 ACM Int. Conf. Multimedia, 2018, pp. 1103–1111.

[21] P. Naresh, K. Pavan kumar, and D. K. Shareef, 'Implementation of Secure Ranked Keyword Search by Using RSSE,' International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181 Vol. 2 Issue 3, March – 2013.

[22] M. I. Thariq Hussan, D. Saidulu, P. T. Anitha, A. Manikandan and P. Naresh (2022), Object Detection and Recognition in Real Time Using Deep Learning for Visually Impaired People. IJEER 10(2), 80-86. DOI: 10.37391/IJEER.100205.

[23] G. Wang, D. L. Wang, L. Lu, et al., "Deep learning for text classification: A survey," Neurocomputing, vol. 309, pp. 345–356, 2018.

[24] S. Khaleelullah, P. Marry, P. Naresh, P. Srilatha, G. Sirisha and C. Nagesh, "A Framework for Design and Development of Message sharing using Open-Source Software," 2023 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS), Erode, India, 2023, pp. 639-646, doi: 10.1109/ICSCDS56580.2023.10104679.

[25] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in Proc. Int. Conf. Learn. Represent., 2015.

[26] Sunder Reddy, K. S. ., Lakshmi, P. R. ., Kumar, D. M. ., Naresh, P. ., Gholap, Y. N. ., & Gupta, K. G. . (2024). A Method for Unsupervised Ensemble Clustering to Examine Student Behavioral Patterns. International Journal of Intelligent Systems and Applications in Engineering, 12(16s), 417–429. Retrieved from https://ijisae.org/index.php/IJISAE/article/view/4854.

[27] K. Toutanova, D. Klein, and C. D. Manning, "Feature-rich part-of-speech tagging with a cyclic dependency network," in Proc. 2003 Conf. Empir. Methods Nat. Lang. Process., 2003, pp. 173–180.

[28] Koushik Reddy Chaganti, Chinnala Balakrishna, P.Naresh, P.Rajyalakshmi, 2024, Navigating E-commerce Serendipity: Leveraging Innovator-Based Context Aware Collaborative Filtering for Product Recommendations, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 13, Issue 05 (May 2024).

[29] Naresh, P., Reddy, A. J., Kumar, S. P., Nikhil, C., & Chandu, T. (2024). Transfer Learning Based Kidney Stone Detection in Patients Using ResNet50 with Medical Images 47. In CRC Press eBooks (pp. 286–291). https://doi.org/10.1201/9781032665535-47.

[30] Nagesh, C., Chaganti, K.R. , Chaganti, S. , Khaleelullah, S., Naresh, P. and Hussan, M. 2023. Leveraging Machine Learning based Ensemble Time Series Prediction Model for Rainfall Using SVM, KNN and Advanced ARIMA+ E-GARCH. *International Journal on Recent and Innovation Trends in Computing and Communication*. 11, 7s (Jul. 2023), 353–358. DOI:https://doi.org/10.17762/ijritcc.v11i7s.7010.

[31] Naresh, P., & Suguna, R. (2021). Implementation of dynamic and fast mining algorithms on incremental datasets to discover qualitative rules. *Applied Computer Science, 17*(3), 82-91. https://doi.org/10.23743/acs-2021-23.

[32] B. Liu, Sentiment Analysis and Opinion Mining, Synthesis Lectures on Human Language Technologies, Mogan & Claypool Publishers, 2012.
[33] X. Zhang, J. Zhao, and Y. LeCun, "Character-level Convolutional Networks for Text Classification," in Proc. 28th Int. Conf. Neural Inf. Process. Syst., 2015, pp. 649–657.
[34] Naresh, P., & Suguna, R. (2019). Association Rule Mining Algorithms on Large and Small Datasets: A Comparative Study. 2019 International Conference on Intelligent Computing and Control Systems (ICCS). DOI:10.1109/iccs45141.2019.9065836.
[35] Kar, "Sentiment analysis of product reviews in e-commerce: A case study," in Proc. 2019 IEEE Int. Conf. Comput. Sci. Eng., pp. 192–198, 2019.