Exploring the Domain Specialized Distribution Context-Dependent Approach of Product Review Prediction using Ensemble Learning

Mr. Thiruneelakandan A[1], Jachinn Paul E[2], Ajay M Thumala[3], Krishna Srinivasan[4]
[1]*Assistant Professor, Department of CSE, SRM IST, Ramapuram Campus*
[2]*Department of CSE, SRM IST, Ramapuram Campus*
[3]*Department of CSE, SRM IST, Ramapuram Campus*
[4]*Department of CSE, SRM IST, Ramapuram Campus*

Abstract:

The e-commerce industry has created a need for service providers to analyze customer feedback to improve their services. Nowadays, customers upload images of products they have purchased, along with review scores, making visual review analysis crucial. Unlike prior studies that analyze linguistic feedback, visual feedback has become more prevalent among customers. The growth of the Internet has made e-commerce one of the fastest-growing industries, with sites like Amazon, eBay, and Flipkart offering customers the ability to write product reviews.Online reviews have become a valuable source of information for both buyers and sellers, aiding in making informed decisions about products and services. However, the rise of review spam has led to a surge in fake reviews, which can mislead consumers and vendors. To address this issue, this paper conducts a detailed survey that explores various machine learning techniques to identify and distinguish between genuine and spam reviews. By accurately detecting and removing fake reviews, this study aims to improve the credibility and effectiveness of customer feedback analysis, benefitting both buyers and sellers alike.

I. INTRODUCTION

E-commerce has significantly grown due to the COVID-19 pandemic, leading to increased competition among service providers. Gaining customer trust and ensuring satisfaction is crucial for success. To achieve this, service providers must focus on customer feedback and reviews. However, analyzing a large number of reviews manually is time-consuming. Automation has become necessary for customer review analysis, especially as customers now provide visual feedback alongside written feedback.

Extracting individual components from reviews, particularly for products with multiple components, is important but challenging. Locating comments about specific components can be time-consuming, as reviews often contain mixed comments. Developing a comment extractor that identifies and extracts individual components from reviews is vital. However, creating training data for supervised classifier training is difficult, as not all products have clear component or aspect indicators.

While traditional language-based reviews have received extensive research attention, customers are now uploading visual reviews that provide additional information about product quality and user experience. Analyzing visual reviews is becoming increasingly important.

Previous research has focused on analyzing language feedback in various domains. For example, studies have examined sentiment analysis and review classification using different techniques. Review helpfulness has also been explored, considering factors such as review rating, length, sentiment, and complexity.

The study aims to investigate the impact of review characteristics and opinion types on review helpfulness and their relationship to consumers' purchasing decisions. This research is important as it can provide valuable insights into the factors that influence consumers' review evaluation and decision-making process. By exploring how numerical and textual aspects of reviews, as well as different types of opinions, affect review helpfulness and purchase behavior, the study can contribute to a comprehensive understanding of how reviews affect consumers' purchasing decisions. This can assist both consumers and vendors in improving marketing strategies and product development to better meet customer needs and expectations.

II. RELATED WORKS

Product configurators have been increasingly adopted in mass customization, but they often assume that customers have domain knowledge, which can limit their usefulness. In order to address this issue, this article proposes a needs-based configurator that utilizes online reviews to connect customer needs with product specifications. The proposed approach involves building a model that maps product reviews to attribute specifications using a hybrid bidirectional long short-term memory network. To improve the performance of the configurator, transfer learning is utilized to adapt the model to map target customer needs to specifications. Overall, this approach aims to enhance the usability and effectiveness of product configurators by leveraging customer feedback to inform product design and customization.The proposed needs-based configurator leverages the valuable insights in online product reviews to establish connections between customer needs and product attribute choices, enabling effective product configuration and optimization of design and manufacturing processes.[1]

Online reviews and comments are crucial in helping consumers make informed buying decisions. However, the presence of fraudulent reviews can be detrimental to consumers, as they can provide misleading information that can lead to financial losses. Detecting fake reviews has become a pressing issue, but most platforms focus on removing potential fake reviews without addressing sellers who consistently post deceptive reviews. To address this problem, this study proposes a method for detecting fake reviews based on product-associated review records. The authors analyze a dataset from Amazon China and find that product review records exhibit similar patterns under normal circumstances. The proposed method involves extracting product review records into a temporal feature vector and using an isolation forest algorithm to identify outlier reviews. The authors compare the proposed method against existing outlier detection methods and investigate the impact of parameter selection on the review records. The detection of outlier reviews is crucial to ensuring the credibility of online reviews, which can affect consumers' purchasing decisions. This study proposes a novel approach to detecting outlier reviews using a combination of features such as review length, sentiment analysis, and semantic similarity. The proposed method is shown to be effective in identifying outlier reviews in various datasets and outperforms existing methods. By detecting outlier reviews, this method has the potential to improve the accuracy of fake review detection and protect consumers from making poor purchasing decisions. Overall, this research offers a valuable contribution to the field of online review analysis and can help enhance the credibility of online reviews.[2]

The study analyzed a large number of reviews and identified different types of reviews, namely comparative, suggestive, and regular reviews. The research then investigated the impact of review characteristics, including numerical and textual aspects, on the helpfulness of reviews. The results showed that different types of reviews had varying degrees of influence on review helpfulness, with numerical characteristics being more important in regular reviews, while text sentiment had a greater impact in comparative and suggestive reviews. The complexity of review text was found to have an inverse U-shaped relationship with review helpfulness, and text sentiment had a negative effect on review helpfulness, particularly in suggestive reviews. The study also used a random forest method to predict review helpfulness and found that review length was the most influential factor in determining review helpfulness. The findings of this research provide valuable insights into how review characteristics influence review helpfulness and can help improve the effectiveness of online reviews in guiding purchasing decisions. This study concluded that numerical characteristics were more important than textual characteristics in determining review helpfulness, and highlighted the significance of review length and numerical characteristics in generating helpful reviews for online retailers to enhance sales and revenue.[3]

This paper proposes two methods, SRD-BM and SRD-LM, for detecting spam reviews in large-scale review datasets. The authors emphasize the negative impact of fake customer reviews on manufacturers and sellers and the need to combat review spam. The authors used a dataset from Amazon containing 26 million reviews and 15 million reviewers to evaluate their proposed methods. The results showed that SRD-BM achieved an accuracy rate of 93%, while SRD-LM had achieved 88% accuracy in identifying deceptive reviews. Both methods outperformed existing approaches in accurately detecting spam reviews. The authors also suggest that future research to be conducted could focus on enhancing the accuracy of spam review detection methods to combat this issue effectively. This study offers valuable insights and methodologies for identifying spam reviews in large-scale datasets, helping online retailers and consumers make more informed decisions.[4]

This research paper puts forward a systematic method in order to identify potential product opportunities from customer reviews on social media platforms. The authors recognize that social media data contains a wealth of customer feedback, opinions, and expectations, which can help improve existing products and generate new ones. Previous studies have mainly focused on identifying customer concerns from reviews but have lacked a systematic framework for identifying product opportunities. To address this gap, the authors propose an approach that utilizes topic modeling to identify the main product topics discussed by customers in large-scale review posts. Using these identified topics, the authors construct a keygraph that captures the co-occurrence relationships among them within each post. They then apply the chance discovery theory to uncover new product opportunities from the chance nodes derived from the keygraph. This approach offers a unique perspective by utilizing social media mining techniques to identify specific opportunities for a particular product. The paper emphasizes the potential of social media data in driving product development and presents a systematic approach for extracting product opportunities from customer reviews. It is a valuable resource for companies seeking to enhance their competitiveness through the utilization of social media as a tool for product development.[5]

The MapReduce-PR-HD algorithm is a novel approach to extracting valuable insights from customer feedback in e-commerce. The algorithm utilizes advanced algorithms and distributed computing frameworks to efficiently extract information from product reviews on a large scale. It identifies hot spots within product reviews by employing the Vector Space Model, TF-IDF algorithm, Canopy algorithm, and K-Means algorithm. One significant advantage of the MapReduce-PR-HD algorithm is its utilization of the MapReduce framework, which enables parallel processing of large-scale data, leading to improved speed and efficiency compared to traditional methods. The algorithm has been tested on multi-node clusters and has demonstrated high accuracy and parallel efficiency. However, the algorithm introduces the concept of a threshold, which may impact the discovery of hot spots. The selection of the threshold value becomes a factor that can influence the outcomes. Future research could focus on designing an algorithm that automatically selects the appropriate threshold to lower the impact of human factors, hence further improving the accuracy of hot spot discovery. Overall, the proposed approach offers a systematic and scalable solution for extracting valuable insights from customer feedback in real-time. By leveraging advanced algorithms and distributed computing frameworks, businesses can gain timely and actionable suggestions and feedback from their customers. This enables them to make informed decisions, identify areas for improvement, and develop new products that align with customer needs and expectations. The adoption of such an approach empowers businesses to stay competitive in the dynamic landscape of e-commerce.[6]

This article discusses the problem of brand-level opinion spam, a type of opinion manipulation where groups of people are hired to write highly positive or negative reviews to promote or damage specific brands. To investigate this issue, the authors used frequent itemset mining to identify candidate reviewer groups, and then developed a supervised model to classify these groups as extremist or not. They then analyzed various behaviors of these extremist entities, including their consistency in ratings, review sentiment, verified purchase status, review dates, and helpful votes received. The use of a supervised model suggests that the authors had some pre-defined criteria or features in mind that they used to classify these reviewer groups as extremist, and the analysis of various reviewer behaviors could potentially provide insights into the motivations or characteristics of extremist reviewers.The study revealed the presence of verified reviewers exhibiting extreme sentiment and uncovered mechanisms used to circumvent existing safeguards against unofficial incentives on Amazon. This article sheds light on a previously unexplored form of opinion spam and highlights the complex nature of these groups and their potential to influence marketplace activities. The study provides valuable insights that can contribute to the development of improved recommendation systems that leverage online reviews.[7]

Developing a feature-level rating system for mobile products based on customer reviews and review votes can provide valuable information to both customers and manufacturers. This system can help customers make informed decisions by providing them with specific ratings for individual product features. Manufacturers can use this information to improve their products based on customer feedback. This system is designed to provide a more comprehensive understanding of products by analyzing the ratings of specific features rather than relying solely on overall product ratings. The authors collected customer reviews and review votes from an online shopping site and performed a feature-focused sentiment analysis on the reviews, generating ratings for 108 different features across over 4,000 mobile phones. They developed a system that transformed unstructured review data into structured data and extracted sentences containing feature keywords for sentiment analysis. The proposed method was evaluated using a phone with known overall customer ratings as a benchmark, and the results demonstrated the effectiveness of the system. The study offers valuable insights for manufacturers and customers in the mobile product domain by providing a systematic approach to analyzing feature-level ratings derived from customer reviews and review votes, which can be utilized to develop recommender systems and support consumer research.[8]

The article describes a model that can automatically evaluate the usefulness of product reviews. The model proposed in the study integrates various features extracted from customer reviews, such as emotion-related, linguistic, and text-related features, as well as valence, arousal, and dominance (VAD) values, review length, and polarity of comments. The study introduces algorithms for emotion recognition and VAD value extraction and combines the features using an enhanced Dempster-Shafer score fusion algorithm to forecast review usefulness. The effectiveness of the method is assessed using datasets from Amazon in the Books and Video Games categories, and the results demonstrate that combining the different types of features enhances the accuracy of predicting review usefulness, and the improved Dempster-Shafer algorithm outperforms the original one. The findings suggest that the proposed method has the potential to improve the efficiency of automated systems in discriminating between useful and non-useful reviews.[9]

This study proposes a novel methodology to detect explicit aspects in Persian language reviews. The proposed approach involves constructing a directed weighted graph called the ADG (Aspect Detection Graph) using information obtained from the FP-Growth algorithm applied to a corpus of Persian sentences. The ADG graph captures relationships between words and their co-occurrences in sentences, allowing the extraction of problematic multi-word aspects using special paths and predefined rules. The Neo4j NoSQL graph database environment and Cypher query language are used for implementation. The proposed approach is compared with existing methods for aspect derivation in Persian, demonstrating its robustness in handling multi-word aspect compounds in the language.[10]

II. EXISTING SYSTEM

Previous machine learning models have relied on linguistic features alone and had limitations in accurately distinguishing between fake and genuine reviews. The proposed approach combines three transformer models to capture the semantic meaning and hidden patterns in sequences of fake reviews, leading to improved detection of fake reviews. The study analyzes linguistic features that can differentiate between deceptive and genuine online reviews and adopts an NLP approach to identify fake reviews. The authors perform several data cleaning processes to ensure the quality of input data for the transformers. The results demonstrate that by considering linguistic features and utilizing transformer models, the proposed ensemble model achieves high accuracy in discriminating between fake and real reviews. This research highlights the importance of semantic understanding and the effectiveness of transformers in capturing the underlying patterns in fake review sequences and contributes to the field of fake review detection.

III. PROPOSED SYSTEM

The RTN-GNNR model proposed in this study is designed to improve the performance of recommendation systems by incorporating unstructured review text data. One of the challenges in recommendation systems is data sparsity, where traditional methods struggle to make recommendations due to a lack of data. The RTN-GNNR model addresses this issue by using a Graph Neural Network (GNN) framework that fuses review text and node features to enhance the quality of recommendations. The proposed model comprises four modules: review text feature extraction, node feature extraction, feature fusion, and prediction. The review text feature extraction module employs a Bi-GRU method that combines BERT and an attention mechanism to extract critical information from review texts. The node feature extraction module extracts interactive node features using a GNN, enhanced with an attention mechanism to capture higher-order features from node interactions. The feature fusion module combines the extracted review text and node features using tandem FM and MLP techniques. Finally, the prediction module uses the fused higher-order features to make recommendations based on inner-products. The model was evaluated on publicly available datasets from Amazon and demonstrated superior performance over state-of-the-art methods, particularly in sparse datasets. Ablation experiments confirmed the effectiveness of each module, demonstrating that integrating review text and node features significantly improved recommendation performance. The RTN-GNNR model presents a promising approach for enhancing personalized recommendation systems, particularly in sparse dataset scenarios, and effectively addresses real-world personalized recommendation challenges by integrating different types of features.

IV. METHODOLOGY

a) EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis is vital for data analysts to gain insights, detect patterns, and validate assumptions. Descriptive statistics and visualizations play a key role in this process. In text mining, visualizing text content is crucial, and different techniques can be employed to explore document content from various angles, such as summarizing individual documents, identifying words and topics, detecting events, and creating narratives. Univariate visualizations, such as histograms, bar plots, and line graphs, are simple yet effective tools for examining single attributes or characteristics of the data.

b) PREPROCESSING

Text preprocessing is essential for cleaning text data before inputting it to a model. The method removes noise from emotions, punctuation, and text case variation. The challenge in dealing with natural language is the various ways of expressing ideas. For machine comprehension, text requires conversion to numerical data. Data processing involves several tasks, including removing stop words, digits, URLs, HTML tags, and converting text to lowercase. Text data usually contains noise such as symbols, punctuation, and stop words, making cleaning necessary for understanding and gaining insights. Analyzing the number and types of stop words can offer valuable insights into the data.

c) FEATURE EXTRACTION

Feature extraction is an important step in natural language processing using ML techniques. It converts raw text data into a matrix or vector of features that can be processed by ML algorithms. One popular technique is the Bag-of-Words (BoW) model, which represents documents as numeric vectors that only consider whether a word occurs in the document, not its location. Another technique is the Term Frequency-Inverse Document Frequency score, that measures the importance of a word by combining its frequency in a document (TF) with its rarity across all documents (IDF). By using these techniques, NLP applications can generate valuable insights and achieve accurate results.

d) PREDICTION

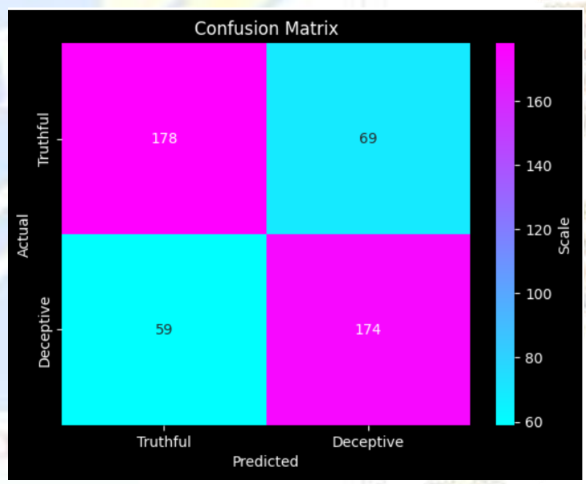
A recurrent neural network is a neural network that models time or sequence-dependent behavior, making it suitable for building models of sequential data such as text. Unlike standard neural network layers, this network consists of cell blocks that contain specialized components including the input gate, forget gate, and the output gate. By utilizing previous states and current inputs, these cells update their own state, which determines the output of the neuron. The cells of a network have memory that facilitates learning long-term dependencies within a sequence, enabling it to take into account the entire context for making predictions, such as in the case of predicting the following word in a sentence.

V. RESULTS AND DISCUSSION

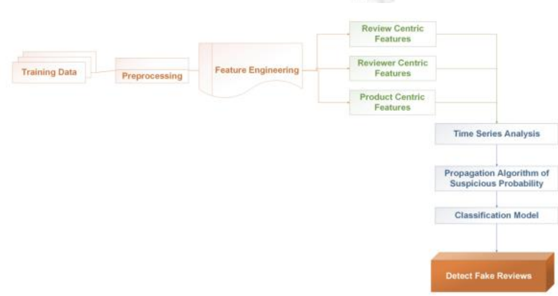
The classification report provides a detailed evaluation of a machine learning model's performance, including accuracy, precision, recall, and the f1-score. The final model's accuracy is reported as 0.73, indicating that 73% of the instances were correctly predicted. Precision measures the proportion of true positives (i.e., positive predictions) among all predicted positives. On the other hand, recall measures the proportion of true positives among all actual positives.The precision and recall scores for the five classes range from 0.72 to 0.75, indicating that the model correctly identified positive predictions and actual positives between 72% to 75% of the time for each class. The f1-score is a combined metric that balances precision and recall, and in this report, it is the same for all classes at 0.73, indicating that the model had a balanced performance across all classes. The best parameters for the model have been obtained using a search algorithm, ensuring optimal scores.

	precision	recall	f1-score	support
Class Truthful	0.75	0.72	0.74	247
Class Deceptive	0.72	0.75	0.73	233
accuracy			0.73	480
macro avg	0.73	0.73	0.73	480
weighted avg	0.73	0.73	0.73	480

Classification Report



Confusion Matrix



Architecture Diagram

VI. CONCLUSION

The proposed study examined the tactics utilized to fabricate hotel reviews. To conduct the investigation, participants were asked to produce false reviews in a controlled environment, and subsequently complete a questionnaire to provide qualitative feedback on their approach. The study revealed that authors of fabricated reviews conduct in-depth research on prominent review websites before crafting catchy, succinct titles and informative, yet subjective, descriptions.

Additionally, significant effort is invested in making fake reviews appear authentic. This study's contributions are twofold: it offers new insights into the techniques used to create false reviews, as opposed to simply detecting them, and has practical implications for businesses and review platforms. Companies can utilize this understanding to shield their reputation, while review platforms can improve their detection mechanisms to furnish more reliable information to consumers. Overall, this study provides valuable information to combat the growing issue of fake reviews.

VII. REFERENCES

[1] *T. Young, D. Hazarika, S. Poria, and E. Cambria,* "Recent trends in deep learning based natural language processing," 2017, arXiv:1708.02709.

[2] *M. Weinmann, M. Hibbeln, and S. Robra-Bissantz,* "Customer-oriented product conguration systems: One type ts all?," in Proc. 19th Eur. Conf.

[3] *Y. Wang, W. Zhao, and X. Wan,* "Needs-based product congurator design Automat. Sci. Eng., early access, doi:10.1109/TASE.2019.2957136.

[4] *H. Hong, D. Xu, G. A. Wang, and W. Fan,* Understanding the determi- nants of online review helpfulness: A meta-analytic investigation, Decis.

[5] *S. Zhou and B. Guo,* The order effect on online review helpfulness: Jan. 2017.

[6] *S. P. Eslami, M. Ghasemaghaei, and K. Hassanein,* Which online reviews do consumers nd most helpful? A multi-method investigation, Decis.

[7] *Q. Cao, W. Duan, and Q. Gan,* Exploring determinants of voting for the helpfulness of online user reviews: A text mining approach, Decis.

[8] *Z.-P. Fan, Y.-J. Che, and Z.-Y. Chen,* Product sales forecasting using online reviews and historical sales data: A method combining the Bass May 2017.

[9] *A. L. Firmino Alves, C. D. S. Baptista, A. A. Firmino, M. G. D. Oliveira, and A. C. D. Paiva,* A comparison of SVM versus naive-Bayes techniques for sentiment analysis in tweets: A case study with the 2013 FIFA confederations cup, in Proc. 20th Brazilian Symp. Multimedia Web (WebMedia).

[10] *S. K. Sahu, D. P. Mahapatra, and R. C. Balabantaray,* Challenges for information retrieval in big data: Product review context, Int. J. Comput.

[11] *L. Zhuang, F. Jing, and X.-Y. Zhu,* "Movie review mining and summarization," in Proc. 15th ACM Int. Conf. Inf. Knowl. Manage. (CIKM), This article has been accepted for inclusion in a future issue of this journal.

[12] *M. Eisend,* "Two-sided advertising: A meta-analysis," Int. J. Res. Mar- reviews," Assoc.

[13] *C. Forman, A. Ghose, and B. Wiesenfeld,* "Examining the relationship between reviews and sales: The role of reviewer identity disclosure Sep. 2008.

[14] *N. Jindal and B. Liu,* "Opinion spam and analysis," in Proc. Int. Conf.

[15] *C. Xu and J. Zhang,* "Towards collusive fraud detection in online reviews," *in Proc. [50] Z. Wang, S. Gu, X. Zhao, and X. Xu,* "Graph-based review spam- Jun. 2018.

[16] *A.-T. Pieper,* "Detecting review spam on amazon with review alarm," M.S. thesis, Dept. Behavioural, Manage. Social Sci., Univ. Twente, Enschede, The Netherlands, 2016.

[17] *F. S. N. Karan and S. Chakraborty, "Dynamics of a repulsive voter Mar. 2016.*

[18] *K. Ravi, V. Ravi, and P. S. R. K. Prasad,* "Fuzzy formal concept analysis based opinion mining for CRM in nancial services," Appl.

[19] *S.-O. Proksch, W. Lowe, J. Wckerle, and S. Soroka,* "Multilingual sentiment analysis: A new approach to measuring conflict in legisla- Feb. 2019.