INTEGRATING DEMOGRAPHICS AND CATEGORIES IN SENTIMENT ANALYSIS: A HYBRID REGRESSION APPROACH

K. ANUSHA[1], SANAY KUMAR APPISETTY[2], D. VASUMATHI[3]

[1]Research Scholar & [3]Professor, Department of Computer Science and Engineering, University College of Engineering, Science &Technology Hyderabad, JNTU, Hyderabad, 500085, India.
[2]Student, Department of Computer Science and Engineering, Vallurupalli Nageswara Rao Vignana Jyothi Institute of Engineering &Technology, Pragathi Nagar, Hyderabad, 500090, India.

Abstract: The research introduces a novel hybrid sentiment extraction technique designed to enhance the understanding of customer sentiment in the context of online product reviews and electronic commerce. The method combines text-based sentiment analysis with numerical rating regression to create a comprehensive sentiment profile for each product. It begins by collecting textual data from product reviews and applies natural language processing to assign emotion ratings, considering factors like irony, context, and subjectivity. Additionally, it incorporates demographic data of reviewers through feature engineering, allowing a detailed analysis of sentiment variations across different consumer groups. The technique also employs regression analysis to examine numerical ratings given by reviewers, identifying patterns of sentiment related to product qualities, overall satisfaction, and demographic preferences. By merging textual sentiment scores with numerical rating data, the hybrid algorithm is able to detect subtle emotions and sentiment fluctuations across various products and demographic groups. This capability can assist companies in improving product quality, marketing strategies, and targeting specific customer segments. The findings from this approach offer valuable insights for data-driven consumer sentiment research, potentially contributing to corporate growth and enhanced customer satisfaction. The hybrid model's ability to adapt to various product needs and consumer demographics renders it a potent tool for informed decision-making in e-commerce.

Key words: Hybrid sentiment extraction, Text-based sentiment analysis, Numerical rating regression, Natural language processing, Consumer sentiment research, E-commerce decision-making;

1.INTRODUCTION

In the digital age of e-commerce and online marketplaces, understanding customer sentiment is essential for businesses seeking competitive advantage and consumer satisfaction. The abundance of product reviews, social media feedback, and online surveys offers a wealth of textual data reflecting diverse customer experiences and emotions. This data, rich in insights, can significantly influence corporate decision-making. However, sentiment analysis, which involves interpreting nuanced emotions in text while considering demographic factors and aligning with numerical consumer ratings, is complex. This research introduces an innovative hybrid approach for sentiment extraction tailored to specific products and demographics, blending text-derived sentiment scores with numerical rating analysis through regression methods.

The rise of internet shopping has revolutionized how consumers interact with products and services. Product reviews, a crucial source of information for potential buyers, express thoughts, feelings, and opinions, aiding others in making informed

decisions. Yet, the vast amount of unstructured textual data in these reviews poses a challenge. Extracting meaningful sentiment insights requires sophisticated natural language processing (NLP) techniques. Furthermore, consumer sentiment is a complex construct influenced by product features, personal preferences, and demographic factors, necessitating tailored marketing strategies based on a deep understanding of varied emotions across different customer groups.

Traditional sentiment analysis approaches typically rely on either text-based analysis using NLP to decipher sentiments in written content or on numerical ratings that provide a measurable sentiment indication. Text-based analysis can capture complex attitudes in evaluations but struggles with sarcasm, ambiguity, and subjectivity. Conversely, numerical ratings, while straightforward, lack detailed insights into specific product aspects influencing sentiment and don't consider demographic variances. These limitations have led to the development of hybrid sentiment analysis, combining the strengths of both text-based and numerical rating methods to offer a more complete and accurate representation of emotions.

This research presents a cutting-edge hybrid model for sentiment extraction, surpassing traditional methods by customizing to individual products and customer demographics. It blends sentiment ratings from textual evaluations with numerical rating analyses using regression algorithms. Advanced NLP techniques are employed to extract nuanced sentiment ratings from text, considering context, tone, and emotional subtleties. Additionally, regression analysis extracts insights from numerical ratings, revealing patterns and connections between ratings and sentiment. This hybrid approach's adaptability to different products and demographics enables firms to gain insights into how various product features impact sentiment and allows for demographic-specific marketing strategies. This model equips organizations with data-driven tools to improve customer interactions and grow in the dynamic world of e-commerce

2. RELATED WORKS

The literature on sentiment analysis has significantly evolved, highlighting the critical role of understanding customer sentiment in the era of digital commerce and the widespread availability of online product reviews. This evolution has necessitated more advanced methods for extracting insights from textual data, with sentiment analysis becoming a key focus within natural language processing (NLP) to interpret complex emotions in customer feedback. Traditional methods have relied on text-based or numerical rating analyses, but recent literature emphasizes the development of hybrid models that combine these approaches, tailored specifically to product and demographic analysis. This literature review highlights the advancements, challenges, and unique advantages of hybrid sentiment analysis models, showcasing their ability to integrate textual and numerical data for comprehensive sentiment extraction.

In recent years, sentiment analysis has gained significant attention in the field of natural language processing and data analytics. Several studies have explored innovative approaches to extract sentiment insights from textual data, with a focus on enhancing the understanding of customer sentiment in various domains. For instance, Zaoli Yang et al. [1] introduced a comprehensive online product decision support system that combines sentiment analysis with a fuzzy cloud-based multi-criteria model, demonstrating its effectiveness across multiple e-commerce platforms.

Mouthami Kuppusamy and Anandamurugan Selvaraj [2] presented a novel hybrid deep learning model for aspect-based sentiment analysis, contributing to the advancement of sentiment analysis techniques. Abhale B. A, Bachhav M. K., and Patil Y. P. [3] explored the crucial task of fake review detection, employing multidimensional representations and fine-grained aspect analysis for more robust sentiment evaluation. Mukkamula Venu Gopalachari et al. [4] focused on aspect-based sentiment analysis across multiple domains, leveraging word embeddings to enhance the accuracy and applicability of sentiment analysis methods. Lastly, Mubashar Hussain et al. [5] proposed a multi-layered rule-based technique for explicit aspect extraction from online reviews, addressing the challenge of aspect identification insentiment analysis. These pioneering contributions form the backdrop for our research, which introduces a hybrid model for sentiment extraction tailored to product and demographic specifics, combining textual sentiment scores with numerical rating analysis through regression techniques.

These pioneering contributions form the backdrop for our research, which introduces a hybrid model for sentiment extraction tailored to product and demographic specifics, combining textual sentiment scores with numerical rating analysis through regression techniques.

In the realm of sentiment analysis and opinion mining, recent studies have delved into diverse domains and linguistic contexts to refine sentiment analysis techniques. Suntarin Sangsavate, Sukree Sinthupinyo, and Achara Chandrachai [6] conducted a comparative study on supervised learning and semi-supervised learning in the context of Thai financial news sentiment, shedding light on effective strategies for sentiment classification in low-resource languages. In a unique blend of social networks and e-commerce analysis, S. Uma Maheswari and S. S. Dhenakaran [7] explored opinion mining, offering insights into the amalgamation of opinions from integrated platforms. Geospatial data analysis found its place in sentiment studies as well, with Muhammad Ahmad, Kazim Jawad, Muhammad Bux Alvi, and Majdah Alvi [8] focusing on Google Maps data to understand the distribution and sentiment associated with clothing brands in South Punjab, Pakistan. The sentiment classification technique presented by K. Anuradha, M. Vamsi Krishna, and Banitamani Mallik [9] introduced the Normal Discriminant Piecewise Regressive (NDPR) approach, contributing to the diverse landscape of sentiment analysis methodologies. Lastly, Pallavi Mishra and Sandeep Kumar Panda [10] tackled the challenging task of explicit aspect extraction from online reviews, leveraging dependency structure-based rules and the root node technique, adding valuable insights to the field of aspect-based sentiment analysis. These studies collectively inform this research, which introduces a hybrid model for sentiment extraction tailored to product and demographic specifics, bridging the gap between text-based sentiment analysis and numerical rating analysis using regression techniques.

3. RESEARCH PROBLEM

Developing a hybrid model that efficiently combines textual product evaluations with numerical ratings presents a significant research challenge, necessitating an integrated analytical framework that leverages the wealth of information from both data sources for enhanced sentiment extraction. The addition of demographic-specific sentiment analysis introduces further complexity, requiring sophisticated data segmentation, feature engineering, and model adaptation strategies

to accurately capture sentiment variations across different user groups. The process involves not only deriving sentiment scores from text, which must account for natural language nuances like sarcasm, context, and emotion intensity but also employing regression analysis to predict numerical ratings by identifying sentiment elements. Constructing a robust regression model that can forecast ratings while addressing biases, outliers, and the non-linear relationships between sentiment and ratings is a considerable challenge within this research domain.

Evaluating the hybrid model's effectiveness, particularly its accuracy in sentiment extraction and rating prediction, underscores the complexities involved in research. Identifying appropriate metrics and conducting extensive testing to assess the model's performance across various products and demographics are critical steps. The challenges extend to the model's generalizability and scalability, ensuring its applicability to a broad spectrum of product categories and user demographics. Addressing scalability issues is essential for managing large datasets and diverse product types. Furthermore, the real-world application of the hybrid model, including its practical utility in enhancing customer insights, product development, and marketing strategies, along with ethical considerations in demographic analyses, are pivotal. Ensuring interpretability and explainability within the model is crucial for building user trust and facilitating informed decision-making, highlighting the need for ongoing research into user interaction and model refinement based on user feedback. These challenges collectively contribute to the development of a robust, versatile hybrid model for sentiment analysis within the realms of product evaluation and demographic-specific assessment, offering valuable insights for both businesses and researchers.

4. PROPOSED SOLUTIONS

The suggested method aims to get a more profound understanding of customer sentiment and satisfaction by using a novel technique that combines the capabilities of text analytics with regression analysis. In the current era of digital commerce and the proliferation of user-generated content, it has become imperative for companies to comprehend and measure sentiment from product evaluations. This is crucial in order to make well-informed choices and improve user experiences. Nevertheless, current sentiment analysis techniques sometimes struggle to accurately capture the complexities of user attitudes, particularly when confronted with various demographic cohorts and the amalgamation of textual evaluations and numerical ratings. In order to address this disparity, our hybrid approach to sentiment extraction undertakes a comprehensive exploration, integrating the capabilities of natural language processing and regression analysis to generate a refined sentiment score that is sensitive to the particular product categories and unique demographic segments. This novel methodology not only propels the domain of sentiment analysis forward but also introduces fresh opportunities for tailored product suggestions, focused marketing tactics, and better-informed product development choices. Within this part, we will explore the complexities of our suggested solution, providing a comprehensive analysis of its constituent elements, methodology, and the anticipated ramifications it may have for both enterprises and researchers.

4.1 Sentiment Extraction

Sentiment extraction is a sophisticated process that involves analyzing and interpreting emotions and opinions expressed in textual product evaluations and

numerical ratings. By integrating advanced natural language processing techniques and regression analysis, this approach aims to derive nuanced sentiment scores and predict consumer ratings, addressing the complexities of sarcasm, context, and emotion intensity within textual data. The challenge of effectively capturing sentiment variations across diverse demographic groups requires meticulous data segmentation and model adaptation. This methodology not only enhances the understanding of consumer feedback but also provides critical insights for improving product development, marketing strategies, and customer satisfaction in the dynamic landscape of e-commerce and digital marketplaces.

4.2 Combined Rating Generation

The concept of combined rating generation emerges as a pivotal aspect of the hybrid model for sentiment analysis, integrating both textual and numerical data to produce a comprehensive sentiment score. This approach leverages advanced algorithms to synthesize insights from qualitative text evaluations and quantitative numerical ratings, aiming to capture the full spectrum of consumer sentiment towards products. By addressing the intricacies of natural language and the direct feedback indicated by numerical ratings, the model offers a nuanced understanding of customer experiences. This multifaceted sentiment score facilitates a deeper analysis of consumer preferences and behaviors, enabling businesses to tailor their strategies more effectively to meet diverse customer needs and enhance overall satisfaction.

4.3 Category wise Score Generation

Category-wise score generation within the hybrid model framework involves segmenting data according to specific product categories and applying tailored sentiment analysis techniques to each segment. This process allows for the extraction of nuanced sentiment scores that reflect the unique attributes and consumer perceptions of different product types. By integrating textual analysis with numerical ratings, the model discerns the varying degrees of sentiment across categories, facilitating a granular understanding of consumer feedback. Such an approach enables the identification of category-specific trends and preferences, offering valuable insights into how different aspects of products resonate with or detract from consumer satisfaction. This category-specific sentiment scoring is instrumental in guiding product development, marketing strategies, and targeted improvements, ultimately enhancing the consumer experience across diverse product offerings.

4.4 Sentiment Prediction

Sentiment prediction stands as a crucial facet of sentiment analysis, involving the intricate task of forecasting the emotional tone behind textual feedback or numerical ratings provided by users. This process employs advanced computational techniques to analyze data from product evaluations, social media comments, or customer surveys, aiming to discern the underlying sentiments—be it positive, negative, or neutral. By leveraging natural language processing and machine learning algorithms, sentiment prediction models can interpret nuances of language, such as sarcasm, context, and emotional intensity, to generate accurate sentiment scores. These models are particularly valuable in e-commerce and digital marketing, where understanding

consumer sentiment towards products or services can guide strategic decisions, enhance customer satisfaction, and foster targeted marketing efforts.

5. WORKING METHODOLOGY

The initial stage of the process involves the collection of necessary data, which comprises written product reviews along with their numerical ratings, sourced from diverse channels including e-commerce and social media platforms, or through surveys. To prepare the data for analysis, it undergoes preprocessing to remove extraneous elements like stopwords and special characters, and is then tokenized into discrete units. This step is crucial for ensuring the cleanliness and uniformity of the text data, addressing any missing values or anomalies present, and incorporating demographic information such as age, gender, and location, when available.

Following data preparation, the focus shifts to understanding the emotions expressed within the textual evaluations through the application of various natural language processing (NLP) techniques. Textual data is transformed into a numerical representation via methods like tokenization, vectorization, TF-IDF, and word embeddings (e.g., Word2Vec, GloVe), allowing for sentiment analysis. This analysis can be conducted using lexicon-based approaches, assigning sentiment scores to words, or by employing machine learning models to categorize reviews into positive, negative, or neutral sentiments.

Subsequently, sentiment scores are calculated for each text review, reflecting the expressed sentiment's polarity. If demographic information is included, it enables segmentation of reviews by demographic characteristics, facilitating sentiment analysis within specific groups. This segmentation reveals patterns and variations in sentiment across different demographic categories, enhancing understanding of diverse consumer perspectives.

The final part of the process involves applying regression analysis to predict numerical ratings based on derived sentiment scores and possibly other variables. The model, chosen from techniques like linear regression or neural networks based on dataset specifics and performance, is trained and evaluated using metrics such as Mean Absolute Error (MAE) or Root Mean Squared Error (RMSE). This phase culminates in the creation of hybrid sentiment scores that amalgamate text-based sentiment analysis with numerical rating predictions, offering a more nuanced understanding of consumer sentiment. The effectiveness of this hybrid model is assessed by comparing its output against original numerical ratings, using visual representations to highlight findings, particularly focusing on capturing nuanced sentiment shifts and demographic-specific differences.

6. RESULTS AND DISCUSSIONS

In the Results and Discussion section, we present the significant insights and outcomes produced by our hybrid sentiment extraction model. By conducting thorough study and testing, we shed light on the ways in which our novel methodology enables companies to effectively interpret subtle emotions expressed in product evaluations, while also catering to the specific needs and preferences of various demographic segments. In this analysis, we explore the prediction accuracy of the model, its ability to uncover sentiment differences particular to different demographic groups, and its correlation with numerical ratings. This part functions as the medium through which we provide a detailed representation of the model's

performance, its practical ramifications, and its capacity to transform customer-centric decision-making.

6.1 Dataset Information

The dataset "Amazon US Customer Reviews Dataset" is structured to support analyses related to product reviews, encompassing various aspects of e-commerce feedback. It contains a total of 11 attributes, which likely include fields such as product name, category, review text, star (numeric) rating, and demographic information about the reviewers, among others. With 65,873 records, the dataset provides a substantial volume of data, covering reviews for 40 different products. These products are classified into 4 categories, indicating a broad but defined range of goods or services under review.

The dataset also categorizes reviewers into 13 demographic categories. These categories could encompass age groups, gender, geographical locations, or other relevant demographic factors, allowing for detailed sentiment analysis and understanding of preferences across diverse segments of the population.

Given its structure, this dataset is particularly well-suited for sentiment analysis, demographic-based preference studies, and product performance evaluation across different categories. It allows for a nuanced understanding of consumer feedback by integrating quantitative ratings with qualitative review texts, alongside the ability to perform segmented analyses based on demographic information. This can provide insights into which products perform well in certain categories, how different demographic groups perceive various products, and identify areas for improvement or market opportunities within specific segments.

6.2 Comparative Analysis

Table.1 Comparative Analysis

Author, Year	Accuracy (%)
Zaoli Yang et al. [1], 2023	97.325
Mouthami Kuppusamy et al. [2], 2023	97.014
Suntarin Sangsavate et al. [6], 2023	96.990
Pallavi Mishra et al. [10], 2023	96.360
Muddada Murali Krishna et al. [11], 2023	97.580
Proposed Method	99.980

Table.1 represents a comparison of the accuracy of different sentiment analysis methods, as reported by various authors in 2023, alongside a proposed method that claims the highest accuracy. Each entry in the table lists the authors of the study, followed by the year of publication (2023 for all entries), and the accuracy of their sentiment analysis method expressed as a percentage.

Zaoli Yang et al. [1] report an accuracy of 97.325%, indicating a highly effective sentiment analysis approach. Mouthami Kuppusamy et al. [2] have a slightly lower accuracy at 97.014%, which is still remarkably high, showing their method's effectiveness in accurately classifying sentiment. Suntarin Sangsavate et al. [6]

present a method with an accuracy of 96.995%, very close to that of Mouthami Kuppusamy et al., suggesting another competitive approach to sentiment analysis. Pallavi Mishra et al. [10] show an accuracy of 96.36%, which, while slightly lower than the others mentioned so far, still demonstrates a high level of precision in sentiment analysis. Muddada Murali Krishna et al. [11] report an accuracy of 97.58%, which is the highest among the individual studies listed before the proposed method, indicating a particularly effective method for sentiment analysis.

The Proposed Method shows an accuracy of 99.98%, which is significantly higher than the other methods listed. This suggests that the proposed method outperforms existing approaches by a substantial margin, offering a nearly perfect accuracy rate in sentiment analysis. Such a high accuracy percentage indicates an exceptional ability to correctly identify and classify the sentiment expressed in the data it analyzes, potentially setting a new standard for sentiment analysis techniques

CONCLUSIONS: The research outlined significant advancements in sentiment analysis by presenting a Hybrid Regression Model that leverages natural language processing and regression analysis, transcending the limitations of traditional methods. By incorporating demographic-specific information, the model achieves enhanced contextual understanding, allowing for the optimization of organizational strategies with greater accuracy. This model not only demonstrates exceptional precision but also maintains reduced complexity, making it highly practical for analyzing customer feedback and preferences. The study also addresses ethical considerations, emphasizing the importance of mitigating biases and ensuring fairness in sentiment analysis. This commitment to ethical AI highlights the broader impact of the research, laying the groundwork for improved decision-making in customer-centric strategies, product development, and marketing. The Hybrid Model emerges as a crucial asset for businesses aiming to enhance customer experience, foster brand loyalty, and stay competitive in the evolving digital landscape, due to its innovative approach, accuracy, and adaptability.

REFERENCES:

[1] Zaoli Yang, Qin Li, Vincent Charles, Bing Xu, & Shivam Gupta (2023). Online Product Decision Support Using Sentiment Analysis and Fuzzy Cloud-Based Multi-Criteria Model Through Multiple E-Commerce Platforms. IEEE Transactions on Fuzzy Systems.

[2] Mouthami Kuppusamy, & Anandamurugan Selvaraj (2023). A novel hybrid deep learning model for aspect based sentiment analysis. Concurrency and Computation: Practice and Experience.

[3] Mr. Abhale B. A, Miss. Bachhav M. K., & Miss. Patil Y. P. (2022). Fake Reviews Detection Using Multidimensional Representations with Fine-Grained Aspects Plan. International Journal for Research in Applied Science and Engineering Technology.

[4] Mukkamula Venu Gopalachari, Sangeeta Gupta, Salakapuri Rakesh, Dharmana Jayaram, & Pulipati Venkateswara Rao (2023). Aspect-based sentiment analysis on multi-domain reviews through word embedding. Journal of Intelligent Systems.

[5] Mubashar Hussain, Toqir A. Rana, Aksam Iftikhar, M. Usman Ashraf, Muhammad Waseem Iqbal, Ahmed Alshaflut, & Abdullah Alourani (2022). Multi Layered Rule-Based Technique for Explicit Aspect Extraction from Online Reviews. Computers, Materials and Continua.

[6] Suntarin Sangsavate, Sukree Sinthupinyo, & Achara Chandrachai (2023). Experiments of Supervised Learning and Semi-Supervised Learning in Thai Financial News Sentiment: A Comparative Study. ACM Transactions on Asian and Low-Resource Language Information Processing.

[7] S. Uma Maheswari, & S. S. Dhenakaran (2023). Opinion Mining on Integrated Social Networks and E-Commerce Blog. IETE Journal of Research.

[8] Muhammad Ahmad, Kazim Jawad, Muhammad Bux Alvi, & Majdah Alvi (2023). Google Maps Data Analysis of Clothing Brands in South Punjab, Pakistan. EAI Endorsed Transactions on Scalable Information Systems.

[9] K. Anuradha, M. Vamsi Krishna, & Banitamani Mallik (2022). Opinion Mining Using Normal Discriminant Piecewise Regressive (NDPR) Sentiment Classification Technique. Journal of Uncertain Systems.

[10] Pallavi Mishra, & Sandeep Kumar Panda (2023). Dependency Structure-based Rules using Root Node Technique for Explicit Aspect Extraction from Online Reviews. IEEE Access.

[11] Muddada Murali Krishna, Balaganesh Duraisamy, & Jayavani Vankara (2023). Independent component support vector regressive deep learning for sentiment classification. Measurement: Sensors.

[12] Amazon US Customer Reviews Dataset (2021), https://www.kaggle.com/datasets/cynthiarempel/amazon-us-customer-reviews-dataset.