How Trust Signals in Product Reviews Predict Recommendation Behavior: A Behavioral Study Using E-Commerce Data

Michelle Irina Made Enoh *

Posted Date: 6 August 2025

doi: 10.20944/preprints202508.0374.v1

Keywords: text classification; large language models; trust signals; e-commerce; consumer behavior; sentiment analysis; natural language processing

 doi:10.20944/preprints202508.0374.v1

Article

How Trust Signals in Product Reviews Predict Recommendation Behavior: A Behavioral Study Using E-Commerce Data

Michelle Irina Made Enoh

Consumer Behavior and Decision Sciences, Harrisburg University of Science and Technology, Harrisburg, PA 1710 USA; mmade@my.harrisburgu.edu

Abstract

The rise of e-commerce has made online product reviews a cornerstone of the U.S. digital economy, essentially shaping consumer purchasing behavior. However, the integrity of this marketspace is threatened by inauthentic or misleading review content, creating a challenge for consumers who rely on these reviews to make informed decisions. This study investigates how specific language patterns within review texts, termed "trust signals," can predict a consumer's recommendation behavior more effectively than traditional sentiment analysis methods. Firstly, baseline predictive models like Random Forest and XGBoost were developed using numerical data such as product ratings, achieving high accuracy but demonstrating that the models rely heavily on the star rating alone. To address this limitation and analyze the unstructured text, a robust Large Language Model (Mistral-7B) was fine-tuned on a large dataset of e-commerce reviews for a women's clothing retail brand. The fine-tuned model was able to predict recommendation behavior with over 92% accuracy based solely on the review text. Furthermore, an n-gram analysis identified that specific phrases related to product fit ("true size," "runs small") and quality ("soft comfortable," "fabric soft") were some of the most powerful trust signals. These findings demonstrate that advanced machine learning models can be trained to detect complex, context-rich cues in language, providing a more sophisticated and reliable method for understanding the key elements that build consumer trust. This research offers a valuable methodology for enhancing the transparency of the digital marketplace, which is vital for the nation's continued economic stability.

Keywords: text classification; large language models; trust signals; e-commerce; consumer behavior; sentiment analysis; natural language processing

---

E-commerce has changed the game in how consumers shop for products and services. In today's American economy, the online marketplace has become the major cornerstone with e-commerce sales accounting for a huge and ever-growing portion of the nation's retail landscape. According to data from the United States Department of Commerce (2024), e-commerce sales in 2023 alone exceeded one trillion dollars, proving its massive role in modern economic activity. This digital shift has changed how people decide what to buy. Since many physical stores have been closing down in recent years, consumers can no longer try or touch items physically prior to purchasing, thereby leaving them to rely mostly on online reviews as social proof. Moreover, these testimonials from different users have become a main source of information that directly influences what people purchase (Chevalier & Mayzlin, 2006).

Because these reviews are so important, their trustworthiness is a matter of national economic interest. A reliable review system helps everyone by promoting fair competition and empowering shoppers. Therefore, the integrity of this system is under constant threat from fake or biased reviews that can mislead consumers and distort the market (Malbon, 2013). Consequently, the main challenge for both shoppers and sellers are to figure out how to tell a real review recommendation from a fake

 doi:10.20944/preprints202508.0374.v1

one. The challenge is the fact that other methods of analyzing text only focuses on counting keywords thereby failing to understand the context of the review text. For example, a basic tool cannot tell the difference between a negative review which states "I wanted to love this dress, but it runs small" and "I love how this dress runs true to size". From this example, both reviews use the word "love" but they convey opposite outcomes. A positive one and a negative one. This limitation demonstrates the need for more advanced methods that are able to understand language as humans do.

In this study, we address this issue by identifying specific key phrases "trust signals" which predicts whether a consumer will recommend a product. We argue that key phrases related to fit and quality are more reliable predictors than general positive and negative words. In addition, we use a Large Language Model (LLM) on a large e-commerce dataset called Women's Clothing E-Commerce Reviews available on Kaggle to see if they can learn to detect these trust signals accurately. Moreover, a successful model would offer a solid new way to evaluate the actual meaning of online reviews in addition to offering a useful tool for safeguarding consumers and boosting the country's digital economy.

1. Literature Review

This study sits at the intersection of e-commerce, computational language, and consumer psychology. It looks at previous research in five main areas to provide context for this study which are; how online reviews affect consumer recommendations, the evolution of Natural Language Processing (NLP) for text classification, the role of trust in online reviews, the psychology of consumer trust , and the economic impact of fake and unauthentic review. By understanding what is already known allows us to pinpoint the precise knowledge gap that this study attempts to fill.

1.1. The Role of Trust in Online Reviews

Because in online shopping consumers are unable to physically inspect products when they shop, they can only rely on reviews to ascertain an item's quality and how reliable it is. These reviews influence is the foundation of consumer trust. According to research, a review's credibility is influenced by a number of variables, such as the tone of the review (whether positive or negative), the quantity of reviews, and the text's particular content (Flanagin et al., 2014). Since consumers use reviews and average ratings as a quick way to determine a product's popularity and quality, studies have long shown that products with more reviews have higher sales (Duan et al., 2008).

Nevertheless, consumer trust is more complicated than a simple star review. Studies have shown that consumers value the actual content of reviews for how helpful it is towards making a good buying choice. For example, according Mudambi & Schuff (2010), reviews that are more detailed and specific are considered more helpful and are given more weight by shoppers. In addition, also of note is the "negativity bias," which suggests that shoppers may find negative reviews more believable because they seem more critical and honest (Ahluwalia, 2002).

Overall, this shows specific words and terms used in a review are very important. Therefore, the major problem is not just to measure if a review is positive or negative but to identify trust signals in the form of key phrases which build trust and influence consumer's final buying decision.

1.2. From Keywords to Context: The Evolution of Text Classification

Researchers use Natural Language Processing otherwise known as NLP to analyze review text. Initially, the methods employed for this were quite straightforward. For example, methods like the "bag of words" model were only limited to counting the number of times negative or positive keywords appeared in a review (Pang & Lee, 2008). While this approach was good at the start, the main issue was its difficulties distinguishing between a review which says "the fabric is a great deal for a cheap price" because it will only identify the word "cheap" as a negative signal.

However, according to Vaswani et al. (2017), everything changed with a major advancement in Natural Language Processing (NLP) which came with transformer-based models introduced in the

 doi:10.20944/preprints202508.0374.v1

paper "Attention Is All You Need". These new models including BERT were better at reading and understanding the context of sentences (Devlin et al., 2018). More so, this technology has evolved into what is best known today as Large Language Models (LLMs). Because these LLMs have been trained using a vast amount of text data available on the internet, it is better equipped to understand language deeper than other models. As a result, this makes it possible to adapt them for complex tasks like ours with a process called fine-tuning (Brown et al., 2020).

1.3. Recommendation Behavior as a Function of Word-of-Mouth

For the longest time, traditional marketing came with word-of-mouth whereby a consumer physically speaks highly of a product or service to friends and relatives thereby influencing them positively to try them. With the evolution of traditional marketing to digital marketing, and physical stores into e-commerce, coupled with technological advancement, traditional word-of-word transitioned to online word-of-mouth in the form of online reviews. A positive experience is the main reason why a consumer recommends a product. Also, it is clear that online reviews can be translated to a form of modern electronic word-of-mouth (eWOM) which is a powerful driver of consumer behavior and sales (Godes & Mayzlin, 2004). Within a review, when the final recommendation is "I would buy this again,", this is a direct statement of this behavior.

Therefore, it is imperative to understand the link between the review text and the final recommendation as it is of great economic, and commercial interest. This is because by understanding this link can help businesses better understand what drives recommendation thereby helping them improve their products and manage consumer expectations better. This will promote a transparent and fair market ultimately benefiting the national economy.

1.4. The Psychology of Consumer Trust

Consumer's trust is the foundational pillar of every successful profit and non-profit making company. This trust goes far beyond a single transaction resulting to profits but it is the basis for creating a loyal community of consumers who become brand advocates and ambassadors at no cost. For this reason, companies that focus on long term success often heavily invest on building and maintaining brand trust in the mind of their target consumers. Basic marketing principles have long established that it is more cost-effective to retain a loyal customer than it is to acquire a new one (Reichheld & Sasser, 1990).

The psychology of consumer's trust is centered on the brand's ability to consistently deliver on the promise made. Moreover, consumers grant their trust when a product or service delivers on its promise by providing the features, usability and attributes it was advertised to possess. When this promise is met, consumers are more likely to become loyal, repeat consumers. In this heavily e-commerce reliant economy, online review texts have become the main source for new and potential consumers to determine if a brand is true to its claims before making a purchase. This therefore highlights that it is important for review text to be authentic because it is main source of the language signals used in predicting recommendation behavior.

According to the work from Mayer, Davis, and Schoorman (1995), an academic definition of trust is the willingness of a consumer to be vulnerable to the actions of a seller, based on the expectation that the seller will perform a particular action important to the consumer, irrespective of the ability to monitor or control that seller. This definition hinges on three main components namely the willingness to be vulnerable, expectation that the consumer will act beneficially, and the fact that this happens irrespective of the ability to control them.

In essence , for Mayer, Davis, and Schoorman, trust is a conscious decision to rely on another party, accepting the inherent risks because one believes the other party will act in a way that is beneficial or at least not harmful, regardless of the ability to oversee or regulate their behavior. This willingness to accept vulnerability is based on perceptions of the seller's or product's characteristics. Researchers have identified three key dimensions that form the basis of a consumer's trust:

 doi:10.20944/preprints202508.0374.v1

1. **Ability (or Competence):** This refers to the consumer's belief that the brand or seller has the skills and expertise to deliver on its promises. In the context of e-commerce, this often translates to the perceived quality of the product does it function as advertised? Is it well-made? (Gefen, Karahanna, & Straub, 2003).
2. **Benevolence:** This is the consumer's belief that the seller has their best interests at heart and is not solely driven by a profit motive. A benevolent brand is perceived as caring and customer-focused, which can be signaled through good customer service or fair return policies (McKnight, Choudhury, & Kacmar, 2002).
3. **Integrity:** This dimension involves the consumer's perception that the seller adheres to a set of principles that the consumer finds acceptable, such as honesty and fairness. For online reviews, integrity is crucial. Consumers trust reviews that they perceive as honest and unbiased representations of a product experience (Malbon, 2013).

In the absence of direct interaction, consumers look for signals to assess these above-mentioned dimensions. Online reviews have become one of the most powerful sets of signals. A review that provides specific, tangible details about product quality (e.g., "the fabric is soft and durable") directly addresses the ability dimension. Also, a review that mentions a positive customer service experience signals benevolence. Finally, the perceived honesty of the review itself speaks to integrity. Therefore, the "trust signals" identified in this study are not just random phrases, they are the language use of these core psychological dimensions of trust. By analyzing these signals, we can gain a deeper, more refined understanding of the cognitive and emotional processes that drive a consumer's final recommendation behavior.

1.5. The Economic Impact of Fake and Unauthentic Reviews

Even though genuine reviews are detrimental in the digital marketplace, the integrity of this environment is continually being threatened by the presence of fake and unauthentic content. This issue is not minor because it is a significant problem that can disrupt markets and cause consumer's harm. Research has shown that a large percentage of online reviews can be deceptive, ranging from fake positive reviews posted by sellers to promote their own products, to negative smear campaigns intended to harm competitors (Luca & Zervas, 2016). We cannot neglect the economic consequences of this deceit because it is substantial. Fake reviews create information asymmetry, where consumers are led to believe that a product is higher quality that it actually is. Consequently, this leads to misallocated spending, where consumers purchase inferior goods resulting to financial loss and a reduction in overall market efficiency (Malbon, 2013).

Furthermore, the presence of fake reviews goes against the very essence of the digital market system which is trust. When consumers can longer trust the authenticity of online reviews, the value of the whole digital marketplace reduces, and this reduction potentially leads to a decrease in e-commerce activity entirely. This shows that there is a critical need for methods of text review analysis that look beyond star ratings. To conclude, the ability to detect subtle language signals as investigated in this study represents progress towards identifying reviews that are more likely to be authentic and trustworthy, thereby helping to minimize the harmful economic effects of deceptive content.

1.6. The Research Gap: Beyond Sentiment to Predictable Trust Signals

In summary, we have seen from past research that trust is fundamental for online reviews to be impactful. Also, modern Large Language Models have become better at understanding language context, and what people write in reviews directly influences their recommendation behavior. However, a significant research gap still exists. Most Natural Language Processing studies have focused on analyzing a review text classifying it as either positive, negative or neutral without grasping specific words or phrases which predict signals of trust that could lead to a consumer making a recommendation.

 doi:10.20944/preprints202508.0374.v1

This is exactly where this study comes in. Instead of focusing on general sentiment analysis, we investigate deeper into specific phrases which serve as reliable trust signals. In our approach, the main idea was that phrases related to real product features such as fit (true size), quality (fabric soft), appearance (looks great) is a better way to predict recommendation than simply looking at generic words such as good or bad. We tested this by training and fine-tuning a Large Language Model (LLM) to determine whether such a model can learn to accurately detect these subtle signals. In essence, this research uses an advanced model not just to classify reviews, but as a tool to understand the language of trust and how it directly influences consumer behavior.

2. Methods and Materials

This study was conducted in two main stages using a multi-phase quantitative methods approach to investigate how trust signals in product reviews predict recommendation behavior. The initial stage's focus was to establish a performance baseline using traditional machine learning models with structured data whereas the second stage consisted of a more robust and advanced step consisting of fine-tuning a Large Language Model (LLM) to analyze the unstructured review text. Finally, this entire research was designed using one fixed random seed (42005) for reproducibility for all data splitting and modeling procedures.

2.1. Dataset

The research used the publicly available "Women's Clothing E-Commerce Reviews" dataset, from the Kaggle.com platform. This dataset includes 23,486 rows of anonymous customer reviews for a women's clothing retailer and 10 feature variables. Each row corresponds to a customer review, and includes the variables:

- **Clothing ID:** Integer Categorical variable that refers to the specific piece being reviewed.
- **Age:** Positive Integer variable of the reviewer's age.
- **Title:** String variable for the title of the review.
- **Review Text:** String variable for the review body.
- **Rating:** Positive Ordinal Integer variable for the product score granted by the customer from 1 Worst, to 5 Best.
- **Recommended IND:** Binary variable stating where the customer recommends the product where 1 is recommended, 0 is not recommended.
- **Positive Feedback Count:** Positive Integer documenting the number of other customers who found this review positive.
- **Division Name:** Categorical name of the product high level division.
- **Department Name:** Categorical name of the product department name.
- **Class Name:** Categorical name of the product class name.

Data Format and Preprocessing. The dataset was provided in a comma-separated values(.csv) format. Before analysis, rows with missing values in either the Review Text or Recommended IND columns were removed to ensure data integrity. No other transformations were applied at this stage.

Dataset Rationale and Characteristics. This dataset was chosen because it contains real world data of authentic consumer reviews containing genuine language signals of trust and dissatisfaction from an anonymous but existing e-commerce retailer. In addition, the size of this dataset is meaningful enough for this robust analysis while remaining manageable for training advanced models with the available computational resources. The initial exploratory data analysis (EDA) revealed several key characteristics of the dataset. Firstly, a significant class imbalance was observed in the target variable, with over 82% of the reviews having a positive recommendation (see Figure 1). This positivity bias is very common in e-commerce data but presents a challenge for predictive modeling that must be accounted for at evaluation.

 doi:10.20944/preprints202508.0374.v1

Secondly, the analysis of reviewer demographics showed that the majority of contributors were between the ages of 30 and 50, indicating that the findings are most representative of this age bracket (see figure 2). Finally, the dataset is specific to a single department (women's clothing), which makes it an excellent case study but also suggests that its specific verbal patterns may not be directly generalizable to other product categories.

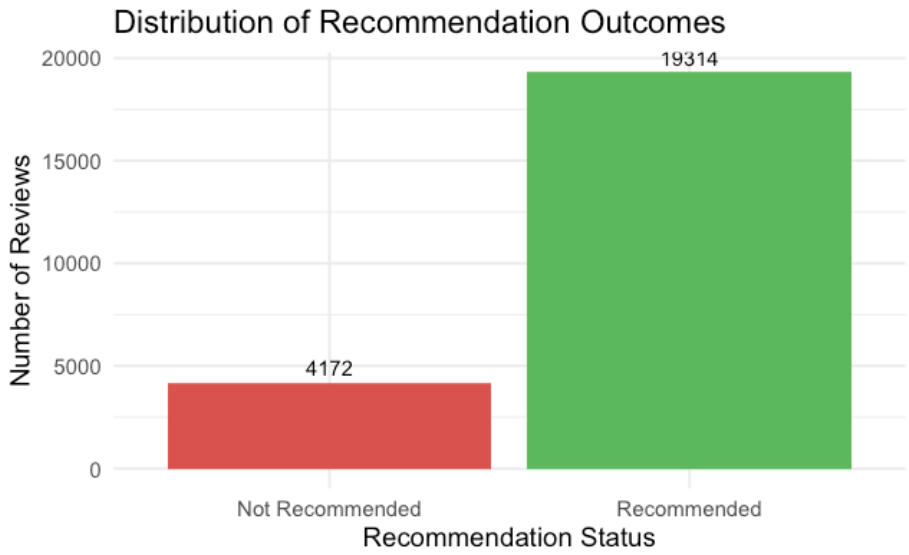


Figure 1. *Distribution of Recommended Outcomes.*

The figure one above clearly shows the obvious imbalance in the dataset. Out of 23,486 reviews, 19,314 (approximately 82.3%) were positive recommendations whereas only 4,172 (approximately 17.7%) were negative therefore not recommended. This distribution is skewed and a critical finding for this study as it confirms the positivity bias often discovered in online reviews data. This proofs that such dataset requires the use of evaluation metrics beyond simple accuracy, such as the F1-score, to properly assess model performance on the minority class.

 doi:10.20944/preprints202508.0374.v1

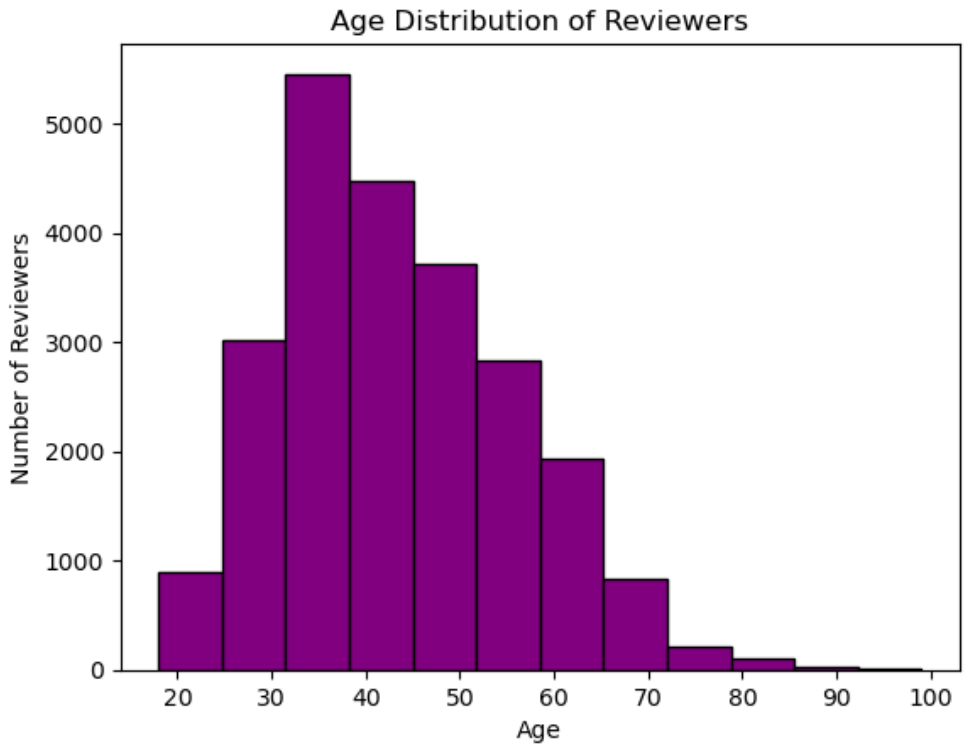


Figure 2. *Age Distribution of Reviewers.*

The demographic analysis of the reviewers, shown in Figure 2, reveals that the data is most representative of consumers between the ages of 30 and 50. This concentration within a specific age bracket is a notable characteristic of the dataset, and suggests that the identified trust signals may be most relevant to this demographic. For this behavioral study, the key variables used were the unstructured Review Text and the binary target variable, Recommended IND, where a value of 1 indicates a positive recommendation and 0 indicates a negative one. Numerical features such as Age, Rating, and Positive Feedback Count were also used in the initial modeling stage.

2.1.1. Stage 1 Baseline Modeling and Traditional Methods

To begin this analysis, three widely used machine learning models including RandomForest, Gradient Boosting, and XGBoost were implemented to establish a baseline for prediction accuracy. At this stage, the main goal was to determine if we could only use the structured, numerical data available in the dataset such as Age, Rating, to predict recommendation behavior effectively.

Prior to model training, a correlation matrix of trust signals was conducted to examine whether any variables were highly correlated in ways that could compromise the model interpretability. As shown in the correlation matrix of trust signals (Appendix A), the relationship among all variables involved, namely Age, Rating, and Positive Feedback were weak, indicating very low multicollinearity. The analysis revealed that each trust signal provided unique, non-overlapping information. This finding is a good sign for multivariate modeling, as it suggests each feature can contribute independently to the model's predictions.

Data Splitting and Preprocessing. For the baseline models, the dataset was split into training set (60%), validation set (20%), and test set (20%). This is to ensure that the models are trained on one portion of the data, tuned on another, and finally evaluated on a completely unseen part of the data to provide an unbiased measure of the performance. In addition, a correlation analysis of the numerical features (Age, Rating, Positive Feedback Count) was conducted, which showed very low correlation between variables, indicating that each feature provided unique information.

Model Implementation. As previously mentioned, three models were chosen for their robustness in classification tasks as follows:

 doi:10.20944/preprints202508.0374.v1

- **Random Forest:** This model was selected because it is an ensemble method that builds multiple decision trees and merges them to get a more accurate and stable prediction. It is also effective at providing clear feature importance scores.
- **Gradient Boost:** This model was chosen for its high predictive power. It works by sequentially adding new models that correct the errors of the previous ones, creating a strong final predictor.
- **XGBoost:** This is a highly efficient and scalable implementation of Gradient Boosting, XGBoost was included for its well-known performance and speed in machine learning competitions and real-world applications.

After training for each of these models was done, a feature importance analysis was performed to identify which numerical features were predictive.

2.1.2. Stage 2 Advanced Text Classification with a Fine-Tuned LLM

The limitation of relying on solely numerical data let to the second stage of this research which focused on the unstructured data within the review text.

Model Selection and Fine-Tuning Environment. The model selected for this task was Mistral-7B, a Large Language Model known for its sophisticated understanding of language and context. To make the training of this large model feasible, we utilized the Unsloth library within a Google Colab environment. We had to use Google Colab Pro to secure enough computational resources to run this model successfully. Unsloth is a specialized tool that optimizes memory usage and significantly speeds up the fine-tuning process, making it possible to train large models on readily available hardware.

Data Preprocessing and Prompt Formatting. Before training, the review text and its corresponding recommendation label (Recommended or Not Recommended) were formatted into a clear, instruction-based prompt. This prompt structure is crucial, as it trains the model to perform a specific task. In this case, this meant classifying the sentiment of a given review. This step transforms the raw text into a structured format suitable for instruction fine-tuning. The data was then tokenized, converting the text prompts into a numerical format that the model can process.

Model Training. The pre-processed and tokenized dataset was used to fine-tune the Mistral-7B model. The training was trained for three epochs, with a learning rate of 5e-5 and an effective batch size of 32 completing in approximately 28 minutes on an NVIDIA A100 GPU. During this process, the model learns to identify the specific language patterns, or "trust signals," within the review text that are predictive of the final recommendation behavior. (See Appendix L)

Evaluation Metrics. The model's performance was evaluated using a standard set of metrics for classification tasks. These included:

- **Accuracy:** The overall percentage of correct predictions.
- **Precision:** The proportion of positive predictions that were actually correct.
- **Recall:** The proportion of actual positives that were correctly identified.
- **F1-Score:** A single score that balances how often the model is correct (precision) with how well it finds all the positive cases (recall).
- **Confusion Matrix:** A table that visualizes the performance of the model by showing the counts of true positives, true negatives, false positives, and false negatives.
- **Qualitative Analysis:** In addition to these quantitative metrics, a qualitative analysis of specific review examples was conducted to provide a deeper, more nuanced understanding of the model's predictive behavior.

For transparency and reproducibility, data availability, code availability and the computation tools and ai assistance used in this study are provided at the end of this paper.

 doi:10.20944/preprints202508.0374.v1

3. Results

This research's results are presented in two phases, mirroring the modeling procedure used. The first section establishes a baseline using traditional machine learning models on numerical features. Meanwhile, the second section details the performance of the fine-tuned Large Language Model (LLM) on the unstructured review text.

3.1. Phase 1 Baseline Model Performance

To establish a performance baseline, three traditional machine learning models (Random Forest, Gradient Boosting, and XGBoost) were trained to predict the "Recommended IND" using only the numerical features such as Age, Rating, and Positive Feedback Count. All three models achieved a high accuracy of approximately 94% on the validation set. (see Appendices H, I, and J). Prior to training, a correlation analysis was conducted on the numerical predictors. As shown in Appendix A, the matrix revealed very low correlation between all variables, confirming that each feature provided unique information for the models.

A feature importance analysis was then conducted for these models, and the results were consistent across all three. As shown in Table 1, The Rating feature was hugely the most important predictor, accounting for over 86% of the decision-making weight in the Random Forest model and over 98% in the boosting models. Age and Positive Feedback Count on the other hand had a minimal impact. This indicates that while traditional models are effective, they rely almost entirely on the product rating and do not capture insights from the review text itself.

Table 1. *Feature Importance Table for Baseline Models.*

Signal	Random Forest	Gradient Boosting	XGBoost
Rating	0.0856	0.9915	0.9891
Age	0.0856	0.0052	0.0058
Positive Feedback	0.0459	0.0033	0.0051

3.2. Phase 2 Fine-Tuned LLM Performance

The second phase focused on analyzing the review text using the fine-tuned Mistral-7B model. This approach aimed to determine if an advanced AI model could predict recommendation behavior based solely on the language of the reviews.

Model Training and Evaluation. The fine-tuning process was stable and successful. The model's training loss consistently decreased over the training epochs, indicating that it was effectively learning the patterns in the data without simply memorizing it. This is illustrated in Figure 3.

 doi:10.20944/preprints202508.0374.v1

Figure 3. *Model Training Loss Curve.*

The training loss curve for the fine-tuned *mistral 7B* model shows a consistent and stable decline over the course of the 432 training steps. Also, the curve begins at an initial loss of approximately 2.38. Early in the training, there is a sharp decrease in loss indicating that the model is learning at the beginning. This loss later continues to decrease slowly suggesting that the model is refining its internal representations. Furthermore, at approximately step 200 to 420, the curve plateaus at a loss value of approximately 1.33 to 1.40 showing the training is almost convergent. This indicates that the model successfully learned from the training data without significant overfitting.

Upon completion of training, the model was evaluated on the validation set of 500 reviews. The model achieved a high overall accuracy of 92.2%. While slightly lower than the baseline models' accuracy, it is important to note that this result was achieved using only the complex, unstructured text data, without access to the powerful Rating feature.

The detailed performance metrics, which account for the dataset's class imbalance, are presented in Table 2. The model performed exceptionally well on the majority "Recommended" class (F1-Score of 0.95) and achieved a strong F1-Score of 0.77 for the more challenging minority "Not Recommended" class.

Table 2. *Classification Report for Fine-Tuned LLM.*

Class	Precision	Recall	F1-Score	Support (N)
Recommended	0.94	0.96	0.95	411
Not Recommended	0.80	0.74	0.77	89
Weighted Avg	0.92	0.92	0.92	500

A more detailed breakdown of these predictions is illustrated in Figure 4, which presents the confusion matrix.

 doi:10.20944/preprints202508.0374.v1

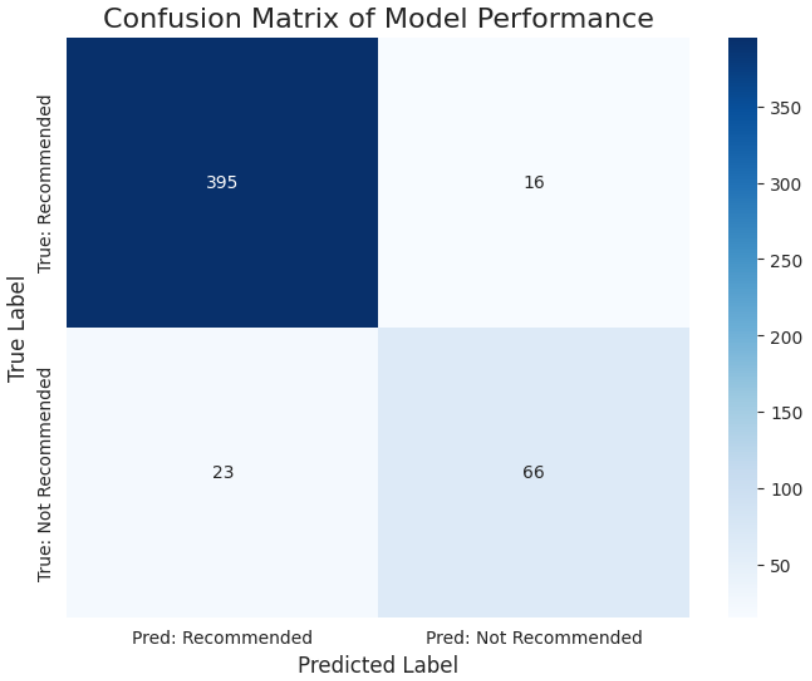


Figure 4. *Confusion Matrix of LLM Performance.*

The matrix shows that the model correctly identified 401 of the 411 "Recommended" reviews, and 66 of the 89 "Not Recommended" reviews. The primary source of error was in misclassifying a small number of "Not Recommended" reviews as "Recommended" (23 instances), while being highly reliable in identifying positive reviews (only 10 errors).

Identification of Language Trust Signals. To understand the key phrases and language patterns the model learned, an n-gram analysis was conducted on the dataset to identify the most commonly used words in each recommendation class. These phrases represent the "trust signals" that the model learned to identify and associate with each class. Figure 5 shows clearly illustrates the distinct patterns revealed in this analysis.

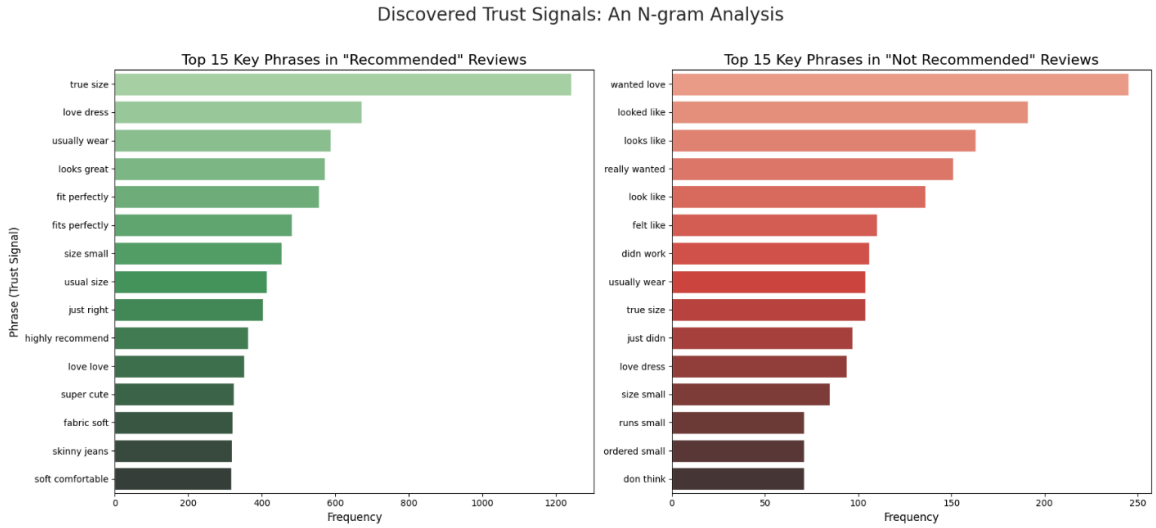


Figure 5. *Discovered Trust Signals from N-gram Analysis.*

The most common phrases in "**Recommended**" reviews were extremely positive and related to physical product attributes. Key phrases included "true size", "fits perfectly", "usual size", and "just right", indicating that product fit is a primary driver of a positive recommendation.

 doi:10.20944/preprints202508.0374.v1

On the other hand, the most frequently used phrases in "**Not Recommended**" reviews were related to failed expectations. The most common phrases were "runs small," "looked like," "felt like", "looks like", "ordered small", and "size small," again highlighting the key importance of fit. Other key negative phrases included "don think", and " didn work", (as written in the original review) signaling poor product quality. These results provide direct, data-driven evidence of the key language signals that correlate with recommendation behavior.

Qualitative Analysis of Model Predictions. To provide a deeper and refined understanding of this model's performance, a qualitative analysis of specific review examples was conducted. This analysis highlights the model's ability to understand context and identifies areas where it still faces challenges. Table 3 presents a selection of these examples.

Table 3. *Qualitative Analysis of Sample Model Predictions.*

Review Text	True Label	Model Prediction	Analysis
"Very quick delivery... one thing is the nice little button in the back fell off... other than that i love the print and everything about the dress. true to size."	Recommended	Recommended	**Correct.** The model correctly identified the strong positive trust signal "true to size" and was not misled by the minor negative detail about the button.
"Got this in the sky color. returned it because the top is way too sheer... for the price, material just felt flimsy & cheap. it does run small..."	Not Recommended	Not Recommended	**Correct.** The model correctly focused on the key negative trust signals "cheap" and "run small," ignoring less important details like the color.
"Loved the material and the style. unfortunately, my fears came true... there wasn't enough support... i am sad i won't be keeping this though. the material is heavenly!!!"	Not Recommended	Not Recommended	**Correct (Nuanced).** This is a key example of the model's sophistication. Despite the presence of strong positive words like "Loved" and "heavenly," it correctly identified that the core functional issue ("not enough support") was the deciding factor, leading to a negative recommendation.
"I've rarely met a maeve item i didnt like... i tried on the 6 and it was a great fit except for the chest area... and the butt/hips area, making	Not Recommended	Recommended	**Incorrect.** This example highlights a limitation. The review contains complex and mixed signals about fit. The model likely over-weighted the positive phrase "great fit"

 doi:10.20944/preprints202508.0374.v1

Review Text	True Label	Model Prediction	Analysis
this dress a little less shift and a little baggy in the abdominal region..."			and was unable to correctly interpret the more nuanced negative details, leading to an incorrect prediction.

Note: The table shows review texts as written by customers and sampled from the held-out validation set. The "Analysis" column provides an interpretation of the model's predictive decision for each case.

4. Discussion

The purpose of this research was to determine how trust signals in product reviews predict recommendation behavior. The finding of this research demonstrates that while high predictive accuracy can be achieved by relying solely on numerical ratings, a fine-tuned large language model (LLM) on the other hand, can achieve a similar performance by using only the unstructured text of consumer reviews. This suggests that the detailed text or words consumers choose to describe their experiences independent of their star ratings carry meaningful signals about their satisfaction and likelihood to recommend. In other words, the linguistic details embedded in review text can serve as reliable indicators of consumer sentiment and recommendation intent.

4.1. Interpretation of Results

The main outcome of this research is that certain specific key phrases are greatly predictive of consumer's definitive recommendation. The n-gram analysis identified that phrases related to product fit such as "true size", "size small", "just right", "usual size", "fit perfectly" , "fits perfectly" and perceived quality such as "fabric soft", "soft comfortable", are the most prevalent signals. This is a very important insight which shows that consumer's trust and associated recommendation behavior are not just solely based on unclear or vague positive or negative emotions. On the contrary, they are closely linked to whether the product fulfills a set of specific, verifiable expectations. When a product is "true to size", it fulfills a crucial expectation, fosters confidence, builds trust, resulting in a recommendation. When it "runs small," that trust is betrayed, thereby failing to meet expectations and leading to a negative outcome.

The success of the fine-tuned Mistral-7B model, with 92.2% accuracy on textual analysis, illustrates that modern AI can transcend mere keyword counting to comprehend the contextual significance of these trust signals. For instance, the model must learn to differentiate between a review which states that "I ordered my true size but it was too tight," and one which says, "This sweater is true to size and fits perfectly". The model successfully decodes the language of customer trust, as evidenced by its high performance, which demonstrates its ability to make these complex distinctions.

4.2. Qualitative Analysis Insights

Although the model's high accuracy is clear from the quantitative results, the qualitative analysis offers deeper insight into its strengths and limitations. The analysis of sample reviews as presented in Table 3 of the results section unveils several key insights. The sample reviews in Table 3 highlight how the model interprets both overt and subtle trust signals when making predictions. To begin, the model showed its exceptional ability to identify the main sentiment of a review, even when conflicting signals were present. For instance, in a positive review that mentioned a minor flaw such as a button falling off, the model correctly prioritized the positive trust signal "true to size" to make a correct "Recommended" prediction.

The most interesting part is that the model proved its ability to understand context in reviews with mixed sentiment. This can be noticed in a case where a review text that contained strong positive words such as "Loved" and "heavenly" but eventually came to the conclusion that the product was

 doi:10.20944/preprints202508.0374.v1

"Not Recommended" because of lack of support. In this case, a simple keyword-based model would have probably failed here. However, the fine-tuned LLM accurately identified that the functional issue was the deciding factor, showcasing a sophisticated human-like understanding of the text.

Finally, even though the qualitative analysis confirmed the impressiveness of the model's abilities, it also highlighted its limitations. In one incorrect prediction, the model was faced with a very complex review text which contained contradictory statements about fit. The model classified incorrectly a "Not Recommended" review as "Recommended" probably because it over weighted a positive phrase "great fit" and was unable to correctly interpret the subtler negative details. This shows that though highly effective, the model can still be challenged with language that is unclear or complex.

4.3. Practical Implications and National Interest

The findings of this research have important practical implications for maintaining the integrity of the American e-commerce industry, which is clearly of national interest and importance.

Enhanced Consumer Protection. This study's methodology establishes a path for the development of more sophisticated tools that safeguard consumers from inauthentic or misleading reviews. By focusing on real trust signals rather than merely sentiment, such technologies could more effectively identify reviews that lack particular, reliable details, allowing consumers to make safer and more informed purchasing decisions. This directly improves the economic well-being of the American people.

Improving Business Practices and Fair Competition. Many businesses can make use of this research to better understand the important factors that influence consumer's satisfaction in the digital e-commerce environment. The findings of this study indicate that getting "fit" and "quality" right is essential to consumers. Moreover, by analyzing review text for these specific trust signals, companies can gain practical insights to improve their products and significantly reduce costly returns. This not only protects their profit margins, but most importantly, preserves valuable consumer trust.

Strengthening the Digital Economy. Consumer trust is the foundation of the U.S. digital economy because it is essential for customer retention. This perfectly aligns with a core principle of modern marketing, which states it is more costly to attract a new customer than to keep the current one satisfied (Kotler & Keller, 2016). Therefore, to maintain such trust in today's highly digitalized e-commerce age, it is important to create methods which can automatically assess the predictive value of online review text, and also determine the authenticity of user generated content. This research introduces such a methodology. Furthermore, this research study is a significant step towards reliability and transparency with regards to our digital market which is essential for a continuous economic growth and stability.

Limitations. It is important to acknowledge the limitations of this study. First, the dataset used is from a single domain (women's clothing). Therefore, the particular trust signals mentioned here especially the fit related ones might not be applicable to other product categories like electronics or books. Testing this methodology on a more diverse range of product reviews in order to find more universal trust signals could be one way to find a solution. Secondly, the qualitative analysis highlighted that the model can still be challenged by highly complex or contradictory language. While the model is very accurate, this shows there is still room for improvement in handling the most nuanced cases of human expression.

Future Research. The findings of this study open several promising paths for future research. One potential study would be to apply this fine-tuning methodology to datasets from different product domains. This would help determine whether the key trust signals are universal or if they are specific to certain product types. For example, are reviews for electronics driven by signals related to "battery life" and "processing speed" in the same way that clothing reviews are driven by "fit"?

Another important future study would be to train the model to detect more subtle linguistic cues, such as sarcasm or suspected inauthentic (fake) reviews. By fine-tuning the model on a dataset

 doi:10.20944/preprints202508.0374.v1

labeled for these characteristics, it could be possible to create an even more robust tool for assessing the true trustworthiness of online content, which would be a significant extension of the current work.

5. Conclusions

This study aimed to investigate how trust signals in product reviews predict recommendation behavior which is an essential factor in maintaining consumer's trust in today's United States' Digital E-commerce Space. Most importantly, this research showed that a fine-tuned Large Language Model(LLM) can predict a consumer's recommendation with an accuracy of over 92% by analyzing the unstructured text of reviews from a large dataset of e-commerce text reviews.

Furthermore, this outcome aligns with traditional model approaches that solely rely on numerical ratings. However, it also proves that powerful predictive signals can be derived from the complexity and richness of review text even when the cues are subtle. This study's most important contribution is the discovery and identification of relevant trust signals. These trust signals are specific phrases related to product fit and quality that predict if a customer will recommend an item. This approach is more advanced than traditional sentiment analysis because it demonstrates that artificial intelligence can be trained to detect nuanced context rich cues embedded in consumer language instead of just overall negative or positive sentiment. Crucially, a qualitative analysis of specific predictions confirmed the model's sophistication, demonstrating its ability to accurately interpret mixed sentiment reviews where keyword-based methods would probably fail. Therefore, this ultimately provides a more reliable method for understanding consumer behavior.

In conclusion, this study adds significant value by providing a new methodology for analyzing user-generated content that has direct applications for consumer protection and the enhancement of the digital marketplace. By moving beyond simple ratings to understand the language of trust, this work contributes to a more transparent and reliable e-commerce ecosystem, which is of substantial interest to the economic health and stability of the nation.

Specifically, the tool was used for:

1. **Code Generation and Debugging:** Assisting in writing and troubleshooting Python code for the fine-tuning of the Mistral-7B model using the Unsloth and Hugging Face Transformers libraries. It also supported the creation of data visualizations using matplotlib and seaborn.
2. **Manuscript Drafting and Refinement:** Assisting in the initial drafting, paraphrasing, and editing of the manuscript to enhance clarity, conciseness, and conformity with APA 7th edition style guidelines.

The author independently designed the research methodology, conducted all analyses, interpreted the results, and made all final decisions regarding content and conclusions. The AI tool did not contribute to the formulation of research questions, theoretical framework, or scientific interpretation. The author retains full responsibility for the integrity and accuracy of this work.

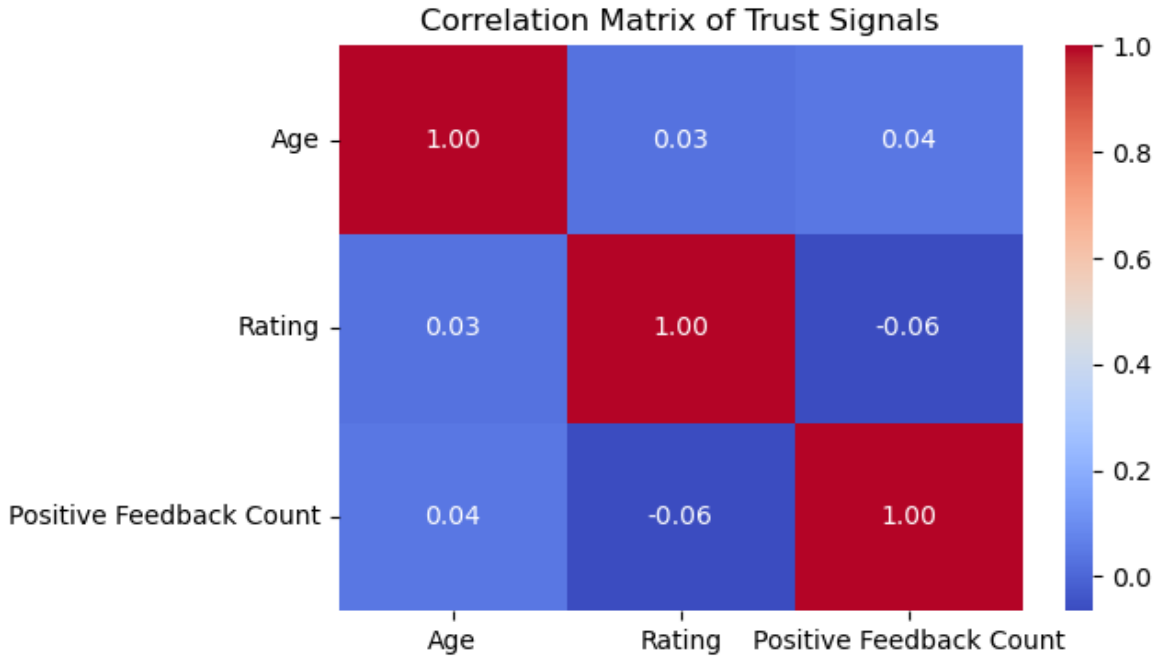
Data Availability Statement: The dataset used in this study is publicly available. The "Women's Clothing E-Commerce Reviews" dataset was sourced from the Kaggle repository and is available at the following URL www.kaggle.com/datasets/nicapotato/womens-ecommerce-clothing-reviews.

Code Availability Statement: The complete code used for data analysis, model training, and fine-tuning in this study is publicly available in a GitHub repository. This repository includes the Google Colab notebook used for the analysis. The repository is also available at the following URL https://github.com/MADEENOH/ANLY699-Thesis-Trust-Signals.git.

Computational Tools and AI Assistance: The author began with a comprehensive exploratory data analysis and developed baseline machine learning models (e.g., Random Forest, XGBoost) to address the research question. Building on this foundation, a generative artificial intelligence (AI) tool Google's Gemini was employed to assist with advanced model fine-tuning and manuscript preparation. The AI's function was strictly limited to programming and writing support, without influencing the research design or scientific conclusions.

 doi:10.20944/preprints202508.0374.v1

Appendix A



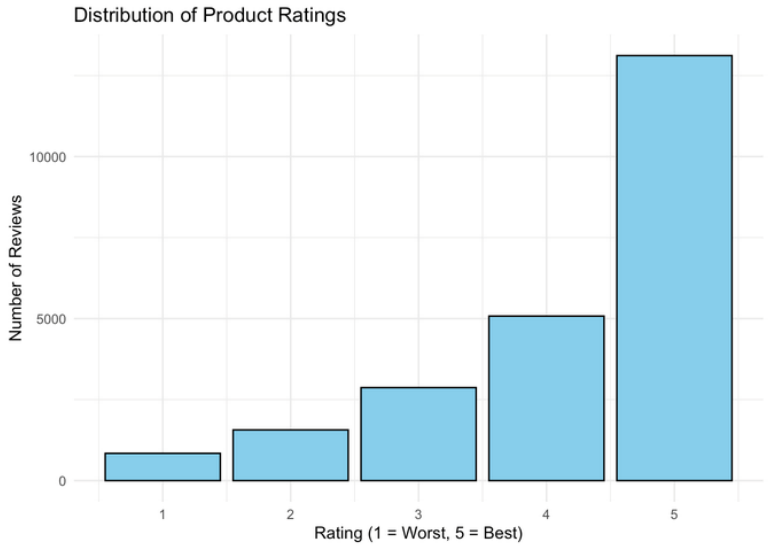
Appendix B

Rating Distribution

Plot 1: Rating Distribution

```
library(ggplot2)

ggplot(df_clean, aes(x = Rating)) +
  geom_bar(fill = "skyblue", color = "black") +
  labs(title = "Distribution of Product Ratings",
       x = "Rating (1 = Worst, 5 = Best)",
       y = "Number of Reviews") +
  theme_minimal()
```



The plot shows that Most customers gave a 5-star rating, followed by 4 and 3 stars. Very few gave 1 or 2 stars.

There could be a positivity bias where satisfied customers are more likely to leave reviews

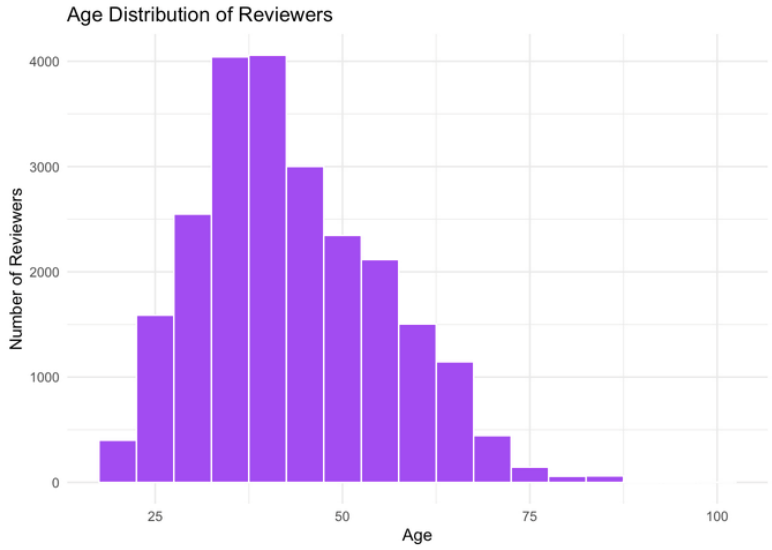
Appendix C

 doi:10.20944/preprints202508.0374.v1

Age Distribution

Plot 2 Age Distribution

```
ggplot(df_clean, aes(x = Age)) +
  geom_histogram(binwidth = 5, fill = "purple", color = "white") +
  labs(title = "Age Distribution of Reviewers",
       x = "Age",
       y = "Number of Reviewers") +
  theme_minimal()
```



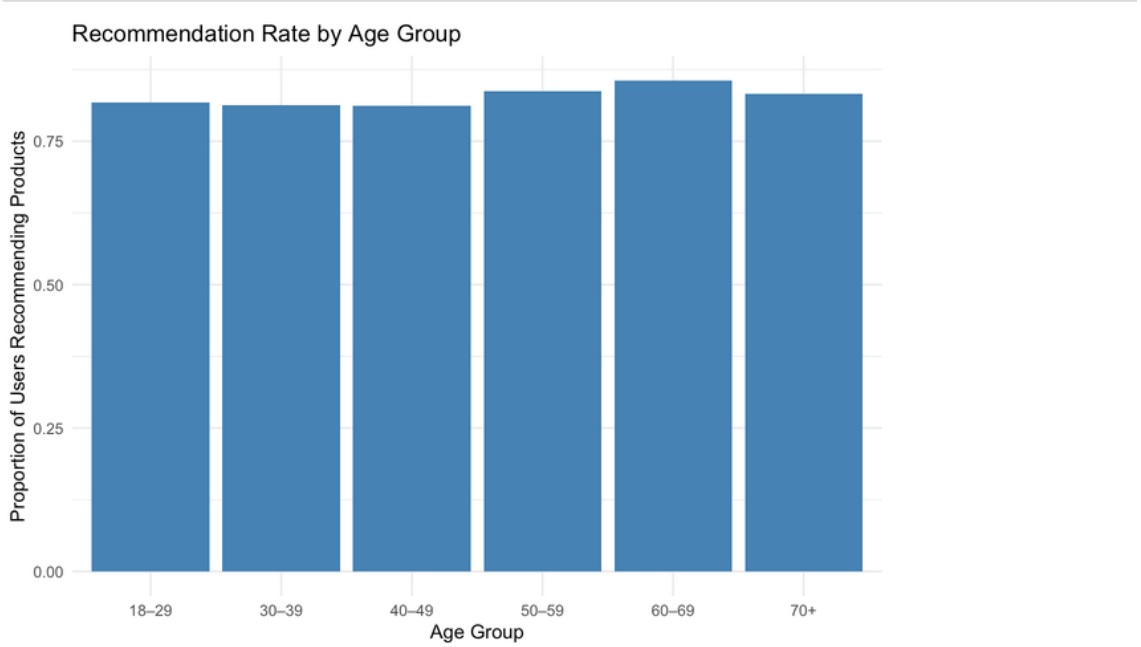
The plot shows the following:

- The most active reviewers are in their 30s to early 40s.
- There's a steep drop-off in review frequency past age 60.
- Very few reviewers are under 25 or over 75.

Appendix D

 doi:10.20944/preprints202508.0374.v1

Recommendation Rate by Age Group

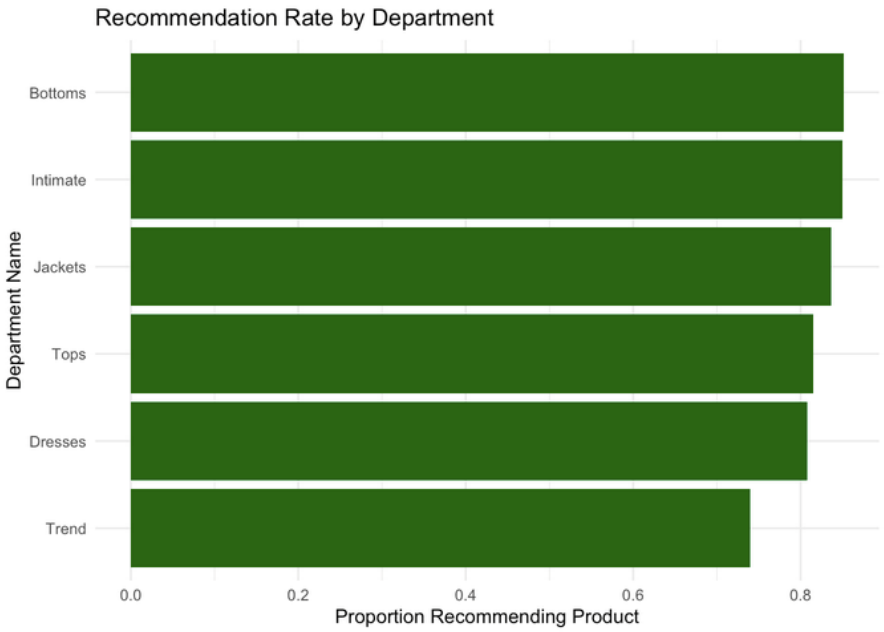


- All age groups are highly likely to recommend products (rates ≈ 80–85%).
- Consumers aged 50–69 show a slightly higher tendency to recommend, possibly reflecting more confidence or satisfaction.
- The "NA" bin is successfully removed as it was there previously

Appendix E

Recommendation Rate by Department

 doi:10.20944/preprints202508.0374.v1

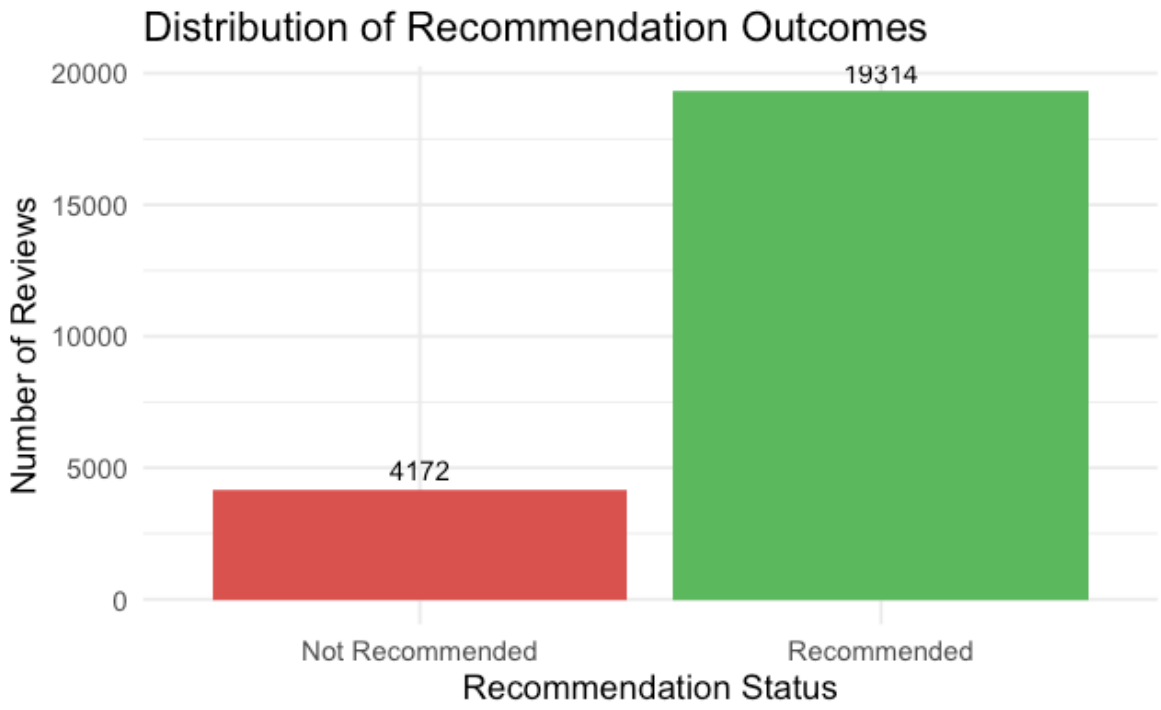


- "Bottoms" and "Intimate" lead in consumer trust — highest recommendation rates.
- "Trend" department lags slightly, possibly indicating more experimental or fashion-forward items that divide opinion.

Overall, all categories are above 75%, but this breakdown gives us a marketing-level lens on which departments drive satisfaction.

Appendix F

Distribution of Recommendation Outcomes



Appendix G

Logistic Regression Model

 doi:10.20944/preprints202508.0374.v1

Step 3: Train First Model — Logistic Regression

```
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import classification_report, confusion_matrix

# Create and train the model
logit = LogisticRegression()
logit.fit(X_train, y_train)

# Predict on validation set
y_val_pred = logit.predict(X_val)

# Evaluate performance
print("Logistic Regression Evaluation (Validation Set):")
print(confusion_matrix(y_val, y_val_pred))
print(classification_report(y_val, y_val_pred))
```

```
Logistic Regression Evaluation (Validation Set):
[[ 740  103]
 [ 192 3662]]
              precision    recall  f1-score   support

           0       0.79      0.88      0.83       843
           1       0.97      0.95      0.96      3854

    accuracy                           0.94      4697
   macro avg       0.88      0.91      0.90      4697
weighted avg       0.94      0.94      0.94      4697
```

Logistic Regression Summary

Model: Logistic Regression

Target: `Recommended IND` (1 = yes, 0 = no)

Features: `Age`, `Rating`, `Positive Feedback Count`

- **Accuracy:** 94% (Validation set)
- **Strength:** Excellent at predicting recommendations (Class 1)

Appendix H

Random Forest Model

 doi:10.20944/preprints202508.0374.v1

Step 4: Random Forest Classifier

```
]: from sklearn.ensemble import RandomForestClassifier

# Train the model
rf = RandomForestClassifier(n_estimators=100, random_state=42005)
rf.fit(X_train, y_train)

# Predict on validation set
y_rf_val = rf.predict(X_val)

# Evaluate performance
from sklearn.metrics import classification_report, confusion_matrix
print("Random Forest Evaluation (Validation Set):")
print(confusion_matrix(y_val, y_rf_val))
print(classification_report(y_val, y_rf_val))

# Feature importance
import pandas as pd
rf_importance = pd.Series(rf.feature_importances_, index=X.columns)
rf_importance = rf_importance.sort_values(ascending=False)
print("Feature Importances:")
print(rf_importance)
```

```
Random Forest Evaluation (Validation Set):
[[ 711  132]
 [ 159 3695]]
              precision    recall  f1-score   support

           0       0.82      0.84      0.83       843
           1       0.97      0.96      0.96      3854

    accuracy                           0.94      4697
   macro avg       0.89      0.90      0.90      4697
weighted avg       0.94      0.94      0.94      4697

Feature Importances:
Rating                      0.868524
Age                         0.085623
Positive Feedback Count     0.045853
dtype: float64
```

Appendix I

Gradient Boost Model

 doi:10.20944/preprints202508.0374.v1

Step 5: Gradient Boosting Classifier

```
from sklearn.ensemble import GradientBoostingClassifier

# Train the model
gb = GradientBoostingClassifier(n_estimators=100, random_state=42005)
gb.fit(X_train, y_train)

# Predict on validation set
y_gb_val = gb.predict(X_val)

# Evaluate performance
print("Gradient Boosting Evaluation (Validation Set):")
print(confusion_matrix(y_val, y_gb_val))
print(classification_report(y_val, y_gb_val))

# Feature importance
gb_importance = pd.Series(gb.feature_importances_, index=X.columns).sort_values(ascending=False)
print("Feature Importances:")
print(gb_importance)
```

```
Gradient Boosting Evaluation (Validation Set):
[[ 761   82]
 [ 212 3642]]
              precision    recall  f1-score   support

           0       0.78      0.90      0.84       843
           1       0.98      0.94      0.96      3854

    accuracy                           0.94      4697
   macro avg       0.88      0.92      0.90      4697
weighted avg       0.94      0.94      0.94      4697

Feature Importances:
Rating                     0.991473
Age                        0.005203
Positive Feedback Count    0.003324
dtype: float64
```

Appendix J

XGBoost Model

 doi:10.20944/preprints202508.0374.v1


```
import xgboost as xgb
from xgboost import XGBClassifier

# Train the model
xgb_model = XGBClassifier(use_label_encoder=False, eval_metric='logloss', random_state=42005)
xgb_model.fit(X_train, y_train)

# Predict on validation set
y_xgb_val = xgb_model.predict(X_val)

# Evaluate performance
print("XGBoost Evaluation (Validation Set):")
print(confusion_matrix(y_val, y_xgb_val))
print(classification_report(y_val, y_xgb_val))

# Feature importance
xgb_importance = pd.Series(xgb_model.feature_importances_, index=X.columns).sort_values(ascending=False)
print("Feature Importances:")
print(xgb_importance)
```

```
/Users/michelleirinamadeenoh/anaconda3/lib/python3.10/site-packages/xgboost/training.py:183: UserWarning: [
oost/xgboost/src/learner.cc:738:
Parameters: { "use_label_encoder" } are not used.

  bst.update(dtrain, iteration=i, fobj=obj)
XGBoost Evaluation (Validation Set):
[[ 727  116]
 [ 178 3676]]
              precision    recall  f1-score   support

           0       0.80      0.86      0.83       843
           1       0.97      0.95      0.96      3854

    accuracy                           0.94      4697
   macro avg       0.89      0.91      0.90      4697
weighted avg       0.94      0.94      0.94      4697

Feature Importances:
Rating                      0.989135
Age                         0.005757
Positive Feedback Count     0.005108
dtype: float32
```

Appendix K

Feature importance Comparison Table

Step 7: Feature Importance Comparison Table

```
# Combine all feature importance scores into one table
importance_df = pd.DataFrame({
    'Logistic Coef': logit.coef_[0],
    'Random Forest': rf.feature_importances_,
    'Gradient Boosting': gb.feature_importances_,
    'XGBoost': xgb_model.feature_importances_
}, index=X.columns)

# Display table
importance_df = importance_df.round(4).sort_values(by='Random Forest', ascending=False)
importance_df
```

	Logistic Coef	Random Forest	Gradient Boosting	XGBoost
Rating	3.1859	0.8685	0.9915	0.9891
Age	0.0130	0.0856	0.0052	0.0058
Positive Feedback Count	-0.0111	0.0459	0.0033	0.0051

Feature Importance Summary

This table compares how much each model relies on the three trust signals: Rating, Age, and Positive Feedback Count.

Signal	Most Important Model
Rating	All models — highest weight across the board
Age	More important in Random Forest
Positive Feedback Count	Slight influence in Random Forest and XGBoost

Takeaway:
Rating is consistently the most trusted signal for predicting recommendations. Age and feedback count had minor impact, with Random Forest using them slightly more than the boosting models.

 doi:10.20944/preprints202508.0374.v1

Appendix L

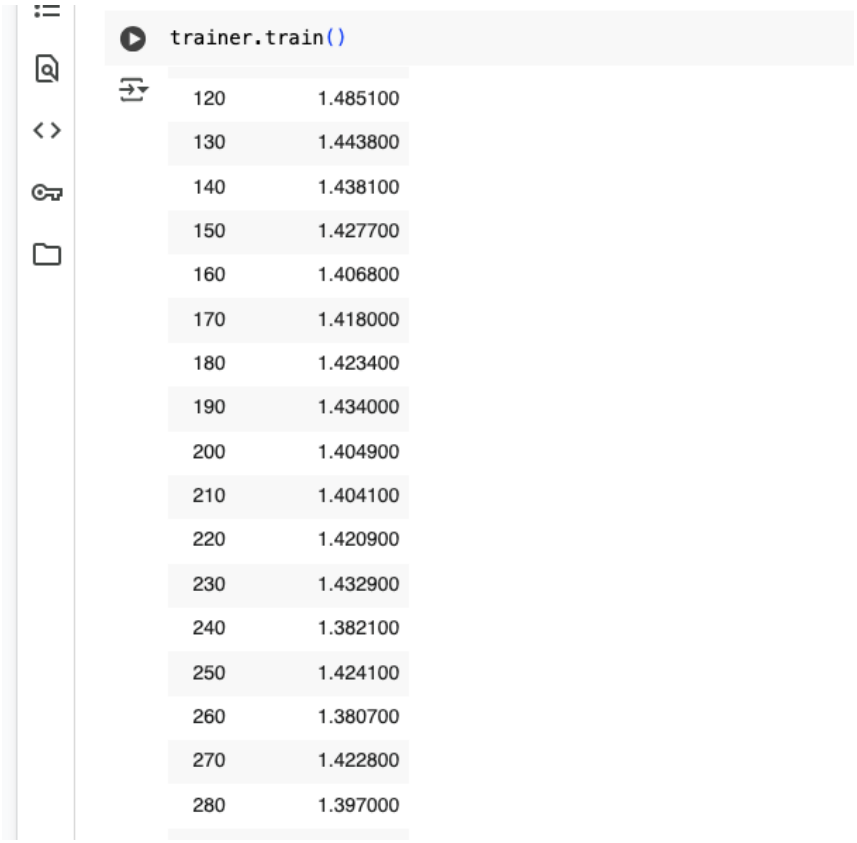
Trainer.train

```
trainer.train()
```

```
==((====))==  Unsloth - 2x faster free finetuning | Num GPUs used = 1
   \\   /|    Num examples = 4,500 | Num Epochs = 3 | Total steps = 423
O^O/ \_/ \    Batch size per device = 8 | Gradient accumulation steps = 4
\        /    Data Parallel GPUs = 1 | Total batch size (8 x 4 x 1) = 32
 "-____-"     Trainable parameters = 41,943,040/7,000,000,000 (0.60% trained)
Unsloth: Will smartly offload gradients to save VRAM!
```

[423/423 28:12, Epoch 3/3]

Step	Training Loss
10	2.373000
20	1.676900
30	1.564400
40	1.583600
50	1.523100
60	1.526400
70	1.480800
80	1.480800
90	1.465900
100	1.459300
110	1.473300
120	1.485100



 doi:10.20944/preprints202508.0374.v1

280	1.397000
290	1.379400
300	1.359500
310	1.357600
320	1.363200
330	1.381300
340	1.342000
350	1.359100
360	1.357700
370	1.357600
380	1.328900
390	1.354000
400	1.337300
410	1.360300
420	1.332500

```
TrainOutput(global_step=423, training_loss=1.445372777627715, metrics={'train_runtime': 1705.266, 'tr
'train_steps_per_second': 0.248, 'total_flos': 1.0117185473396736e+17, 'train_loss': 1.44537277762771
```

References

Ahluwalia, R. (2002). How prevailing attitudes and prior commitment moderate the effects of negative information on evaluations. *Journal of Consumer Research*, *29*(3), 433-440. https://doi.org/10.1086/344421

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877-1901.

Chevalier, J. A., & Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research*, *43*(3), 345–354. https://doi.org/10.1509/jmkr.43.3.345

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Duan, W., Gu, B., & Whinston, A. B. (2008). The dynamics of online word-of-mouth and product sales-An empirical investigation of the movie industry. *Journal of Retailing*, *84*(2), 233-242. https://doi.org/10.1016/j.jretai.2008.04.005

Flanagin, A. J., Metzger, M. J., Pure, R., Markov, A., & Hartsell, E. (2014). Mitigating risk in ecommerce: The role of seller and product characteristics and prior buyer feedback. *Human Communication Research*, *40*(3), 401-429. https://doi.org/10.1111/hcre.12029

Gefen, D., Karahanna, E., & Straub, D. W. (2003). Trust and TAM in online shopping: An integrated model. *MIS Quarterly*, *27*(1), 51-90. https://doi.org/10.2307/30036519

Godes, D., & Mayzlin, D. (2004). Using online conversations to study word-of-mouth communication. *Marketing Science*, *23*(4), 545-560. https://doi.org/10.1287/mksc.1040.0071

Kotler, P., & Keller, K. L. (2016). *Marketing management* (15th ed.). Pearson Education

Luca, M., & Zervas, G. (2016). Fake it till you make it: Reputation, competition, and Yelp review fraud. *Management Science*, *62*(12), 3412-3427. https://doi.org/10.1287/mnsc.2015.2374

Malbon, J. (2013). Taking fake online consumer reviews seriously. *Journal of Consumer Policy*, *36*(2), 139–157. https://doi.org/10.1007/s10603-012-9216-7

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review*, *20*(3), 709-734. https://doi.org/10.5465/amr.1995.9508080335

 doi:10.20944/preprints202508.0374.v1

McKnight, D. H., Choudhury, V., & Kacmar, C. (2002). Developing and validating trust measures for e-commerce: An integrative typology. *Information Systems Research*, *13*(3), 334-359. https://doi.org/10.1287/isre.13.3.334.81

Mudambi, S. M., & Schuff, D. (2010). What makes a helpful online review? A study of customer reviews on Amazon.com. *MIS Quarterly*, *34*(1), 185-200. https://doi.org/10.2307/20721420

Nicapotato. (2018). *Women's E-Commerce Clothing Reviews*. Kaggle. Retrieved July 24, 2025, from https://www.kaggle.com/datasets/nicapotato/womens-ecommerce-clothing-reviews

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, *2*(1–2), 1-135. https://doi.org/10.1561/1500000011

Reichheld, F. F., & Sasser, W. E. (1990). *Zero defections: Quality comes to services*. Harvard Business Review, *68*(5), 105–111.

U.S. Department of Commerce. (2024). *Quarterly retail e-commerce sales, 1st quarter 2024*. U.S. Census Bureau. https://www.census.gov/retail/mrts/www/data/pdf/ec_current.pdf

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.