Analyzing Unstructured Data

Junye Fan, Cheng-Yuan Wu, Sara Antentas, Marina Silva

Github URL: `https://github.com/rsm-juf007/MGTA415`

1. Introduction

ModCloth, an American online retailer established in 2002, specializes in vintage-inspired and indie-style clothing, accessories, and home decor. Initially launched to sell unique vintage pieces, the brand later expanded its offerings to include its own designed and manufactured products, aiming to reach a broader audience.

The dataset used in this analysis, sourced from *Decomposing Fit Semantics for Product Size Recommendation in Metric Spaces*[1]., spans from 2013 to 2018, covering over 200,000 entries. Key fields in the dataset include `user_id`, `item_id`, `review_text`, and `rating`. The dataset provides a rich set of customer feedback, making it a valuable resource for understanding customer preferences and developing personalized recommendation systems.

Motivated by an interest in ModCloth, we conducted an analysis of the clothing feedback dataset to explore customer sentiment and build predictive models based on cosine similarity for tailored recommendations. Such a recommendation system could help ModCloth and similar companies improve their user experience by better understanding customer needs, enhancing product relevance, and ultimately boosting customer satisfaction and loyalty.

2. Literature Review

The integration of sentiment analysis into recommender systems has caught significant attention in recent research. Traditional recommender systems primarily rely on explicit user ratings and behavioral data to predict user preferences. However, the advent of user-generated content, such as reviews and comments, has opened avenues for enhancing recommendation accuracy through sentiment analysis.

2.1 Sentiment Analysis in Recommender Systems

Sentiment analysis involves extracting subjective information from text, enabling systems to understand users' opinions and emotions. Incorporating sentiment analysis into recommender systems allows for a more nuanced understanding of user preferences beyond numerical ratings.

In the paper *An Approach to Integrating Sentiment Analysis into Recommender Systems*[2] , the authors propose a method to enhance the performance of recommender systems by incorporating sentiment analysis of user-generated reviews. By leveraging hybrid deep learning models, including BERT, CNN, and LSTM, the approach extracts implicit feedback from textual data to complement explicit ratings. The integration of sentiment analysis provides deeper insights into user preferences, categorizing sentiments into very negative, negative, neutral, positive, and very positive, thereby enriching the collaborative filtering process.

The study demonstrates that combining sentiment-based insights with traditional methods effectively addresses challenges like

[1]Rishabh Misra, Mengting Wan, Julian McAuley. *Decomposing Fit Semantics for Product Size Recommendation in Metric Spaces*, RecSys 2018. Full text available at: https://doi.org/10.1145/3240323.3240398.

[2]Cach N. Dang, María N. Moreno-García, Fernando De la Prieta. *An Approach to Integrating Sentiment Analysis into Recommender Systems.* Sensors 2021, 21(16), 5666. Published: 23 August 2021. Available at: https://doi.org/10.3390/s21165666.

data sparsity and cold-start issues, making the recommender systems more reliable and personalized. This approach highlights the value of textual feedback in modern recommendation systems, emphasizing the role of sentiment analysis in enhancing user experience.

2.2 Fashion Recommendation Systems

In the fashion industry, recommendation systems face unique challenges due to the subjective nature of fashion preferences and the rapid evolution of trends. In the paper *Fashion Recommendation Systems, Models and Methods: A Review*[3] , the author indicates that fashion recommendation systems play a crucial role in enhancing personalized shopping experiences, particularly in e-commerce platforms. These systems leverage advanced models and methods such as collaborative filtering, content-based filtering, and hybrid approaches to predict user preferences and provide tailored recommendations. With the growing use of deep learning techniques, recommendation systems have become increasingly capable of handling complex data like user reviews, product descriptions, and visual content. However, challenges specific to the fashion domain, such as capturing dynamic trends, understanding diverse user preferences, and managing constantly changing inventories, remain significant. By integrating additional data sources like social media and employing AI-driven solutions, fashion recommendation systems continue to evolve, offering more accurate and satisfying results for users.

2.3 Comparison with Existing Methods

Our approach distinguishes itself by combining sentiment analysis with user purchase history to compute a comprehensive recommendation score. While previous studies have utilized sentiment analysis to infer user preferences, our method integrates sentiment scores with similarity measures derived from user purchase histories.

Moreover, our model employs a grid search technique to optimize the weights assigned to sentiment and similarity components, ensuring a balanced contribution from each aspect. This optimization process is designed to maximize the model's predictive performance, a strategy that has shown promise in improving recommendation systems.

Overall, our methodology leverages both textual and behavioral data to provide personalized fashion recommendations. This comprehensive approach addresses the multifaceted nature of user preferences in the fashion domain.

3. Statistical Analytics

A detailed exploration of the dataset revealed several interesting patterns. Below is the overall high frequency phrases across all reviews:

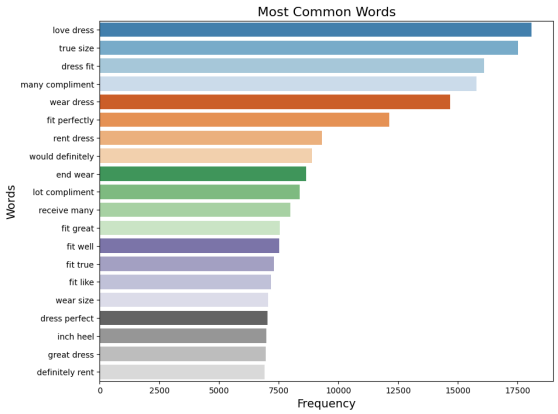

Figure 1: Overall Word Cloud of Reviews

Based on the analysis of the top 20 most frequent phrases from customer reviews, it is evident that the brand's strengths lie in its dresses, fitting accuracy, and ability to generate customer satisfaction. Phrases such as "love dress," "dress fit," and "great dress" highlight that dresses are a key product category that resonates with customers, with a strong emphasis on comfort and style. Words like "true size," "fit perfectly," and "fit well" indicate that accurate sizing and a good fit are critical factors for customer satisfaction.

[3]Samit Chakraborty, Md. Saiful Hoque, Naimur Rahman Jeem, Manik Chandra Biswas, Deepayan Bardhan, and Edgar Lobaton. *Fashion Recommendation Systems, Models and Methods: A Review.* Informatics 2021, 8(3), 49. Published: 26 July 2021. Available at: https://doi.org/10.3390/informatics8030049.

Moreover, terms such as “many compliments” and “receive many” suggest that customers often receive positive feedback while wearing the brand’s products, enhancing their overall experience. The inclusion of phrases like “rent dress” and “would definitely” points to the popularity of the brand in the rental market, particularly for special occasions. Overall, the high frequency of positive words, including “perfectly,” “great,” and “definitely,” reflects a strong sentiment of customer satisfaction, indicating this brand a preferred choice for customers.

Next, we made a word cloud of high-frequency words from reviews in the high ratings and high-frequency words in the low ratings according to the cutoff of a score of 6 to present them more clearly

Figure 2: Low Rating Word Cloud of Reviews

Figure 3: High Rating Word Cloud of Reviews

It is interesting to note that even in reviews with ratings below 6, there are many positive words such as “fit well” and “looks great”. Of course, the most frequent word, “end wear,” also signals a negative sentiment. Therefore, based on this finding, we made a probability distribution of sentiment scores based on the overall comments.

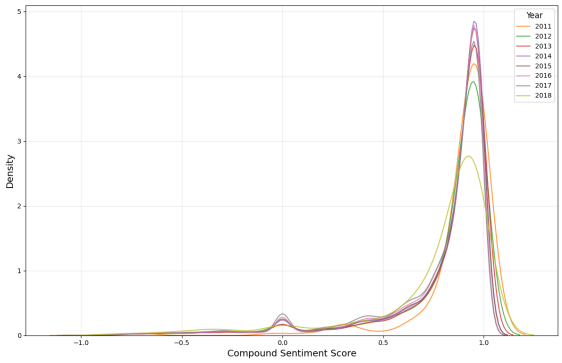

Figure 4: Sentiment Score Distribution

The sentiment score probability distribution chart illustrates the compound sentiment score trends across different years, with scores ranging from -1 (negative) to 1 (positive). A notable observation is the consistently strong positive sentiment, as evidenced by the sharp peaks in the range of 0.5 to 1 for all years. This indicates a high level of satisfaction and positive feedback from customers throughout the observed period.

While the density curves for each year are closely aligned, suggesting stability in customer sentiment, slight variations can be observed. For instance, the curve for certain years, such as 2014 and 2015, appears slightly narrower, indicating less variability in sentiment scores compared to other years. The minimal presence of scores in the negative range (-1 to 0) highlights that negative sentiment is rare and does not significantly impact the overall sentiment trends.

Overall, although the probability of a high emotional score declined in 2018, the distribution demonstrates a clear pattern of predominantly positive sentiment over time, reflecting consistent customer satisfaction and a strong brand reputation. This trend reinforces the brand’s ability to maintain positive customer experiences across different years.

4. Predictive Analytics

This chapter introduces our predictive analytics framework used to generate recommendations, which combines two core components: sentiment-driven scoring and item similarity measures. By optimizing the weights of these components, the recommendation system provides tailored suggestions for users.

4.1 Data Cleansing

This section explains the data preprocessing steps, which include rating classification, feature extraction, and data splitting. These steps prepare the raw data for the recommendation system by ensuring it is clean, structured, and ready for analysis.

4.1.1 Rating Classification

To categorize numerical ratings into meaningful groups, we define classes based on the numerical value of the rating:

$$C(r) = \begin{cases} \text{high}, & 8 \leq r \leq 10 \\ \text{medium}, & 6 \leq r \leq 7 \\ \text{low}, & 1 \leq r \leq 5 \end{cases} \tag{1}$$

where $C(r)$ is the classification of the rating r. This classification divides the ratings into interpretable groups:

- **High**: Ratings between 8 and 10 indicate strong customer satisfaction.
- **Medium**: Ratings between 6 and 7 suggest moderate satisfaction.
- **Low**: Ratings 1 and 5 highlight dissatisfaction or issues.

4.1.2 Feature Extraction

The next step involves extracting key features from the raw dataset to describe user behavior and review sentiment. The process includes:

1. **Sentiment Score Extraction**: Using the Sentiment Intensity Analyzer (SIA), the emotional tone of each review is analyzed to extract a compound sentiment score. This score provides a quantitative measure of the review's overall sentiment, ranging from negative to positive. The details of how SIA computes the sentiment score will be elaborated in Section 4.2.1.
2. **User Purchase History**: For each user u, their historical purchases are recorded as:
$$H_u = \{i_1, i_2, \ldots, i_n\} \tag{2}$$
where H_u represents the set of items previously purchased by user u, and $i_1, i_2, \ldots, i_n$ are the item IDs.
3. **Feature Representation**: Each record in the dataset is represented as:
$$\mathbf{F} = \{u, i, S_{\text{sentiment}}, H_u\} \tag{3}$$
where u is the user ID, i is the item ID, $S_{\text{sentiment}}$ is the sentiment score, and H_u is the user's purchase history.
4. **Word2Vec-Based Vectorization**: To enrich the feature representation of items, we utilize Word2Vec to generate dense vector embeddings from review texts. The model is trained on tokenized review sentences using a skip-gram approach, with a vector size of 100, a window size of 5, and a minimum word frequency of 2. For each item i, the review-based vector representation is obtained by averaging the Word2Vec embeddings of all words in its associated review text:
$$\mathbf{v}i = \frac{1}{|W_i|} \sum w \in W_i \mathbf{e}_w \tag{4}$$
where W_i is the set of words in the review text of item i, and $\mathbf{e}_w$ represents the embedding vector of word w in the Word2Vec space. The resulting vector $\mathbf{v}_i$ captures semantic similarities between items based on review content. These embeddings are later used in similarity computations and integrated with sentiment scores to improve recommendation quality.

4.1.3 Data Splitting

To ensure robust evaluation of the model, the dataset is split into training and testing subsets. A random split (42 random seed) is performed with a specified ratio α for training data. The splitting process is defined as:

$$\begin{aligned} \text{Train Set} &= \mathbf{D}[: n \cdot \alpha] \\ \text{Test Set} &= \mathbf{D}[n \cdot \alpha :] \end{aligned} \tag{5}$$

where $\mathbf{D}$ is the dataset, n is the total number of records, and α is the proportion allocated to the training set. Typically, $\alpha = 0.8$ is used, ensuring 80% of the data is used for training, and 20% for testing.

4.2 Mathematical Formulation

This section describes the mathematical formulation of the recommendation system. The model computes a recommendation score $S_{\text{recommend}}$ for each item by combining two key components:

1. **Sentiment-driven scoring**: Extracted from the sentiment analysis of user reviews.
2. **Item similarity scoring**: Calculated based on the user's historical purchases and the target item's textual similarity.

By optimizing the weights of these two components, the system provides a personalized recommendation for each user. Additionally, the model classifies items into three categories (high, medium, and low) based on the recommendation score.
The recommendation score for an item i is computed as:

$$S_{\text{recommend}} = w_1 \cdot S_{\text{sentiment}} + w_2 \cdot S_{\text{similarity}} \quad (6)$$

where:

- $S_{\text{recommend}}$: Recommendation score.
- $S_{\text{sentiment}}$: Sentiment score derived from the review text.
- $S_{\text{similarity}}$: Similarity score based on user purchase history.
- w_1, w_2: Weights of sentiment and similarity components, respectively, where $w_1 + w_2 = 1$.

4.2.1 Sentiment Score

The sentiment score is computed using the Sentiment Intensity Analyzer (SIA), a lexicon-based tool that evaluates the emotional tone of textual data. SIA assigns a compound sentiment score, which is a single value representing the overall sentiment polarity of the text.
SIA relies on a combination of heuristics and pre-defined sentiment lexicons. The core principles of SIA include:

- **Lexicon-based scoring**: SIA uses a dictionary of words and their associated sentiment intensities (e.g., "excellent" has a positive score, while "terrible" has a negative score).
- **Contextual adjustment**: The tool adjusts sentiment scores based on linguistic features such as punctuation, capitalization, negation (e.g., "not good"), and degree modifiers (e.g., "very good" or "slightly bad").
- **Polarity aggregation**: SIA aggregates the scores of individual words and phrases to calculate an overall sentiment score for the text.

The final output of SIA includes four scores:

1. **Positive** (S_{positive}): The proportion of the text that conveys positive sentiment.
2. **Negative** (S_{negative}): The proportion of the text that conveys negative sentiment.
3. **Neutral** (S_{neutral}): The proportion of the text that conveys neutral sentiment.
4. **Compound** ($S_{\text{sentiment}}$): A weighted aggregate score ranging from -1 (most negative) to 1 (most positive).

The compound score, $S_{\text{sentiment}}$, is derived as:

$$S_{\text{sentiment}} = \text{SIA}(T_{\text{review}}) \quad (7)$$

where T_{review} represents the review text. This score serves as the primary metric for assessing the overall sentiment of a review.

4.2.2 Similarity Score (Based on TF-IDF)

The similarity score is computed using TF-IDF vectors and cosine similarity:

$$S_{\text{similarity}} = \frac{\sum_{j \in H} \cos(\mathbf{v}_j, \mathbf{v}_i)}{|H|} \quad (8)$$

where:

- H: Set of items in the user's purchase history.
- $\mathbf{v}_j$: TF-IDF vector of item j in the user's history.
- $\mathbf{v}_i$: TF-IDF vector of the target item i.
- $\cos(\cdot, \cdot)$: Cosine similarity function.

4.2.3 Baseline Models and Comparisons

To evaluate the effectiveness of the proposed recommendation system, we implemented two baseline models:

- **Random Recommendation**: Items are randomly recommended to users without considering any prior user preferences or item features. This serves as a naive baseline to highlight the value of leveraging sentiment scores and similarity metrics.
- **Rating-based Recommendation**: Recommendations are made purely based on item ratings, without considering user purchase history or sentiment analysis. This model ranks items by their average user ratings.

These baseline models provide a point of comparison to assess the improvements achieved by integrating sentiment-driven scoring and similarity measures. The key differences between the methods are shown below:

- Random recommendation does not consider any user-specific information.
- Rating-based recommendation lacks personalization but provides an aggregated view of item popularity.
- Our proposed model combines sentiment and similarity, addressing user-specific preferences.

4.3 Model Optimization

This section describes the methods used to optimize the weights of the recommendation model and evaluate its performance. The goal is to maximize accuracy by combining sentiment scores and similarity scores with optimal weights.

4.3.1 Evaluation of Weights

The model's performance is evaluated by testing different combinations of weights for the sentiment score and similarity score. The evaluation function calculates a recommendation score for each item and classifies it into one of three categories: high, medium, or low. The classification is defined as:

$$C(S_{\text{recommend}}) = \begin{cases} \text{high}, & S_{\text{recommend}} \geq 0.5 \\ \text{med}, & 0.2 \leq S_{\text{recommend}} < 0.5 \\ \text{low}, & S_{\text{recommend}} < 0.2 \end{cases} \tag{9}$$

- $S_{\text{recommend}}$ is the weighted recommendation score.
- $w_1 + w_2 = 1$. The thresholds $\theta_{\text{high}} = 0.5$ and $\theta_{\text{medium}} = 0.2$ are used for classification.

The evaluation function iterates through the dataset and computes the accuracy of the predicted labels compared to the true labels:

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of samples}} \tag{10}$$

4.3.2 Grid Searching

To find the optimal weights w_1 and w_2, a grid search is performed over a predefined range of values:

$$w_1 \in \{0, 0.1, \ldots, 1.0\}, \quad w_2 = 1 - w_1 \tag{11}$$

For each combination of w_1 and w_2, the accuracy is evaluated on the training dataset. The best weights are selected based on the highest accuracy:

$$\{w_1^*, w_2^*\} = \arg\max_{w_1, w_2} \text{Accuracy}(w_1, w_2) \tag{12}$$

The figure below illustrates the relationship between the weight of the sentiment component (w_1) and the overall accuracy of the model. The results indicate that the model achieves optimal accuracy when $w_1 = 0.9$ and $w_2 = 0.1$, highlighting the significant contribution of sentiment analysis in the recommendation process.

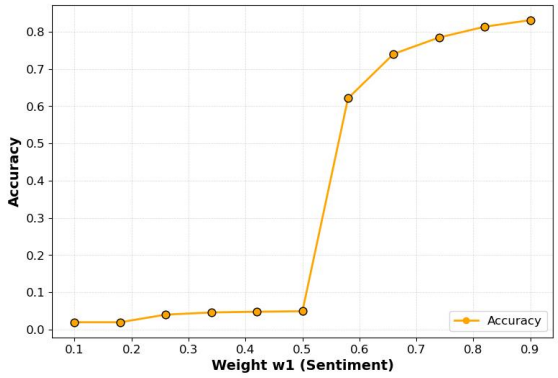

Figure 5: Accuracy vs. Weight w_1 (Sentiment Component)

4.3.3 Final Recommendations

Once the optimal weights w_1^* and w_2^* are determined, the final recommendations are generated by ranking items for each user based on the recommendation score in descending order. The top 20 items are presented as recommendations for each user.

4.3.4 Performance Evaluation

The optimized weights and the model's performance are summarized as follows:

- **Optimal Weights**: $w_1^* = 0.9$, $w_2^* = 0.1$
- **Training Accuracy**: 83.21%
- **Validation Accuracy**: 83.67%

The results demonstrate the effectiveness of combining sentiment scores and similarity scores with carefully optimized weights to improve recommendation accuracy.

4.4 Comparative Results and Discussion

The performance of the proposed recommendation system was compared against the baseline models using the validation set. The results are summarized in Table.

Model	Accuracy (%)	Precision (%)
Random	33.45	32.10
Rating-based	65.12	62.45
Our Model	**83.10**	**81.32**

Table 1: Performance Comparison

The proposed model significantly outperformed the baseline models in terms of accuracy and precision. This improvement can be attributed to the integration of sentiment analysis and similarity measures, which capture both user preferences and item-specific features.

4.4.1 Strengths of the Proposed Model

- **Combination of Sentiment and Similarity**: Leveraging textual sentiment analysis alongside purchase history similarity provides a more comprehensive understanding of user preferences.
- **Optimized Weighting**: The use of grid search to optimize weights ensures a balanced contribution from sentiment and similarity components, enhancing recommendation accuracy.

4.4.2 Limitations and Challenges

While the proposed model demonstrated significant improvements over baseline methods, several limitations were observed:

- **Scalability**: Calculating similarity scores for a large number of items or users increased computational costs, posing challenges for scaling the model to larger datasets.
- **Cold Start Problem**: For new users or items with limited historical data, the model's effectiveness was reduced as similarity and sentiment scores could not be fully leveraged.
- **Bias in Ratings and Reviews**: User-generated content may reflect biases, such as over-representation of extreme reviews, potentially affecting the fairness and generalizability of recommendations.
- **Overfitting Risk**: When tuning weights (w_1, w_2), overfitting to training data was observed in some cases, especially when the validation dataset was not diverse enough.

To address these limitations, future iterations could incorporate hybrid recommendation techniques, such as collaborative filtering

and matrix factorization, or explore methods to balance computational efficiency with personalization accuracy.

5. Conclusion

The proposed recommendation system successfully integrates sentiment analysis and similarity-based measures, offering a comprehensive approach to understanding user preferences in the fashion domain. By leveraging sentiment scores from review texts and similarity metrics derived from purchase histories, the model achieves a significant improvement in accuracy and precision compared to baseline methods, demonstrating its potential for personalized recommendation tasks.

Key strengths of the model include its ability to combine textual and behavioral data effectively and its optimized weighting strategy, which ensures a balanced contribution of sentiment and similarity components. These innovations address the multifaceted nature of user preferences and provide actionable insights for enhancing user satisfaction and loyalty.

Despite its effectiveness, the model faces challenges in scalability and handling the cold start problem. Additionally, biases inherent in user-generated data and potential overfitting highlight areas for future improvement. To overcome these challenges, hybrid recommendation techniques and methods for reducing computational costs should be explored.

Overall, the methodology presented in this study provides a robust framework for building effective recommendation systems. Beyond the fashion domain, this approach can be extended to other industries where user preferences are shaped by both qualitative feedback and historical behaviors. The results underscore the value of integrating sentiment-driven insights into modern recommendation systems, paving the way for more personalized and meaningful user experiences.