Developing an Enhanced Recommendation System for Women's Clothing in E-Commerce: A Comparative Analysis of Traditional Machine Learning and Deep Learning Approaches

Rezky Agung, Fairuuz Nurdiaz Amaanullah

June 12, 2024

1 Introduction

Customer reviews play an important role in the e-commerce sector. These reviews serve as an important tool for understanding customer sentiment. These reviews provide valuable insights into consumer preferences, experiences, and satisfaction levels, thereby enabling e-commerce businesses to make informed decisions. By analyzing customer feedback, companies can identify trends, improve product offerings, and improve the overall shopping experience [1]. This not only helps build a loyal customer base but also drives business growth and competitiveness in a crowded market. Therefore, leveraging customer reviews is crucial for women's clothing e-commerce platforms that aim to stay in tune with their customers' needs and preferences.

However, customer reviews have problems that need to be addressed, such as a variety of different feature types, namely text, categorical, and numeric. A lot of data is missing, such as the text in the review title where the solution must use adjective extraction to get a better meaning [2]. In the feature selection process, there are 10 important features, including: Title, Review Text, Rating, Recommended IND, etc [3]. Selecting the Recommended IND feature as a label is important because it shows whether customers recommend the product, which is a key indicator of customer satisfaction. Using the Rating feature is also very important because it provides direct information regarding customer assessments of products, which helps in analyzing sentiment in review text and title text to decide on business policies.

The problem of data imbalance in customer review datasets requires special handling to ensure that the machine learning model being built is not biased. This imbalance can cause the model to be more likely to predict the majority class, ignoring the minority class which is also important [4]. To balance the dataset, methods such as K-Means or Synthetic Minority Over-sampling Technique (SMOTE) can be used [5]. K-Means can help with clustering and then re-sampling to balance the classes, while SMOTE works by synthesizing new samples from the minority class to increase the number until it is more balanced with the majority class [6]. These two techniques help create a more balanced dataset, ultimately improving the model's performance and accuracy in predicting outcomes in both classes.

The recommendation flow starts with a simple Logistic Regression[7] with TF and TF-IDF [8] feature extraction, also the other comparing text representation using GloVe embeddings[9], to ensure fix length on features some approach are use i.e mean, median, max, midpoint, and absolute max pooling. after that the best pooling are combined to TF-IDF to see whether the performance are get better or not. But ML models still fail to capture complex patterns in textual data. Convolutional neural networks (CNNs) [10] excel at recognizing local patterns, Recurrent neural networks (RNN) [11] is good for processing sequential data, while Long short term memory [12] is able to handle long-term dependencies in text. The combination of CNN-LSTM with GloVe Embedding provides rich word representation and deep contextual understanding, resulting in highly accurate and reliable predictions.

2 Methods and Eksperiments

2.1 Datasets Analysis

The Datasets

The dataset utilized for this analysis is the Women's Clothing E-Commerce Reviews. It comprises reviews authored by genuine customers and has undergone anonymization. It appears that our label distribution is unbalanced.

With a mean of 0.818, it's evident that the majority of instances are labeled as 1. Additionally, positive feedback is concentrated within certain values, suggesting that relying on these features may not be advisable moving forward.

2.2 Preprocessing Data

1. Missing Value Handling

   To handle missing values, we first perform punctuation removal and lowercasing on each data entry. For other missing value titles, we fill them by extracting only the adjectives from the review text, the example is :

Review Text	Absolutely wonderful - silky and sexy and comfortable
Title (After Extract Adjective)	wonderful sexy comfortable

2. Feature Selection

   Rating and Recommended IND exhibit a high correlation score, while the remaining features display low positive or negative correlations. Hence, solely Rating and Recommended IND suffice as features. Furthermore, age is excluded as a feature since product recommendation is independent of age. Ratings carry significant importance for subsequent sentiment analysis, as they provide clarity in distinguishing between positive and negative reviews.

2.3 Features Engineering

Text data is unstructured text also, it's more not straight forward as usual on tabular data, where it can directly fed into model, hence choosing some proper techniques are necessary to ensure better performance on our model.

2.3.1 Features Extraction

1. **Bag-Of-Words Representation**

   BOW is a corpus text representation where using word as features and also each document will represent by a vector. The BoW model is simple and easy to implement, but it has limitations in capturing the meaning of language. We use term frequency (TF) and term frequency - index document frequency (TF-IDF) to ekstraks some features in the data.

2. **Global Vectors for Words Representation**

   The Core Idea is this methods using Matrix Factorization to construct co-occurence matrix X, where X_{ij} denotes how often word j appears in the context of word i. Then this matrix is factorized into word vectors W and context vectors C such their dot product approximates the logarithms of the co-occurence probabilities $X_{ij} \approx exp(W_i^T C_i)$.

2.3.2 Re-balancing Using SMOTE and K Means

Our Analysis has detected some common problem about unbalanced label, unbalance label can lead some major effect that found, our model cannot learn on minority patterns data, hence to address this problem, use Synthetic Minority Oversampling Technique (SMOTE) and K-Means Clusterring to deal with this data, at the and the problem solved with increasing F1-Score on Zero classes prediction.

2.3.3 Pooling

This approach is widely used due to its simplicity and effectiveness in capturing the overall meaning of a text by averaging the word vectors. We use mean pooling with approach can be done by take avarage of the vectors with unfixed lenght sentence n of each data

2.4 Model Workflow

1. **Machine Learning with Logistic Regression Workflow**

   Machine Learning with Logistic Regression Workflow below is diagram of ML workflow of our process. Then the result will evaluate with metrics like accuracy, precision, recall, F1-Score, and loss in figure 1 :

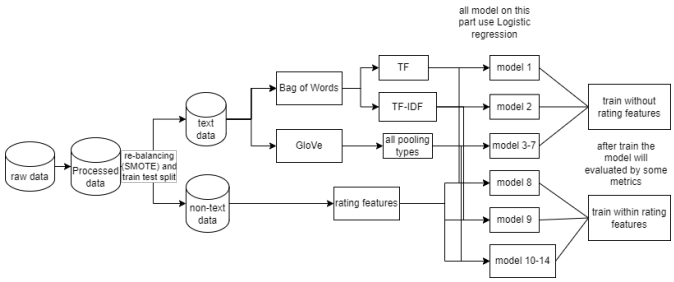

Figure 1: Machine Learning Workflow

2. **Deep Learning CNN-LSTM workflow (proposed method)**

   Deep Learning with Neural Networks Workflow below is diagram of DL workflow of our process. Then the result will evaluate with metrics like accuracy, precision, recall, F1-Score, and loss in figure 2 :

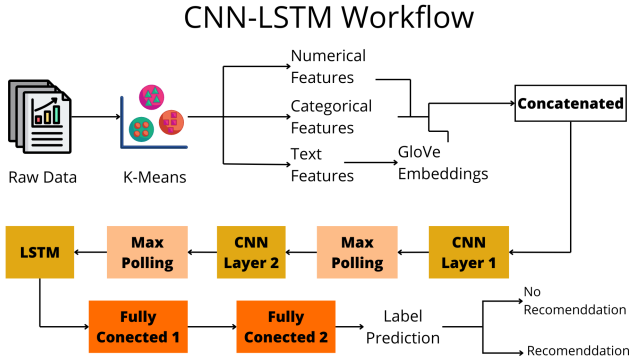

Figure 2: Deep Learning Workflow

2.5 Result and Analysis

1. Result of Machine Learning Approach

   (a) Result of machine learning approach without rating features

Text Representation	Accuracy	Precision	Recall	F1-Score	Loss
TF	0.928	0.927	0.926	0.926	0.2249
TF-IDF	0.925	0.940	0.903	0.921	0.2338
GloVe 200D + mean pooling	0.848	0.857	0.825	0.840	0.3668
GloVe 200D + TF + mean pooling	**0.934**	**0.950**	**0.913**	**0.931**	**0.1959**
GloVe 200D + TF-IDF + mean pooling	0.927	0.944	0.904	0.924	0.2288

Table 1: Performance Metrics for different Text Representation with LogisticRegression Without Rating Features

The results showed minimal overfitting without regularization, with accuracy differences of only 1-3%. TF performed slightly better than TF-IDF in terms of Accuracy and F1-Score. While GloVe 200D with mean pooling performed well, it fell short compared to other methods. Mean pooling consistently outperformed other pooling methods, such as median, midpoint, and abs max pooling. Increasing the dimensionality of GloVe embeddings from 50D to 200D improved performance, with GloVe 200D with mean pooling outperforming lower-dimensional counterparts. The best overall performance was achieved by combining GloVe 200D embeddings with TF and mean pooling, resulting in the highest Accuracy (0.934), Precision (0.950), Recall (0.913), F1-Score (0.931), and lowest Loss (0.1959), effectively leveraging the strengths of both methods.

   (b) Result of machine learning approach with rating features

Text Representation	Accuracy	Precision	Recall	F1-Score	Loss
TF	0.958	0.952	0.962	0.957	0.1272
TF-IDF	0.958	0.968	0.944	0.956	0.1297
GloVe 200D + mean pooling	0.958	0.972	0.940	0.956	0.1375
GloVe 200D + TF + mean pooling	**0.967**	**0.979**	**0.952**	**0.965**	**0.1136**
GloVe 200D + TF-IDF + mean pooling	0.958	0.967	0.947	0.957	0.0598

Table 2: Performance Metrics for different Text Representation with LogisticRegression Without Rating Features

Introducing rating features enhances performance metrics across all text representations, leading to higher accuracy, precision, recall, and F1-scores, and lower loss values. Notably, GloVe 50D with mean pooling significantly improves with the inclusion of rating features, achieving an accuracy of 0.961 and an F1-score of 0.959, indicating that lower-dimensional embeddings can excel with additional informative features. The best-performing combination remains GloVe 200D + TF + mean pooling, with the highest metrics: an accuracy of 0.967, precision of 0.979, recall of 0.952, F1-score of 0.965, and the lowest loss of 0.1136.

2. Result of Deep Learning Approach The table 3 presents the performance of various neural networks models using GloVe embedding in women's clothing e-commerce review analysis, with accuracy, precision, recall, F1-score, and loss metrics measured for each model. The models compared include GloVe CNN, GloVe RNN, GloVe LSTM, and a combination of GloVe CNN LSTM.

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**Loss**
Golve CNN	0.9703	0.9703	0.9703	0.9703	0.1658
Golve RNN	0.9726	0.9726	0.9726	0.9726	0.1397
Golve LSTM	0.9735	0.9735	0.9735	0.9735	0.1298
Glove CNN LSTM	0.9764	0.9764	0.9764	0.9764	0.0688

Table 3: Performance metrics for different models

Based on the table 3, the performance of the neural networks model with GloVe embedding shows very good results in analyzing women's clothing e-commerce reviews. The GloVe CNN model achieved accuracy, precision, recall, and F1-score of 0.9703 with a loss of 0.1658, showing consistent performance but with slightly higher loss than the other models. The GloVe RNN model improves the performance metrics slightly higher with a value of 0.9726 and a reduction in loss to 0.1397, while the GloVe LSTM model shows further improvement with a value of 0.9735 and a loss of 0.1298, indicating better ability in handling long-term dependencies in text data. Finally, the GloVe CNN LSTM model shows the best performance with all metrics of 0.9764 and the lowest loss of 0.0688, combining the strengths of CNN in capturing local features and LSTM in handling sequential context, thus providing the most accurate and efficient results.

3 Conclusion

Clustering is essential for addressing data imbalances, leading to improved prediction accuracy by creating a more balanced dataset, especially when certain classes are underrepresented. Among the 10 available features, the review text, title text, and rating features are the most influential, offering critical insights into customer opinions and satisfaction. Logistic Regression models perform well in label prediction, but Deep Learning models excel in capturing complex data patterns, making them superior for sentiment analysis in women's clothing recommendations. Techniques like data augmentation and transfer learning further enhance Deep Learning models by utilizing pre-trained models and expanding training data without new data collection. Combining clustering, influential textual and rating features, and advanced Deep Learning models forms a robust framework for accurate and reliable sentiment analysis in women's clothing recommendations.

References

[1] X. Lin, Sentiment analysis of e-commerce customer reviews based on natural language processing, in: Proceedings of the 2020 2nd international conference on big data and artificial intelligence, 2020, pp. 32–36.

[2] A. Noor, M. Islam, Sentiment analysis for women's e-commerce reviews using machine learning algorithms, in: 2019 10th International conference on computing, communication and networking technologies (ICCCNT), IEEE, 2019, pp. 1–6.

[3] A. F. Agarap, Statistical analysis on e-commerce reviews, with sentiment classification using bidirectional recurrent neural network (rnn), arXiv preprint arXiv:1805.03687 (2018).

[4] S. S. Rawat, A. K. Mishra, Review of methods for handling class-imbalanced in classification problems, arXiv preprint arXiv:2211.05456 (2022).

[5] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, Smote: synthetic minority over-sampling technique, Journal of artificial intelligence research 16 (2002) 321–357.

[6] P. S. Bradley, K. P. Bennett, A. Demiriz, Constrained k-means clustering, Microsoft Research, Redmond 20 (0) (2000) 0.

[7] M. P. LaValley, Logistic regression, Circulation 117 (18) (2008) 2395–2399.

[8] A. Aizawa, An information-theoretic perspective of tf–idf measures, Information Processing & Management 39 (1) (2003) 45–65.

[9] J. Pennington, R. Socher, C. D. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.

[10] K. O'shea, R. Nash, An introduction to convolutional neural networks, arXiv preprint arXiv:1511.08458 (2015).

[11] S. Grossberg, Recurrent neural networks, Scholarpedia 8 (2) (2013) 1888.

[12] A. Graves, A. Graves, Long short-term memory, Supervised sequence labelling with recurrent neural networks (2012) 37–45.