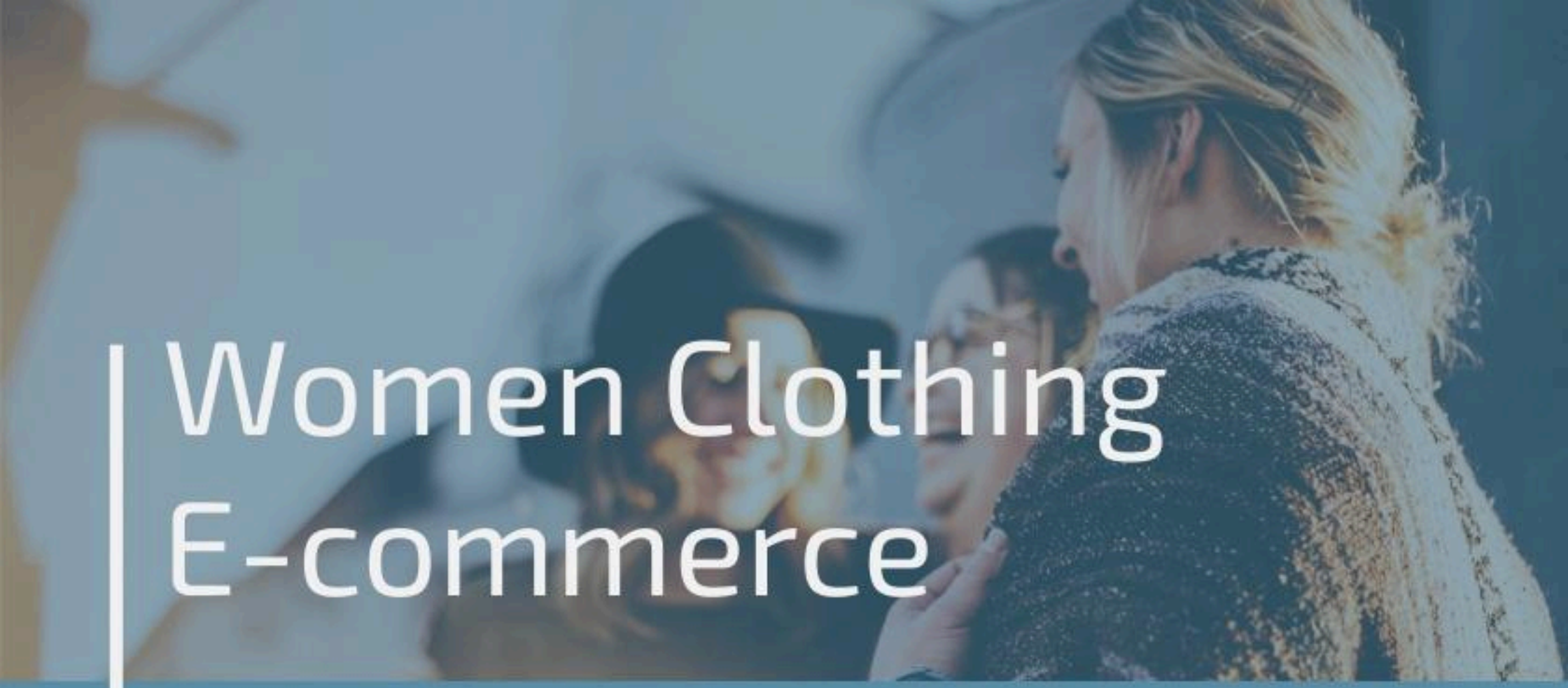

Women Clothing E-commerce

Perform sentiment analysis to explore trends in customer reviews, extract actionable insights to improve online clothing e-commerce

August 2020

Open Door to Sentiment Analysis for E-commerce

The data is a collection of 22641 Rows and 10 column variables. Each row includes a written comment as well as additional customer and product information. The total number of unique words in the dataset is 9810.

The first step is text preprocessing, title and reviews are combined to one columns to reduce missing review information. Then dropping null value rows and removing special symbols, digits, punctuation, pronouns. We explored three ways to define positive and negative reviews: using product ranking, recommendation IND and their interaction. The second step is utilizing two Natural Language techniques to extract feature from reviews: Bag of word which is to weight the words and increase the influence of important words; Word-embedding which is to convert words to vector. The third step is performing powerful machine learning algorithms Logistics Regression and Naive Bayes to classify positive and negative reviews. Logistics Regression utilizes sigmoid function and works best on this binary classification problem. Naive Bayes is an extreme fast and easily interpretable model.

Challenges and problems solving

It's always a challenge for sentiment analysis to classify those reviews having both positive words and negative words but in negative meaning as positive reviews. Instead of manually keeping all negative words, we reversed the polarity of all words after negative words and thus reduce the false negative rate by 10%.

To further improve the model performance and stability, we performed randomize search and K-fold cross validation to fine tuning the model and improved classification ability by 45%.

Visit On point Insights:
http://onpointinsights.co

Bringing Review Score Together

Focus more on target customers in their 30s

Increasing sales through reducing negative reviews

According to the number of reviews for each age group, 30s women has the highest number of reviews. But the average rating(4.17) and positive review rate(81.36%) are not perform well compared to other age groups. Here is a potential big improvement once we focus more on the reviews from 30s

The knits and blouses are the most popular products for women customers in their 30s. However blouses has bad performance with the lowest average rate and positive rate compared to other products. Also according to word cloud: size and color are the two main parts complained by customers. For further development, we recommended retailer focus more on the negative reviews come from blouses brought by 30s women, improve size helper accuracy and color fidelity shown in pictures.

Bring New Opportunities to Jackets

Compared to other department like Dress, Bottons.Jackets has the lowest sales but the positive reviews rate(83.62%) performs well. We suggest that retailer can make more advertisements for Jackets.

95% accuracy to analyze review polarity

Using better review's polarity measurement

By comparing between three methods for polarity measurement: using ranking(1-5) with ranking>=3 as positive, recommended IND(0 or 1) and adding them together with sum>=4 as positive. Ranking and recommended IND are convoluted because each customer differently defines ranking 3 or 4 as either recommended or not recommended. By using the interaction measurement, the accuracy improved 0.5% than using ranking and 0.2% than using recommended IND.So the interaction measurement can give retailer better perspective whether reviews are positive or negative than only use ranking or recommended IND.