Role of Topic Modeling in the Product Reviews Sentiment Analysis

Deyuan Wang, Jieqiao Luo, Yiming Yu

Abstract

The rapid growth of e-commerce market, especially during the period of COVID-19, draws attention from various e-commerce companies around the world. Such growth amplifies the importance of customer reviews. The paper, a comparative analysis, aims at examining the role of topic modeling in the sentiment analysis of product reviews. The study applied two topic modeling techniques, including: K-means Clustering and LDA with Bag of Words to 2018 Women's E-Commerce Clothing Reviews dataset. Then we used three classification models, including Logistic Regression, Naive Bayes, and Support Vector Machine (SVM) to compare the sentiment prediction matrices between the features with topic modeling and that without topic modeling. The results show that although topic modeling has slight direct impact on improving the models' performance to predict the sentiment, it can reflect some insightful patterns behind the evaluation matrices, such as users' reviews overall sentiment, users' preference, and popular categories.

1 Objective

The objective of the project is to explore the role of topic modeling in sentiment analysis. There are two main approaches for the project. First, it is comparative analysis. It includes basic sentiment analysis, sentiment analysis with topic modeling, sentiment analysis with K-means clustering. The results are shown by comparing top frequent key words. The second is the sentiment analysis, rating is assumed as the actual sentiment labels. Accuracy, precision, recall, and F1-score for three methods are measured to evaluate models' performance.

2 Dataset

For the dataset, 2018 women's clothing e-commerce dataset is used. It is from the paper "Statistical Analysis on E-Commerce Reviews, with Sentiment Classification using Bidirectional Recurrent Neural Network". It has 23486 rows and 10 features. Each row corresponds to a customer review and involves three main categorical information. Details of the datasets are shown in the table 1.

Main categorical information	Features	Description
User information	Age	Positive Integer variable of the reviewer's age.
Product information	Clothing ID	Integer Categorical variable that refers to the specific piece being reviewed
	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
Review information	Review Title	String variable for the title of the review. (Since there is over 15% missing review title, it is not used for the analysis)
	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.

Table 1 Descriptions of Features

Exploratory data analysis is proceeded as follows. Firstly, Age distribution for rating level, department name, Recommended IND and class name are consistent. It is shown in the figure 1.

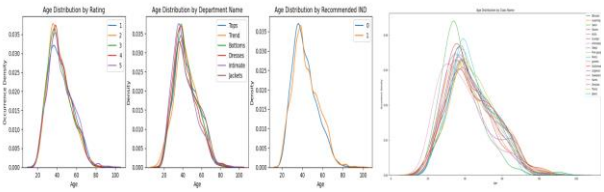


Figure 1 Age distribution for different features

Secondly, rating level in different product categories follows the similar frequency distribution. It is shown in the figure 2.

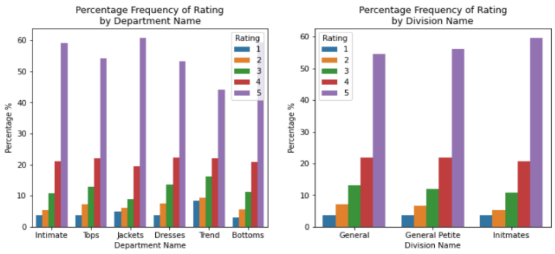


Figure 2 Rating Frequency Distribution for Different Product Features

Thirdly, the recommended IND feature has high positive correlation with rating level. It cannot be used as a predictor feature. It is shown in the figure 3.

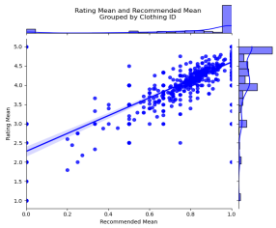


Figure 3 Linear Regression plot for Recommended ID and Rating

3 Background

This section compares various existing word done on sentiment analysis and topic modeling related to online reviews.

Sentiment Analysis for marketing intelligence effort is a latest research focusing on text sentiment analysis on Twitter. Appling TF-IDF tokenization with Naïve Bayes classifier reached over 90% precision and recall (Alamsyah & Saviera, 2021). The text data was extracted from Twitter and preprocessed using a designed pipeline.

Another research on product review Sentiment Analysis proposed a pipeline with two models including Naïve Bayes and Support Vector Machine (SVM) with a more sophisticated system of text preprocessing including Lemmatization, Stemming, and Dependency Parser (Basani et al., 2019). This paper proposed a much complex system that can handle various text preprocessing issues. The machine learning system includes comparing two different models with various measurement. Such machine learning system helped improving our research methodology.

Research from real time Sentiment Analysis of e-commerce websites proposed using Decision Tree, Naïve Bayes, and Maximum Entropy Classification (Sudheer & Valarmathi, 2018). The results concluded that the accuracy of sentiment analysis is related to the number of features selected.

A review on Sentiment Analysis methodology in e-commerce summarized various techniques in Sentiment Analysis from machine learning based approaches to lexicon-based approaches (Marong et al., 2020). Machine learning based approach is divided into supervised learning to unsupervised learning. Lexicon-based approach, however, mostly requires predefined word phrases and opinion words, which increase the complexity of building an automatic system, and faces various issues including slangs and misspelling that cannot be handled as easy as machine learning based approaches.

Topic modeling is another important part in this research. This part introduces previous research of topic modeling in sentiment analysis.

Topic modeling methodologies in sentiment analysis has been summarized in a previous research (Rana et al., 2016). Rana et al. summarized two basic methodologies for topic modeling, probabilistic Latent Semantic Analysis (pLSA) and Latent Dirichlet Allocation (LDA). pLSA, however, is not favorable to most research due to its own property. Therefore, LDA would be more preferable in most cases. Their research also summarized various LDA extended methods that have been proven useful in Sentiment Analysis.

A similar research on Sentiment Analysis in women's e-commerce clothing reviews was conducted using topic modeling (Parveen et al., 2020). Parveen et al. proposed using Latent Dirichlet Allocation (LDA) to conduct review summary classification for Sentiment Analysis. The resulting document-level classification conducted with 5 topics where each topic was automatically calculated using LDA, and each topic with highest important words provided a useful insight of how each group/topic is generated. Such topic classification can be used in further research in Sentiment Analysis.

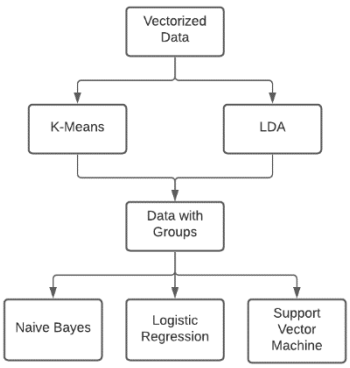
4 Methodology

Pre-processing

Text data need to be preprocessed before feeding into various models. Stop words were removed to simplify the corpus. Snowball stemmer was done using NLTK to simplify words. After tokenization, Count Vectorization was used to vectorize words.

Classification

Before the data is used for sentiment classification, unsupervised machine learning methods were applied first to separate different customers into different groups.



We proposed our workflow of building the classification system. After the text data was vectorized, two unsupervised learning methods were proposed to separate customers into different groups: K-Means and LDA. after the data is grouped with those unsupervised methods, the data is feed into three separate models: Naïve Bayes, Logistic Regression, and Support Vector Machine.

4.1 Unsupervised Methods

K-Means is one of the basic unsupervised methods that groups data based on partitioning observations into k clusters based on nearest mean. It is efficient and fast compared with other much complexed methods. We set K into a range and found the optimal K based on Silhouette Method.

Latent Dirichlet Allocation (LDA) is a topic modeling method that can be used to classify document-level data into different unobserved groups. We set the number of documents to be 7, same as the number of divisions.

4.2 Supervised Methods

Naïve Bayes is one of the simplest text classification methods based on Bayes Theorem. Text data usually words good with naïve bayes because of its assumption that each attribute is not correlated with others.

Logistic regression is mostly useful in binary classification by estimating the log odd of an event. It can also be used in case of multiple categories.

Support Vector Machine (SVM) classifies data into different groups by setting a hyperplane between the closest points in different groups with maximum margin.

5 Results

Based on the K-means clustering outcome, when the number of clusters equals to nine, three models' F1-score are the highest. As the table shown in below, top 10 words in different clusters reflects several patterns.

Cluster 1	love	size	great	wear	fit	color	just	perfect	like	soft
Cluster 2	dress	love	fit	size	like	great	wear	beautiful	just	fabric
Cluster 3	fit	size	just	love	perfect	fabric	color	like	wear	ordered
Cluster 4	size	fit	dress	love	wear	ordered	small	like	just	great
Cluster 5	sweater	love	great	color	soft	like	wear	size	beautiful	just
Cluster 6	like	just	look	fabric	really	wear	color	love	cute	looks
Cluster 7	small	size	wear	medium	runs	ordered	large	fit	like	love
Cluster 8	great	love	wear	fit	color	jeans	comfortable	cute	size	looks
Cluster 9	love	wear	cute	just	beautiful	fabric	perfect	color	size	flattering

Table 2 K-means Clustering Top 10 Keywords

For instance, most adjectives and verbs are positive in different clusters, which fits with the fact that most sentiment in the dataset are positive. In addition, color and size are the two important aspects that users would focus on when they purchase the clothing. Furthermore, users tend to leave more comments on dress, sweater, and jeans.

Besides using K-means clustering techniques, our group also apply Latent Dirichlet allocation (LDA) to the dataset, the number of topics equal to six are optimal. The table below shows top 10 words in different topics.

Topic 1	colors	fits	ordered	great	just	length	fit	love	perfect	size
Topic 2	large	look	just	fit	ordered	cute	petite	xs	size	like
Topic 3	work	great	like	wear	cute	long	love	short	skirt	shirt
Topic 4	got	great	fit	runs	large	wear	size	pants	medium	small
Topic 5	quality	fit	just	love	nice	soft	color	beautiful	flattering	fabric
Topic 6	look	comfortable	like	looks	color	jeans	soft	love	sweater	great

Table 3 LDA Top 10 Keywords

There are some common and differences between K-means clustering results and LDA results. Some patterns are similar, such as “great”, “love”, “fit”, and “color” appears frequently in different topics, showing that most reviews are positive, and users care color and size most. In addition, compared with K-means clustering results, topic 3 and topic 4 reflects that skirts and pants are two additional categories users tend to leave comments on. Moreover, besides color and size, topic 5 shows that users would focus on quality and fabric of clothing.

In order to evaluate the role of topic modeling in sentiment analysis prediction, our group decide to apply Logistic Regression, Naïve Bayes, and Support Vector Machine (SVM) to TF-IDF, bag of words, and bag of words with topic modeling outcomes separately.

For TF-IDF, as the evaluation metrics shown in below, SVM and Logistic Regression models have similar good prediction performance, while Naïve Bayes model has the lowest precision and F1 score.

Evaluation Matrices / Model	Accuracy Score	Precision Score	Recall Score	F1 Score
TfidfVectorizer Logistic Regression Model	0.77	0.994	0.77	0.867
TfidfVectorizer Naive Bayes Model	0.771	0.977	0.771	0.859
TfidfVectorizer SVM Model	0.77	1	0.77	0.87

Table 4 Tf-idf Vectorizer Evaluation Metrics

For bag of words, SVM model has the best performance as TF-IDF and Logistic Regression model has highest accuracy score but lower F1 score compared with SVM models. Naïve Bayes model has the worst performance with lowest accuracy and F1 score.

Evaluation Matrices / Model	Accuracy Score	Precision Score	Recall Score	F1 Score
Bag of Words Logistic Regression Model	0.776	0.914	0.776	0.833
Bag of Words Naive Bayes Model	0.763	0.846	0.763	0.798
Bag of Words SVM Model	0.77	1	0.77	0.87

Table 5 Bag of Words Evaluation Metrics

For bag of words plus K-means clustering outcome, SVM model again has the best performance and Logistic Regression model has medium performance. Naïve Bayes model again has the worst performance with lowest accuracy and F1 score.

Evaluation Matrices / Model	Accuracy Score	Precision Score	Recall Score	F1 Score
K-means(K=9) Bag of Words Logistic Regression Model	0.775	0.913	0.764	0.832
K-means(K=9) Bag of Words Naive Bayes Model	0.764	0.848	0.764	0.799
K-means(K=9) Bag of Words SVM Model	0.77	1	0.77	0.87

Table 6 Bag of Words Plus K-means clustering Evaluation Metrics

For bag of words plus LDA outcome, SVM model again has the best performance and Logistic Regression model has medium performance. Naïve Bayes model again has the worst performance with lowest accuracy and F1 score.

Evaluation Matrices / Model	Accuracy Score	Precision Score	Recall Score	F1 Score
LDA(Topic=7) Bag of Words Logistic Regression Model	0.775	0.912	0.775	0.832
LDA(Topic=7) Bag of Words Naive Bayes Model	0.764	0.849	0.764	0.8
LDA(Topic=7) Bag of Words SVM Model	0.77	1	0.77	0.87

Table 7 Bag of Words Plus LDA Evaluation Metrics

Overall, SVM model has the best performance to predict sentiment, and TF-IDF has best performance among four vectorization methods. However, since the Logistic Regression and Naïve Bayes model learn weights per feature, it will be fairer to compare bag of words and bag of words with topic modeling outcomes. Compared with bag of words, Naïve Bayes model with topic modeling features will have slightly better performance. To summarize, topic modeling has slight impact on improving prediction performance directly,

especially for the Logistic Regression model and SVM model.

6 Discussion

From above results, we can basically have a conclusion that although topic modeling has slight direct impact on improving the models' performance to predict the sentiment, it can reflect some insightful patterns behind the evaluation matrices, such as users' reviews overall sentiment, users' preference, and popular categories. Therefore, the role of topic modeling in predicting sentiment is not limited to improving number in evaluation matrices, but providing extra insightful information related to sentiment.

However, this research study has some limitations. The sentiment variable we created assumes rating has strong relationship with sentiment, and we only apply three classical algorithms to the dataset. In addition, due to the imbalanced sentiment distribution, we cannot gain some insightful ideas about the unsatisfied reviews. Therefore, the conceptual framework in this research and the impact of topic modeling could be extended to more balanced datasets and other e-commerce platforms to better understand the role of topic modeling in predicting sentiment.

References

Alfred. V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling, volume 1*. Prentice-Hall, Englewood Cliffs, NJ.

Alamsyah, A., & Saviera, F. 2018. A Comparison of Indonesia's E-Commerce Sentiment Analysis for Marketing Intelligence Effort (case study of Bukalapak, Tokopedia and Elevenia).

American Psychological Association. 1983. *Publications Manual.* American Psychological Association, Washington, DC.

Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 1981. Alternation. *Journal of the Association for Computing Machinery*, 28(1):114-133. https://doi.org/10.1145/322234.32224.

Marong, M. 2020. Sentiment Analysis in E-Commerce: A Review on The Techniques and Algorithms.

N. Parveen, M.V.B.T. Santhi, L. Ramani Burra et al., Women's e-commerce clothing sentiment analysis by probabilistic model LDA using R-SPARK, *Mate rials Today: Proceedings*, https://doi.org/10.1016/j.matpr.2020.10.064

Prof. K. Sudheer and Dr. B Valarmathi, Real Time Sentiment Analysis of E-Commerce Websites Using Machine Learning Algorithms, *International Journal of Mechanical Engineering and Technology* 9(2), 2018, pp. 180–193.

Rana, T.A., Cheah, Y., & Letchmunan, S. (2016). Topic Modeling in Sentiment Analysis: A Systematic Review. *Journal of ICT Research and Applications, 10*, 76-93.

Yuniarta Basani *et al* 2019 *J. Phys.: Conf. Ser.* **1175** 012103