Statistical Analysis on E-Commerce Reviews, with Sentiment Classification using Bidirectional Recurrent Neural Network

Preprint · March 2018

DOI: 10.13140/RG.2.2.16988.08321

Abien Fred Agarap

De La Salle University

Statistical Analysis on E-Commerce Reviews, with Sentiment Classification using Bidirectional Recurrent Neural Network

Abien Fred M. Agarap*
Department of Computer Science
Adamson University
Manila, Philippines
abien.fred.agarap@adamson.edu.ph

Paul M. Grafilon, Ph.D.†
Department of Computer Science
Adamson University
Manila, Philippines
grafilonpaul@yahoo.com

ABSTRACT

Understanding customer sentiments is of paramount importance in marketing strategies today. Not only will it give companies an insight as to how customers perceive their products and/or services, but it will also give them an idea on how to improve their offers. This paper attempts to understand the correlation of different variables in customer reviews on a women clothing e-commerce, and to classify each review whether it recommends the reviewed product or not and whether it consists of positive, negative, or neutral sentiment. To achieve these goals, we employed univariate and multivariate analyses on dataset features except for review titles and review texts, and we implemented a bidirectional recurrent neural network (RNN) with long-short term memory unit (LSTM) for recommendation and sentiment classification. Results have shown that a recommendation is a strong indicator of a positive sentiment score, and vice-versa. On the other hand, ratings in product reviews are fuzzy indicators of sentiment scores. We also found out that the bidirectional LSTM was able to reach an F1-score of 0.88 for recommendation classification, and 0.93 for sentiment classification.

CCS CONCEPTS

• **Information systems → Data analytics**; **Sentiment analysis**; **Business intelligence**; *Data cleaning*; • **Computing methodologies → Natural language processing**; **Supervised learning by classification**; **Neural networks**;

KEYWORDS

artificial intelligence; artificial neural networks; classification; data analytics; data science; data visualization; deep learning; e-commerce; long short term memory; natural language processing; recurrent neural networks; sentiment classification; supervised learning

1 INTRODUCTION

Companies are starting to turn to social media listening as a tool for understanding their customers, in order to further improve their products and/or services. As a part of this movement, text analysis has become an active field of research in computational linguistics and natural language processing.

One of the most popular problems in the mentioned field is text classification, a task which attempts to categorize documents to one or more classes that may be done manually or computationally. Towards this direction, recent years have shown top interest in classifying *sentiments* of statements found in social media, review sites, and discussion groups. This task is known as *sentiment analysis*, a computational process that uses statistics and natural language processing techniques to identify and categorize opinions expressed in a text, particularly, to determine the polarity of attitude (positive, negative, or neutral) of the writer towards a topic or a product[12]. The said task is now widely used by companies for understanding their clients through their customer support in social media, or through their review boards online.

In this paper, we attempt to analyze the customer reviews on women clothing e-commerce[3] by employing statistical analysis and sentiment classification. We first analyze the non-text review features (e.g. age, class of dress purchased, etc.) found in the dataset, as an attempt to unravel any connection between them and customer recommendation on the product. Then, we implement a bidirectional recurrent neural network (RNN) with long-short term memory (LSTM)[6] for classifying whether a review text recommends the purchased product or not, and for classifying the user review sentiment towards the product.

2 METHODOLOGY

2.1 Machine Intelligence Library

Keras[4] with Google TensorFlow[1] was used to implement the bidirectional recurrent neural network (RNN) with long-short term memory (LSTM)[6] in this study. As for the data preprocessing and handling, the `numpy`[13] and `pandas`[9] Python libraries were used. Lastly, for the data visualization, the `matplotlib`[7] and `seaborn`[14] Python libraries were used.

2.2 The Dataset

The Women's Clothing E-Commerce Reviews[3] was used as the dataset for this study. This dataset consists of reviews written by real customers, hence it has been anonymized, i.e. customer names were not included, and references to the company were replaced with "retailer" by [3].

Table 1 shows the frequency distribution of dataset features and label (`Recommended IND`).

2.3 Data Analysis

In which we analyze the dataset features and their implications on user recommendation and review sentiments. This subsection covers four statistical analyses. Table 2 shows a summary of statistical description of the dataset.

*Student researcher
†Business Intelligence course adviser

Table 1: Frequency Distribution of Dataset Features.

Feature	Unique Count
Clothing ID	1172
Age	77
Title	13984
Review Text	22621
Rating	5
Recommended IND	2
Positive Feedback Count	82
Division Name	3
Department Name	6
Class Name	20

Table 2: Summary of Statistical Description of Dataset Features.

Feature	Mean	Standard Deviation	Type
Clothing ID	919.695908	201.683804	Integer
Age	43.282880	12.328176	Integer
Rating	4.183092	1.115911	Categorical
Recommended	0.818764	0.385222	Categorical
Positive Feedback	2.631784	5.787520	Integer

2.3.1 *Analysis on Univariate Distributions.*

(1) **Age and Positive Feedback Count.** Figure 5 reveals that the most engaged customers in reviewing purchased products were in the age range of 35 to 44. In addition, the figure suggests that they have the most positive reviews on their purchased products. From this, we have two points to consider: (1) the said age group is the most satisfied group in the range of customers, thus, the e-commerce at review must focus on maintaining this segment, and (2) the e-commerce entity can explore why other age groups are comparatively less satisfied than the age group 35 to 44.

(2) **Department Name and Division Name.** Figure 6 shows the frequency distribution of customer reviews per *department* and per *division*. This gives the e-commerce an insight on the customer apparel sizes and clothing being most reviewed, i.e. *General* which refers to clothing size, and *tops* which refers to apparel types.

(3) **Top 60 Clothing ID.** Figure 7 shows the IDs of top 60 reviewed apparel from the e-commerce. The apparels with clothing IDs 1078, 862, and 1094 belong to the *general* division and *dresses* apparely type, with a positive title review of "Beautiful dress" as per [2].

(4) **Class Name.** Figure 8 shows the frequency distribution of apparel classes most reviewed. The top three apparels are *dresses*, *knits*, and *blouses*.

(5) **Rating, Recommendation, and Label.** Figure 9 shows that the dominant reviews were positive, suggesting that the e-commerce fairly satisfies its customers. It may be axiomatic that a review with recommendation implies a higher rating and a positive sentiment. But then again, the processing of sentiments were based on a threshold of higher than rating of 3 for positive, and negative for the rest. We shall look more into this in subsection 2.4.2.

(6) **Word Length.** Figure 10 shows that regardless of the rating in a review, apparel type, or recommendation, the users had qualitatively the same length of words in their reviews.

2.3.2 *Analysis on Multivariate Distributions.*

(1) **Division Name by Department Name.** Figure 11 reveals the dominance of general-sized tops, while Figure 12 supports this inference.

(2) **Class Name by Department Name.** Figure 13 reveals the dominance of dress among apparel types, and supported by Figure 14.

(3) **Class Name by Division Name.** Figure 15 reveals the most reviewed apparel types as general-sized blouses, dresses, and knits. However, 16 shows that most reviews on dresses are from general petite sizes.

(4) **Age by Positive Feedback Count.** Figure 17 shows a small correlation between age and the positive feedback in a review. Based on the figure, the same age group of 35 to 44 seems to be the group that gave most of the positive feedbacks.

(5) **Recommendation by Department Name and Division Name.** Figure 18 corroborates the findings in Figure 11.

(6) **Rating by Department Name and Division Name.** Figure 19 shows consistency in rating distribution.

(7) **Rating by Recommendation.** Figure 20 supports the assumption that a review rating mirrors its recommendation status, i.e. higher rating means recommendation and vice-versa.

2.3.3 *Multivariate Analysis and Descriptive Statistics.*

(1) **Average Rating by Recommendation.** Figure 21 shows consistency on recommendation and rating, i.e. when review has recommendation, the rating is under maximum value of rating; when review has no recommendation, the rating is halved.

(2) **Average Rating and Recommendation by Clothing ID Correlation.** Figure 22 attempts to look at the correlation, if there is any, between the average rating of a product and number of reviews for a product, that is grouped by clothing ID. The correlation matrix suggests there is no such correlation between the variables considered, but it did reveal a relatively strong correlation of 0.8 between rating and recommendation. The mentioned correlation coefficient further substantiates the assumption on connection between rating and recommendation.

2.3.4 *Word Frequency Distributions.*

(1) **Titles.** Figure 23 gives us the most frequent words in a review title. Only the word "flaws" seems to indicate a negative review, but then again, this does not necessarily indicate that the entire product review has a negative sentiment. Take not that this word cloud only accounts for the frequency of words in titles, and does not account for phrases. In other words, there may be counter-words for negative word indicators, but only failed to make it into the word cloud. The same may

be said for positive word indicators in the word cloud, as it does not include any negators if there are any.

(2) **Most Frequent Words in Highly-rated Comments.** Since Figure 24 is a word cloud for reviews with high ratings, it may be assumed that the words in this figure reflects what are written in their respective reviews.
(3) **Most Frequent Words in Low-rated Comments.** Since Figure 25 is a word cloud for reviews with low ratings, it may be assumed that the words in this figure reflects what are written in their respective reviews.
(4) **Word Clouds for Division Names.** Figure 26 shows the most frequent words in product reviews from the "intimates" division; Figure 27 for "general" division; and Figure 28 for "general petite" division. Further investigation on these word clouds may reveal some useful insight on customer acceptability per division.

2.4 Dataset Preprocessing

2.4.1 Text Cleaning. The user review texts were cleaned by eliminating delimiters such as \n and \r found in the texts.

2.4.2 Sentiment Analysis. Instead of manually tagging the review texts, the sentiment analyzer of NLTK[8] was used to automate the process. Thus, leaving behind the intuitive tagging of review texts that had a threshold of rating 3, i.e. if a review rating is greater than or equal to 3, it is considered as a positive feedback, otherwise it is considered a negative feedback. The mentioned manual, intuitive tagging had the flaw of not taking into account some neutral sentiments. Hence, the use of sentiment analyzer by NLTK[8]. See Figure 29 for the frequency distribution of sentiments per recommendation.

2.4.3 Word Embeddings. The GloVe word embeddings[11] were used to map the words in review texts to the vector space.

2.5 Machine Learning

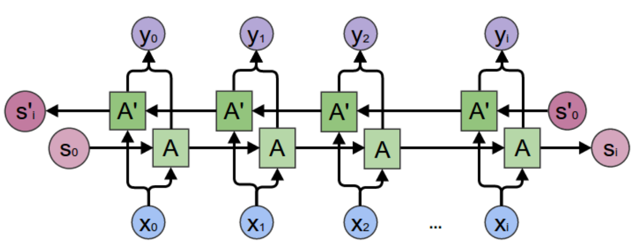

Figure 1: Image from [10]. Computation of a conventional Bidirectional RNN maps input sequences x to target sequences y, with loss $L(t)$ at each time step t. The RNN cells s propagate information forward in time (towards the right) while the RNN cells s' propagate information backward in time (towards the left). Thus at each time step t, the output units $o(t)$ (before applying an activation function to get y) can benefit from a relevant summary of the past in its $s(t)$ input, and from a relevant summary of the future in its $s'(t)$ input.

Given that the problem at hand is a classification task on sentiments, the most appropriate ML algorithm to implement is a recurrent neural network (RNN). However, from literature, we know that a vanilla RNN suffers from *vanishing gradients*. Hence, we used the RNN with long-short term memory (LSTM) units, which was designed to solve the mentioned problem[6]. Furthermore, to better capture the context of words in the review texts, we employed a bidirectional RNN with LSTM (see Figure 1). That is, the model has the capability to learn the context from "past" to the "future" of a text sequence and vice-versa[5]. In turn, giving the model more insight on each review text.

Below are the LSTM gate equations[6], which we implemented using Google TensorFlow[1].

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{1}$$
$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{2}$$
$$\tilde{C}_t = tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{3}$$
$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{4}$$
$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{5}$$
$$h_t = o_t * tanh(C_t) \tag{6}$$

where f is the *forget gate*, which "forgets" non-essential information for the model; i is the *input gate*, which accepts new data input at a given time step s_t; $\tilde{C}$ is the candidate cell state value of each LSTM cell; C is the cell state value to be passed onto the next RNN-LSTM cell; o is the *output gate* which decides what the cell state will output; and h is the cell state output from cell state value and the decided output.

We employed this machine learning model on two text classification problems on the dataset: (1) recommendation classification, which determines whether a review text recommends the reviewed product, and (2) sentiment classification, which determines the tone of the review text towards the purchased product.

2.5.1 Recommendation Classification. A product review has two recommendation states: (1) recommended, and (2) not recommended – a binary classification problem.

2.5.2 Sentiment Classification. A product review has three sentiment states: (1) negative, (2) neutral, and (3) positive – a multinomial classification problem.

3 RESULTS AND DISCUSSION

All experiments in this study were conducted on a laptop computer with Intel Core(TM) i5-6300HQ CPU @ 2.30GHz x 4, 16GB of DDR3 RAM, and NVIDIA GeForce GTX 960M 4GB DDR5 GPU. The dataset was partitioned into a 60/20/20 fashion, i.e. 60% for training dataset, 20% for validation dataset, and 20% for testing dataset.

Table 3 shows the hyper-parameters used by the Bidirectional RNN-LSTM in the experiments, these hyper-parameters were arbitrarily chosen as hyper-parameter tuning implies more computational cost requirement. Table 4 shows the test accuracy and test loss by the Bidirectional RNN-LSTM on both recommendation classification and sentiment classification experiments.

However, take note that the frequency distributions for classes in *recommendation* and *sentiment* are both imbalanced, i.e. there are more *recommended* classes than *not recommended*, and there

Table 3: Hyper-parameters used in Bidirectional RNN-LSTM

Hyper-parameter	Value
Batch Size	256
Cell Size	256
Dropout Rate	0.50
Epochs	32
Learning Rate	1e-3

Table 4: Test Accuracy and Test Loss using Bidirectional LSTM.

Task	Test Accuracy	Loss
Recommendation Classification	≈0.882678	≈0.572342
Sentiment Classification	≈0.928414	≈0.453205

are more *positive* sentiments than there are *negative* and *neutral* combined. This poses a problem as the model shall grow a biased classification towards the class with highest frequency distribution. Hence, we take a look at the statistical report on *recommendation classification* at Table 5.

Table 5: Statistical Report on Recommendation Classification using Bidirectional LSTM.

Class	Precision	Recall	F1-Score	Support
(0) Not Recommended	0.70	0.65	0.68	847
(1) Recommended	0.92	0.94	0.93	3679
Average / Total	0.88	0.88	0.88	4526

Table 5 shows a relatively weaker predictive performance for negative class in the *recommendation classification* problem, as it can also be seen in the confusion matrix in Figure 2 (where 0 represents *not recommended* class, and 1 represents *recommended* class), thus supporting our claim above. To look at the model performance on a relatively fair scheme, we take a look at the ROC curve for the result (see Figure 4).

Table 6: Statistical Report on Sentiment Classification using Bidirectional LSTM.

Class	Precision	Recall	F1-Score	Support
(0) Negative	0.47	0.50	0.49	289
(1) Neutral	0.31	0.18	0.23	22
(2) Positive	0.96	0.96	0.96	4215
Average / Total	0.93	0.93	0.93	4526

Table 6 corroborates our findings on biased classification towards the class with highest frequency distribution, supported by the confusion matrix in Figure 3 (where 0 represents the *negative* class, 1 represents the *neutral* class, and 2 represents the *positive* class).

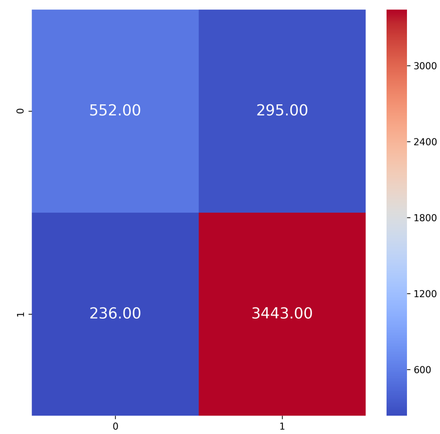

Figure 2: The confusion matrix on recommendation classification.

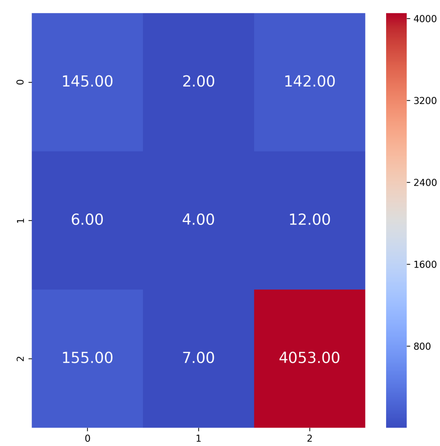

Figure 3: The confusion matrix on sentiment classification.

We can see in this report that the model had a relatively weaker predictive performance for the negative and neutral sentiments.

The empirical evidences presented in this paper indicates a relatively high-performing predictive performance on both recommendation classification and sentiment classification, and this is despite the imbalanced class frequency distribution in the dataset. Such result supports the claim that using Bidirectional RNN-LSTM better captures the context of review texts which leads to better predictive performance. However, to further substantiate this claim, we recommend employing a uni-directional RNN-LSTM on the same classification problems for fair comparison.

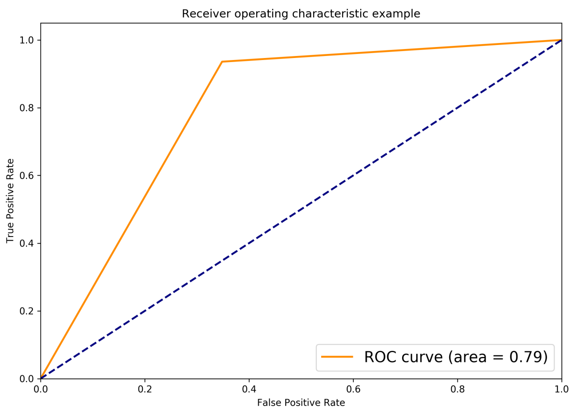

Figure 4: The ROC Curve for binary classification on recommendation indicator.

4 CONCLUSION AND RECOMMENDATION

To further improve the model, hyper-parameter tuning must be performed. This study was limited to an arbitrarily-chosen hyper-parameters due to computational cost restrictions. In addition, k-fold cross validation may give us a better and/or additional insight on the predictive performance of the model.

Despite the limitations on the experiment for this study, it may be inferred that the Bidirectional RNN-LSTM model exhibited high performance (with F1-score of 0.88 for recommendation classification, and 0.93 for sentiment classification). Furthermore, the statistical measures on the classification problem may also be deemed satisfactory.

5 ACKNOWLEDGMENT

A sincere appreciation is given to Nick Brooks for his dataset on *Women's Clothing E-Commerce Reviews*[3], and also for granting us the permission to utilize some of his scripts for data visualizations[2].

REFERENCES

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. (2015). https://www.tensorflow.org/ Software available from tensorflow.org.

[2] Nick Brooks. 2018. Guided Numeric and Text Exploration E-Commerce. (2018). https://www.kaggle.com/nicapotato/guided-numeric-and-text-exploration-e-commerce

[3] Nick Brooks. 2018. Women's E-Commerce Clothing Reviews. (2018). https://www.kaggle.com/nicapotato/womens-ecommerce-clothing-reviews

[4] François Chollet et al. 2015. Keras. https://github.com/keras-team/keras. (2015).

[5] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press. http://www.deeplearningbook.org.

[6] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[7] J. D. Hunter. 2007. Matplotlib: A 2D graphics environment. *Computing In Science & Engineering* 9, 3 (2007), 90–95. https://doi.org/10.1109/MCSE.2007.55

[8] Edward Loper and Steven Bird. 2002. NLTK: The Natural Language Toolkit. In *In Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics. Philadelphia: Association for Computational Linguistics.*

[9] Wes McKinney. 2010. Data Structures for Statistical Computing in Python. In *Proceedings of the 9th Python in Science Conference*, Stéfan van der Walt and Jarrod Millman (Eds.). 51 – 56.

[10] Christopher Olah. 2015. Neural Networks, Types, and Functional Programming. (2015). http://colah.github.io/posts/2015-09-NN-Types-FP/

[11] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*. 1532–1543. http://www.aclweb.org/anthology/D14-1162

[12] EMC Education Services. 2015. *Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data.* John Wiley & Sons, Inc.

[13] Stéfan van der Walt, S Chris Colbert, and Gael Varoquaux. 2011. The NumPy array: a structure for efficient numerical computation. *Computing in Science & Engineering* 13, 2 (2011), 22–30.

[14] Michael Waskom, Olga Botvinnik, Drew O'Kane, Paul Hobson, Saulius Lukauskas, David C Gemperline, Tom Augspurger, Yaroslav Halchenko, John B. Cole, Jordi Warmenhoven, Julian de Ruiter, Cameron Pye, Stephan Hoyer, Jake Vanderplas, Santi Villalba, Gero Kunter, Eric Quintero, Pete Bachant, Marcel Martin, Kyle Meyer, Alistair Miles, Yoav Ram, Tal Yarkoni, Mike Lee Williams, Constantine Evans, Clark Fitzgerald, Brian, Chris Fonnesbeck, Antony Lee, and Adel Qalieh. 2017. mwaskom/seaborn: v0.8.1 (September 2017). (Sept. 2017). https://doi.org/10.5281/zenodo.883859

Appendix

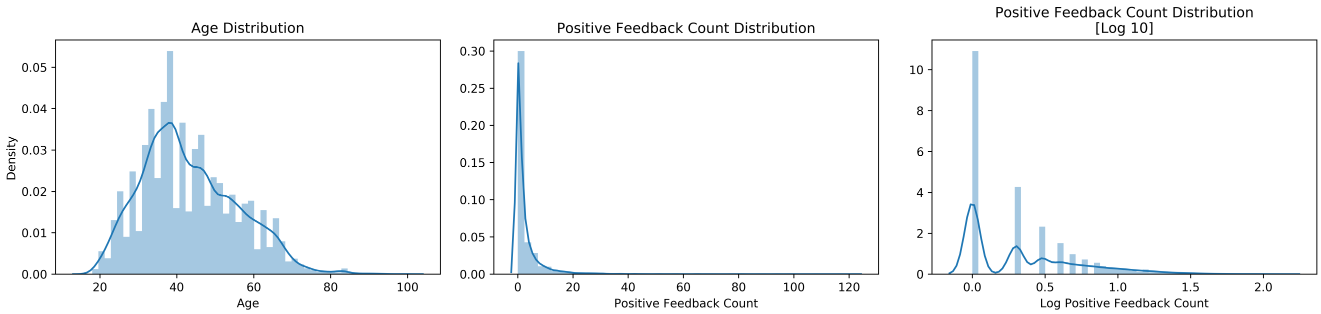

Figure 5: Plot generated based on script by [2]. The frequency distribution of customer age and positive feedback.

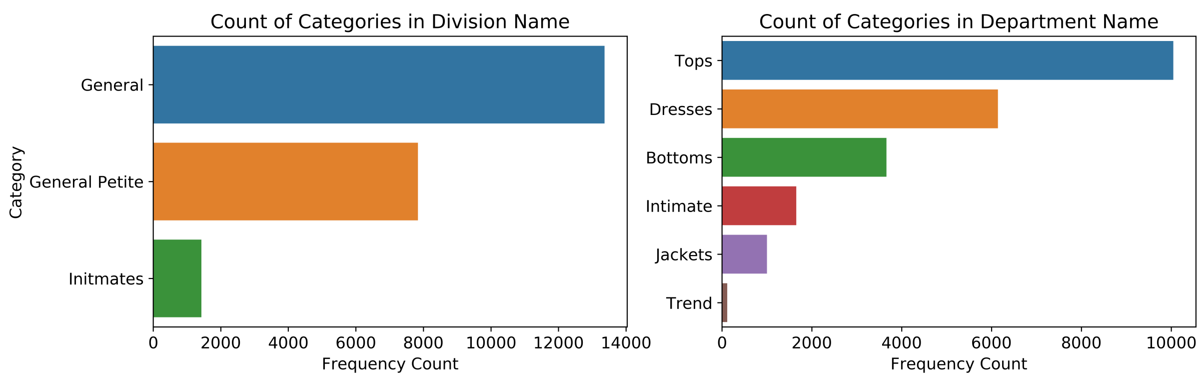

Figure 6: Plot generated based on script by [2]. The frequency distribution of apparels per *division* and *department*.

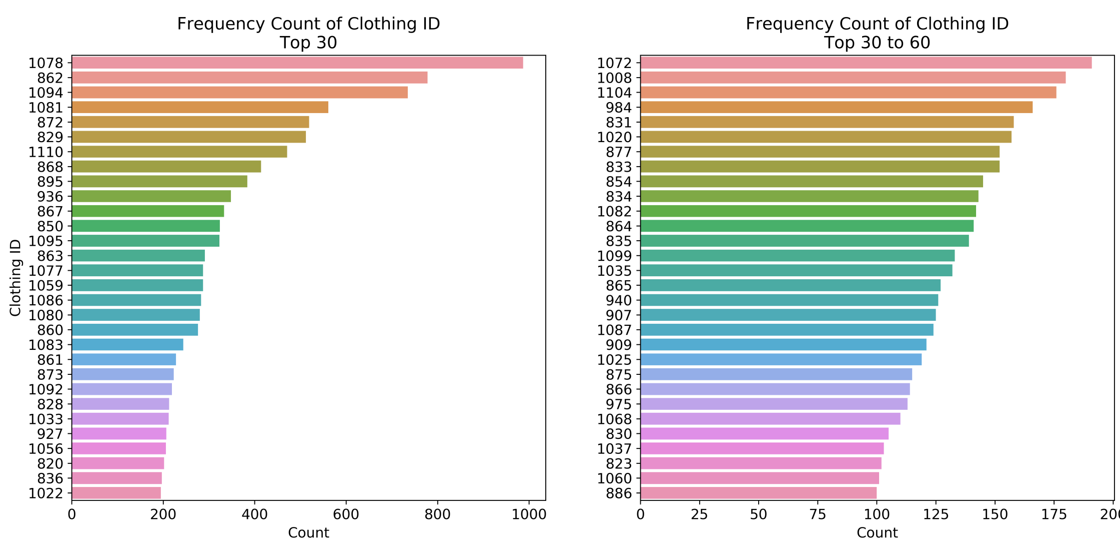

Figure 7: Plot generated based on script by [2]. The frequency distribution of top 60 apparels per *clothing ID*.

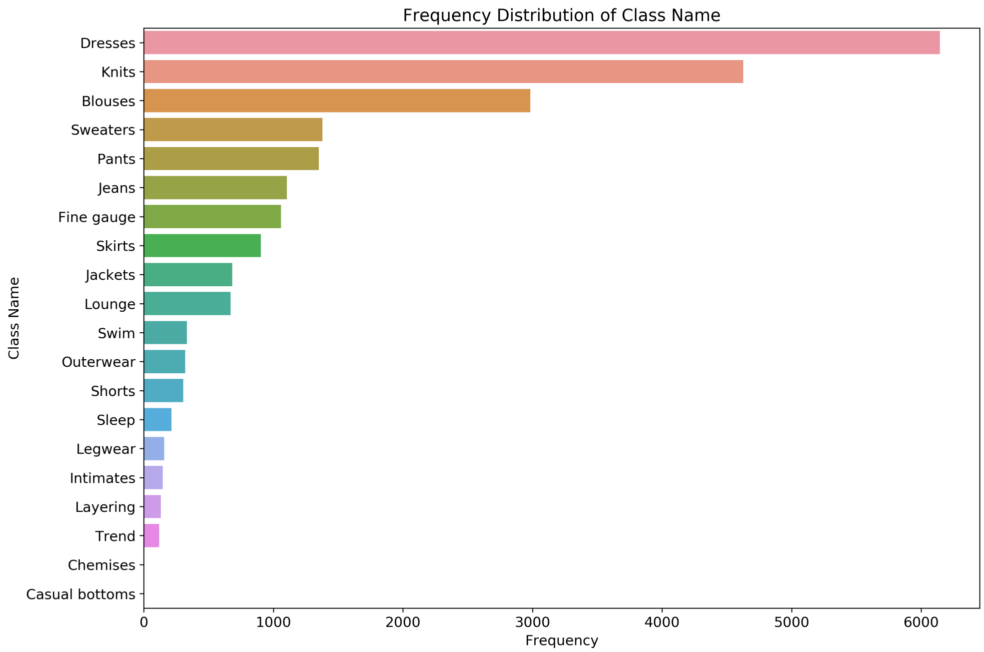

Figure 8: Plot generated based on script by [2]. The frequency distribution of apparels per *class*.

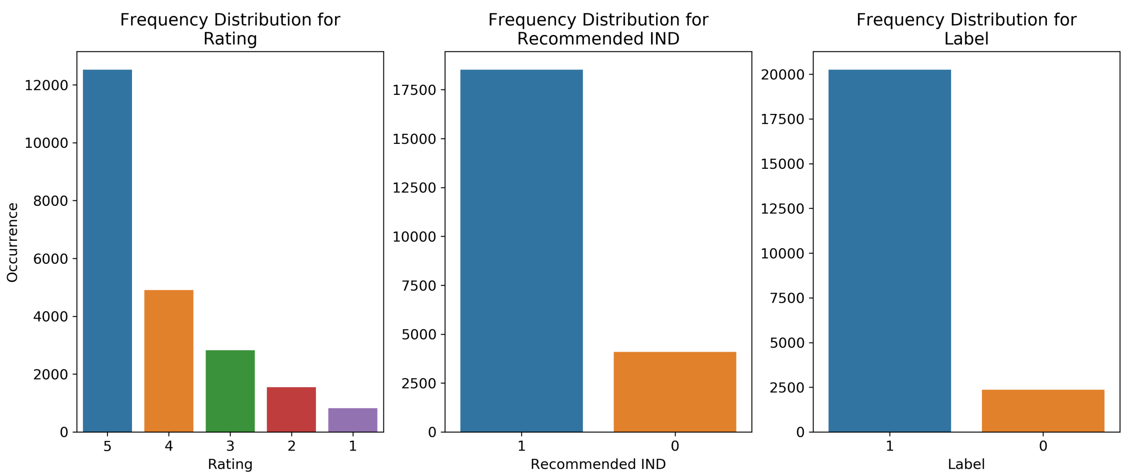

Figure 9: Plot generated based on script by [2]. The frequency distribution of review ratings, recommendation, and labels.

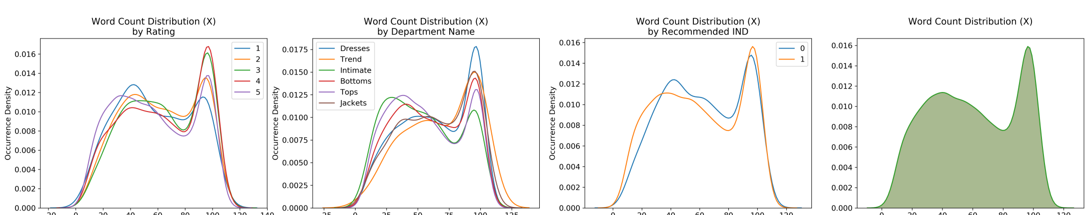

Figure 10: Plot generated based on script by [2]. The word frequency distrubtion in review texts per rating, department, and recommendation.

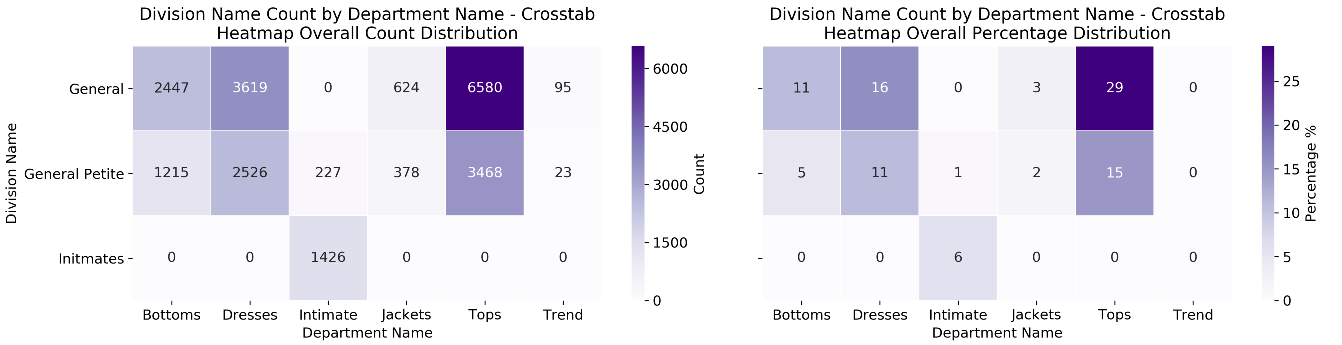

Figure 11: Heatmap generated based on script by [2]. The cross tabulation for apparel per *division* and *department*.

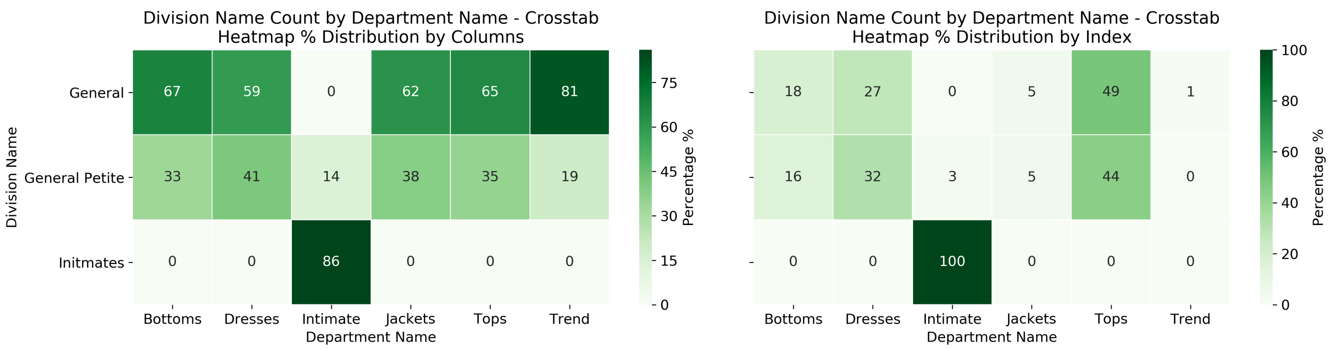

Figure 12: Heatmap generated based on script by [2]. The normalized cross tabulation for apparel per *division* and *department*.

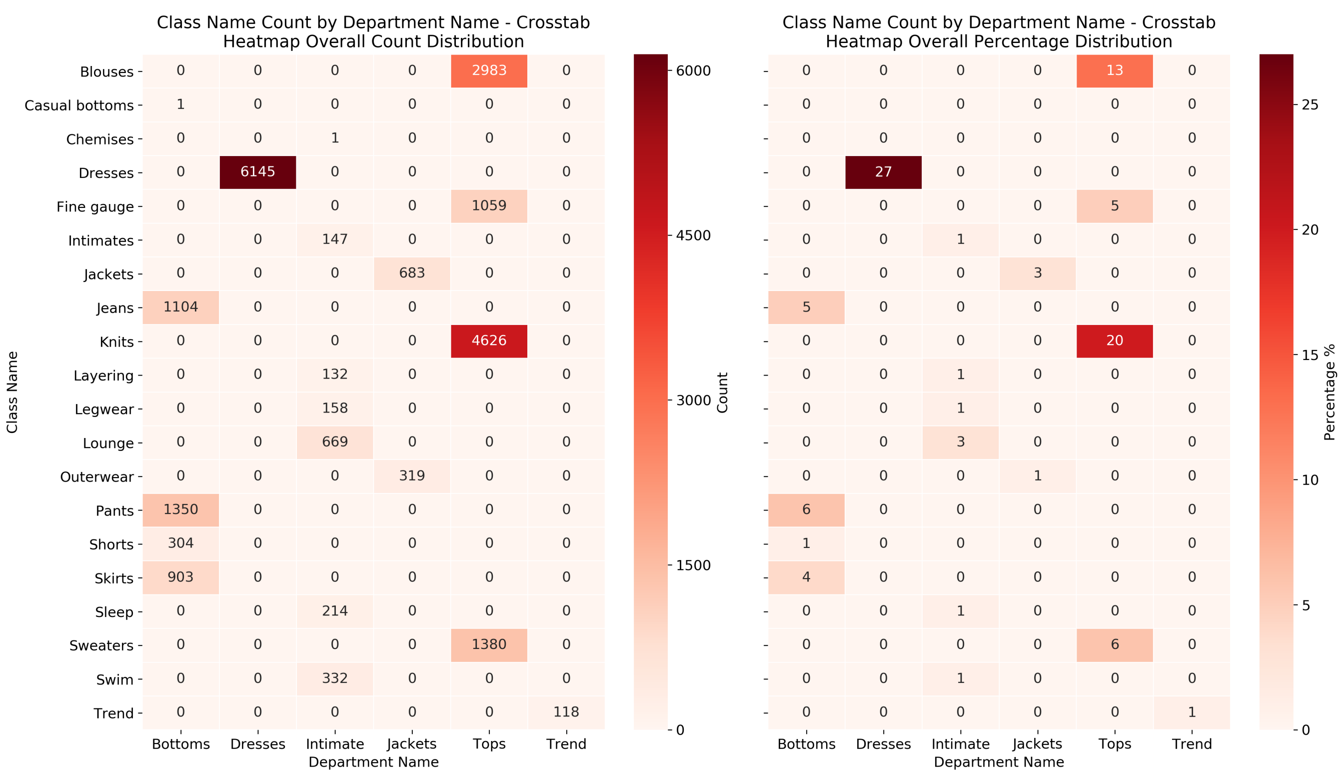

Figure 13: Heatmap generated based on script by [2]. The cross tabulation for apparel per *class* and *department*.

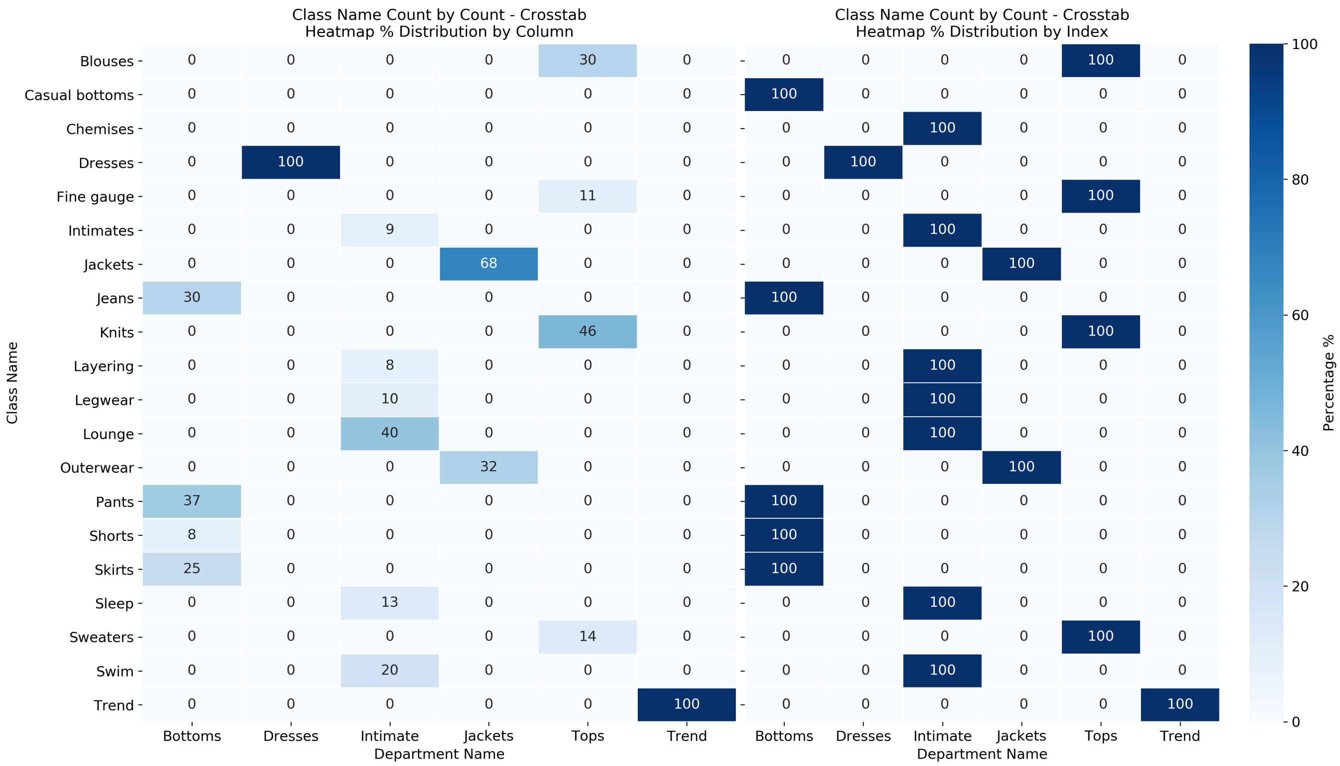

Figure 14: Heatmap generated based on script by [2]. The normalized cross tabulation for apparel per *class* and *department*.

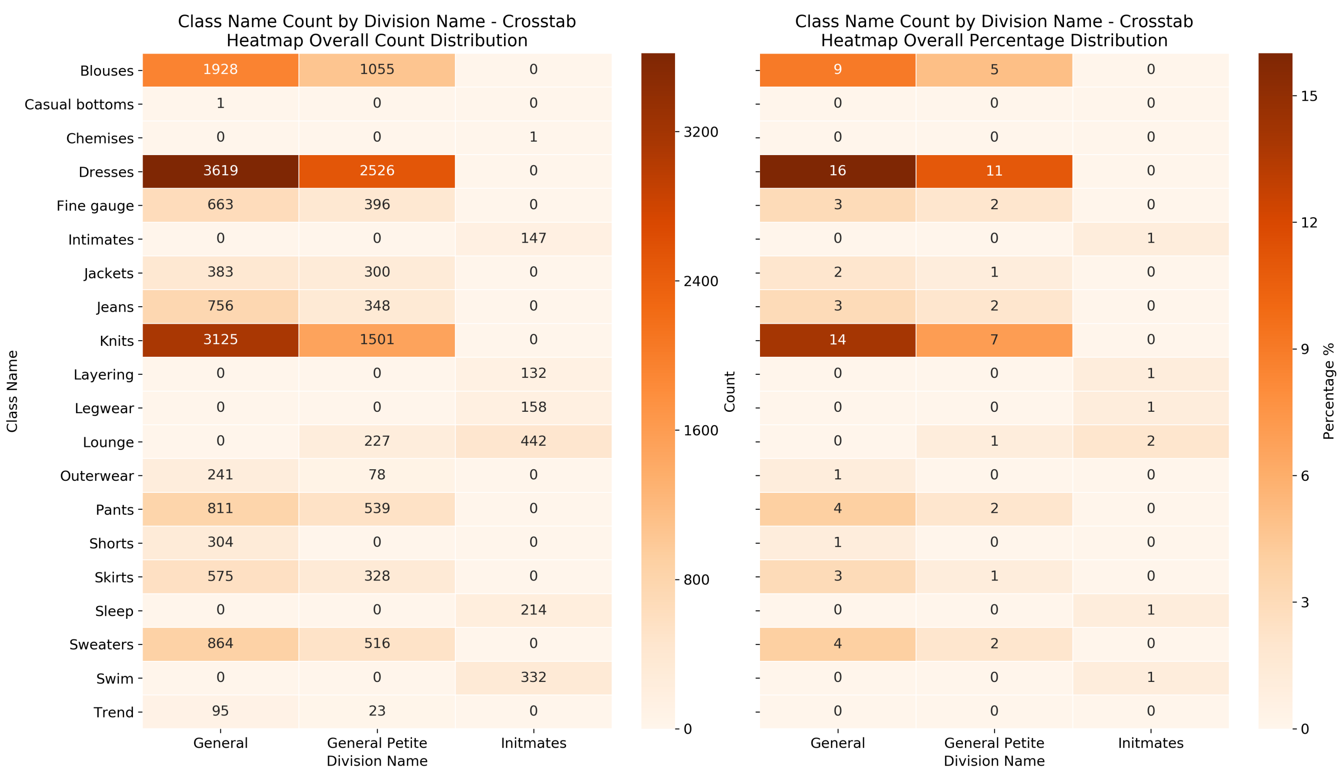

Figure 15: Heatmap generated based on script by [2]. The cross tabulation for apparel per *class* and *division*.

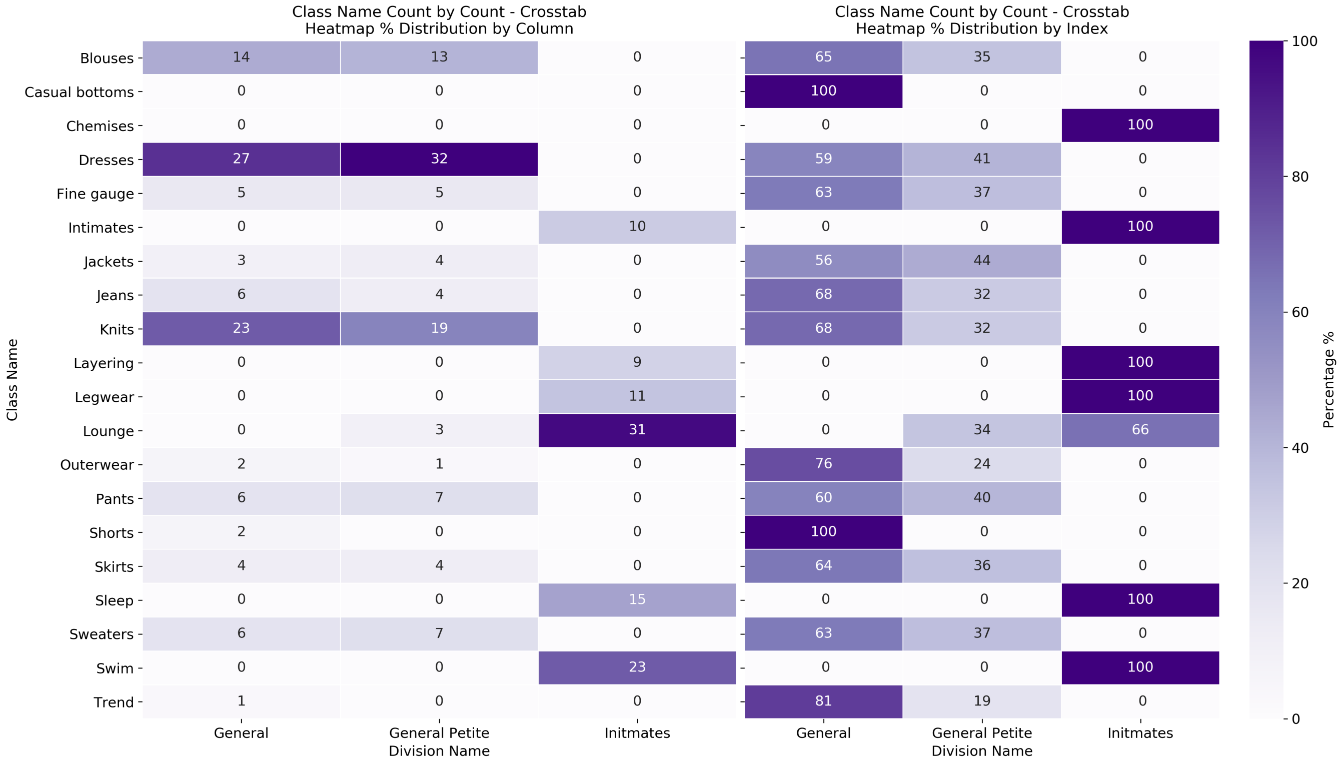

Figure 16: Heatmap generated based on script by [2]. The normalized cross tabulation for apparel per *class* and *division*.

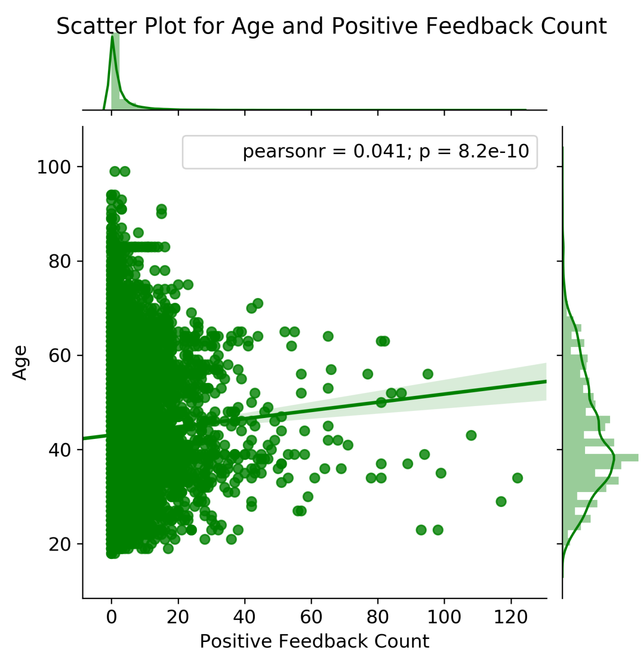

Figure 17: Scatter plot generated based on script by [2]. The scatter plot for age and positive feedback count.

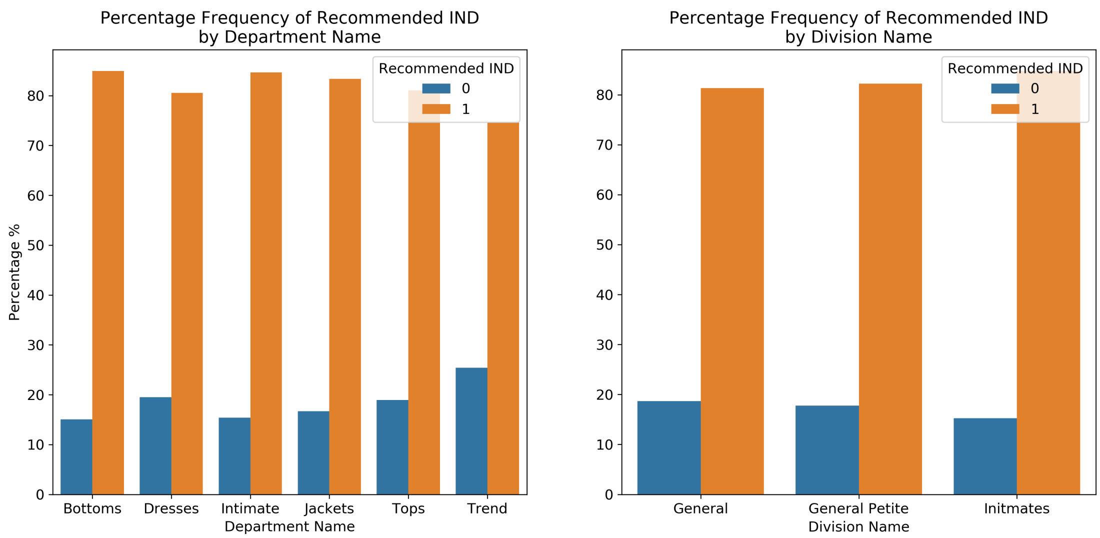

Figure 18: Plot generated based on script by [2]. The percentage frequency of recommendation indicator per review by *department* and *division*.

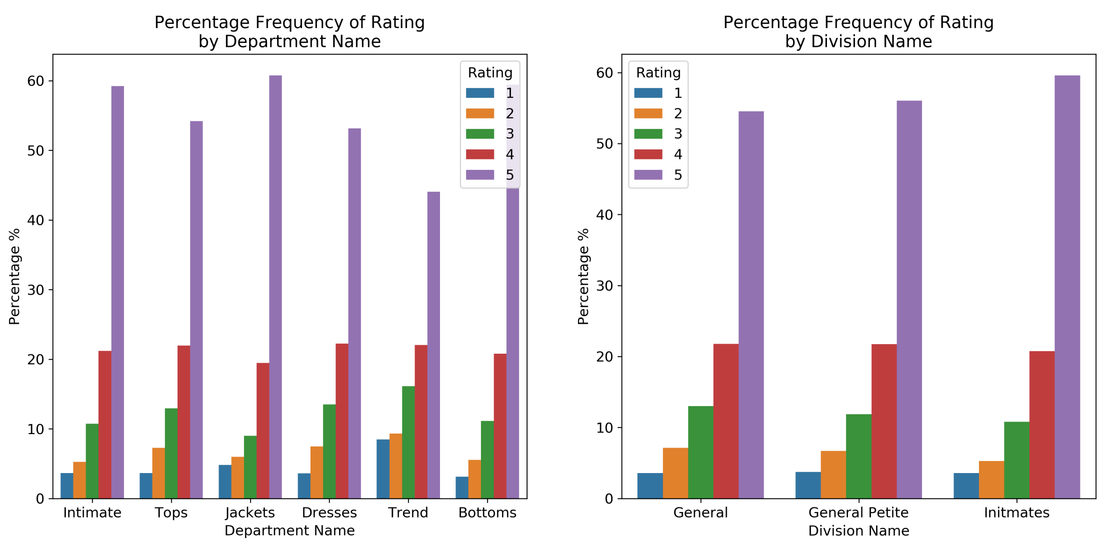

Figure 19: Plot generated based on script by [2]. The percentage frequency of review rating by *department* and *division*.

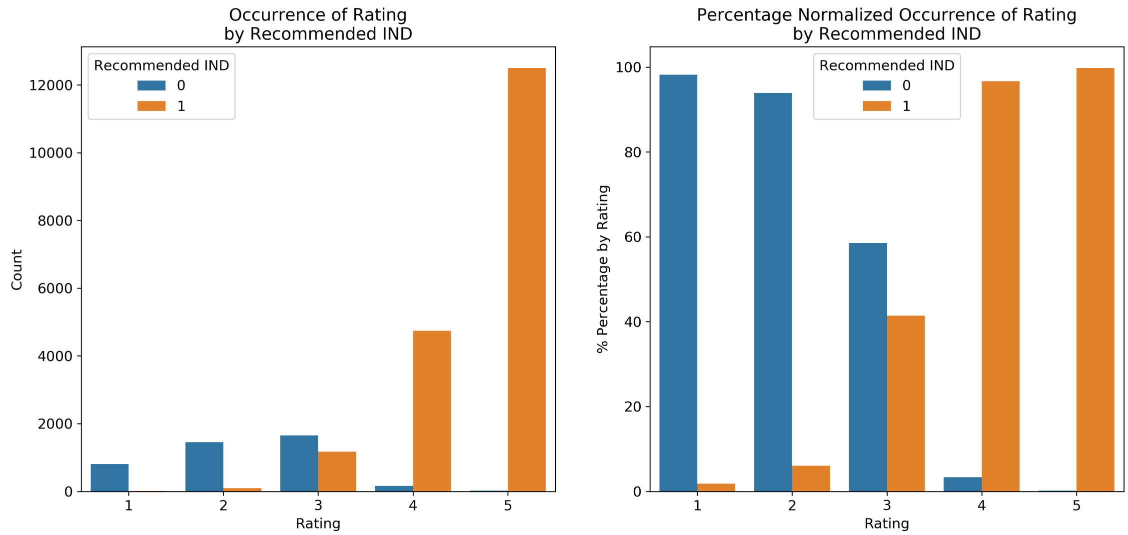

Figure 20: Plot generated based on script by [2]. The frequency of rating by recommendation indicator.

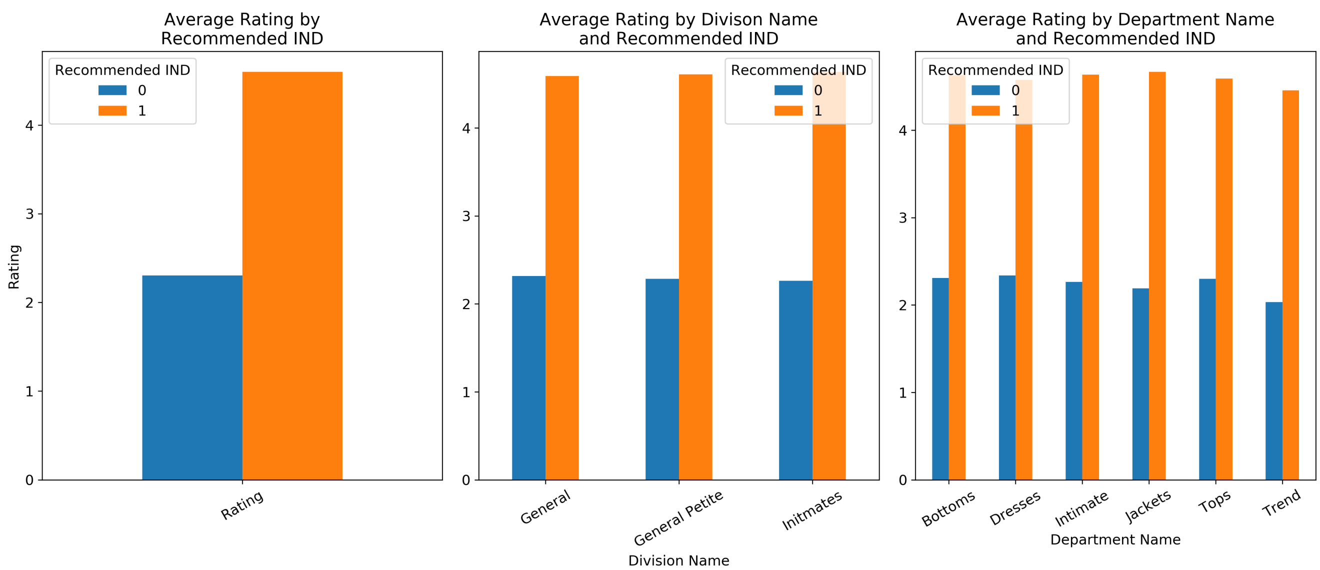

Figure 21: Plot generated based on script by [2]. The average rating frequency by *division department*, and recommendation indicator.

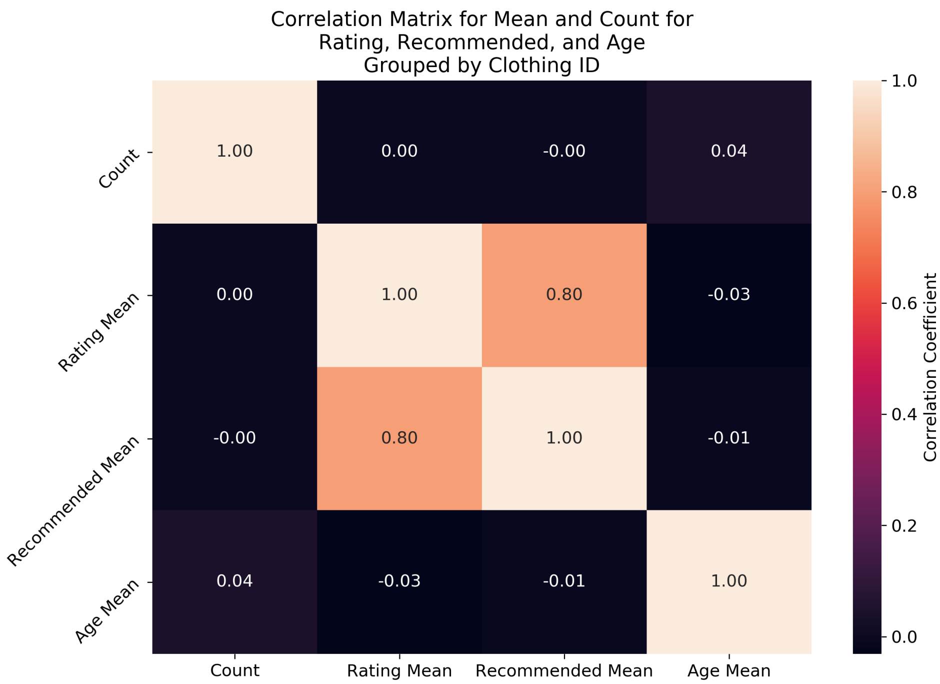

Figure 22: Heatmap generated based on script by [2]. The correlation matrix for average rating and recommendation indicator, grouped by clothing ID.

Figure 23: Word cloud generated based on script by [2]. The most frequent words used for review titles.

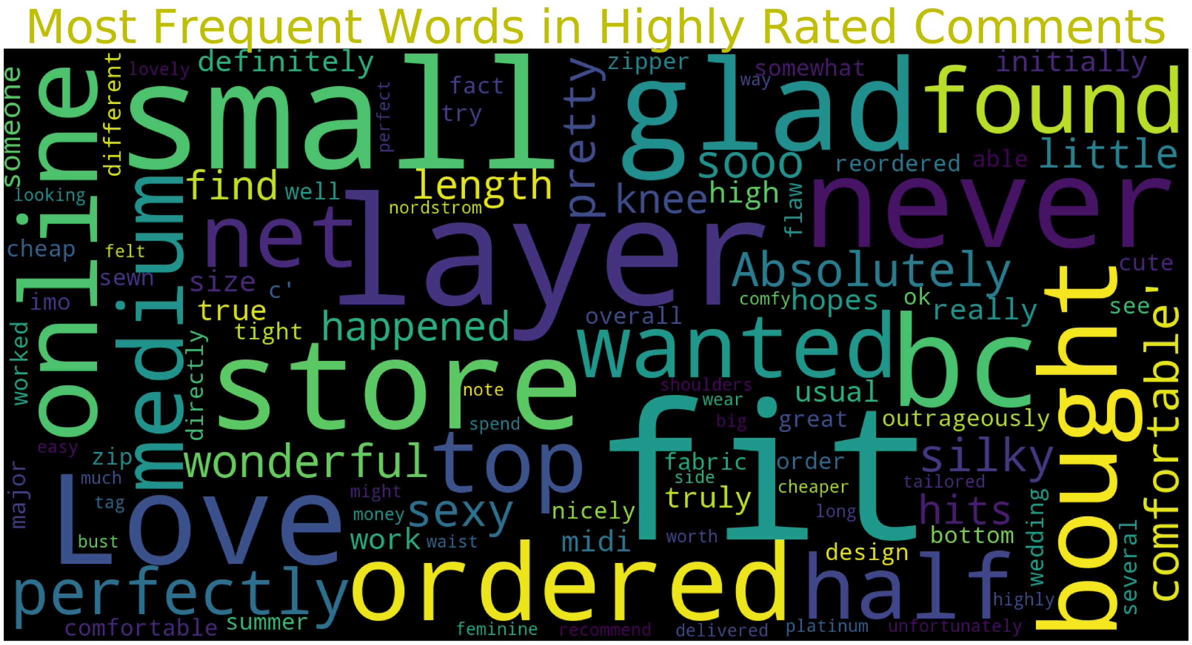

Figure 24: Word cloud generated based on script by [2]. The most frequent words used in review texts with high ratings.

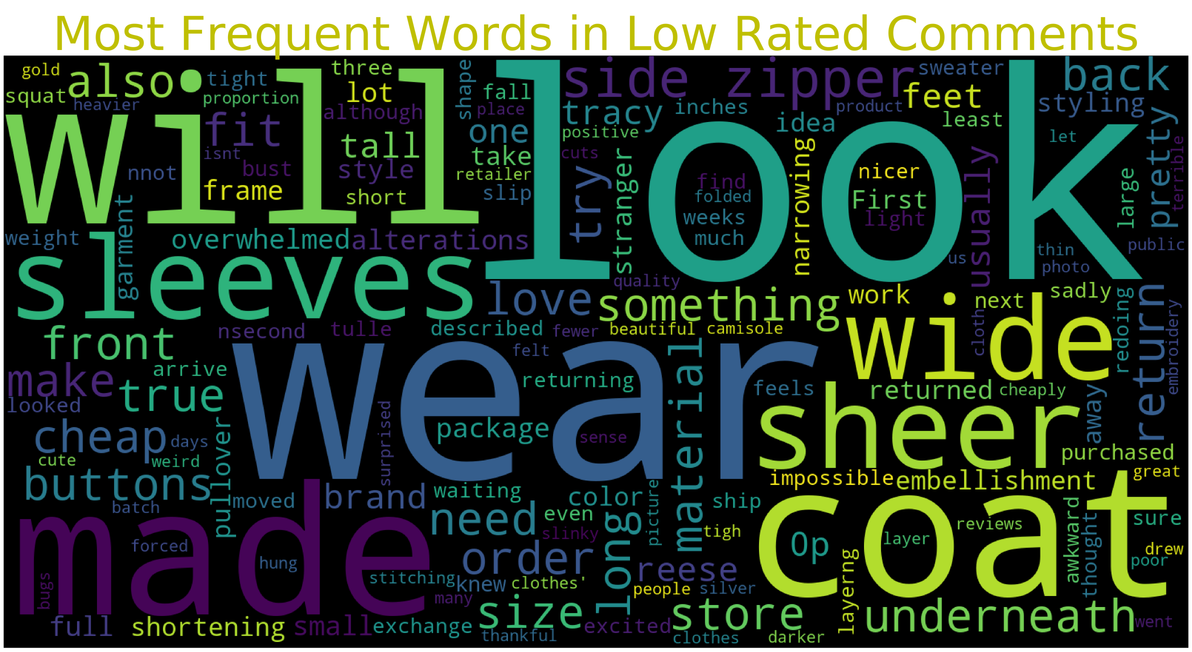

Figure 25: Word cloud generated based on script by [2]. The most frequent words used in review texts with low ratings.

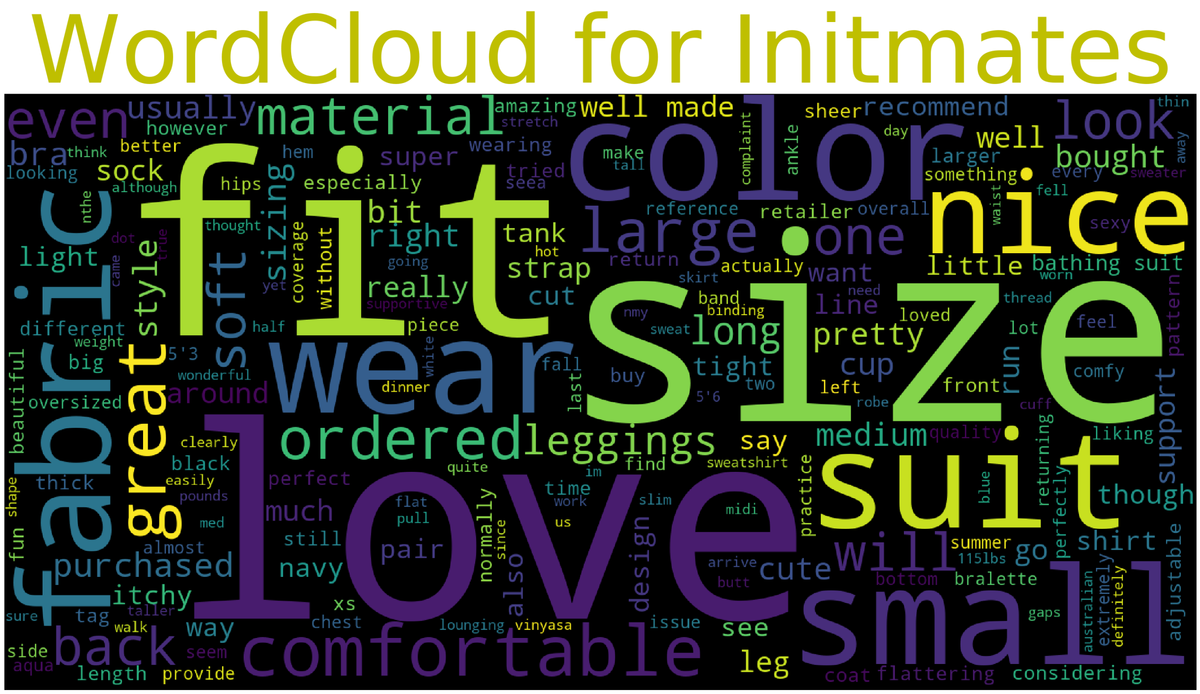

Figure 26: Word cloud generated based on script by [2]. The most frequent words used in review texts in *intimate* apparels.

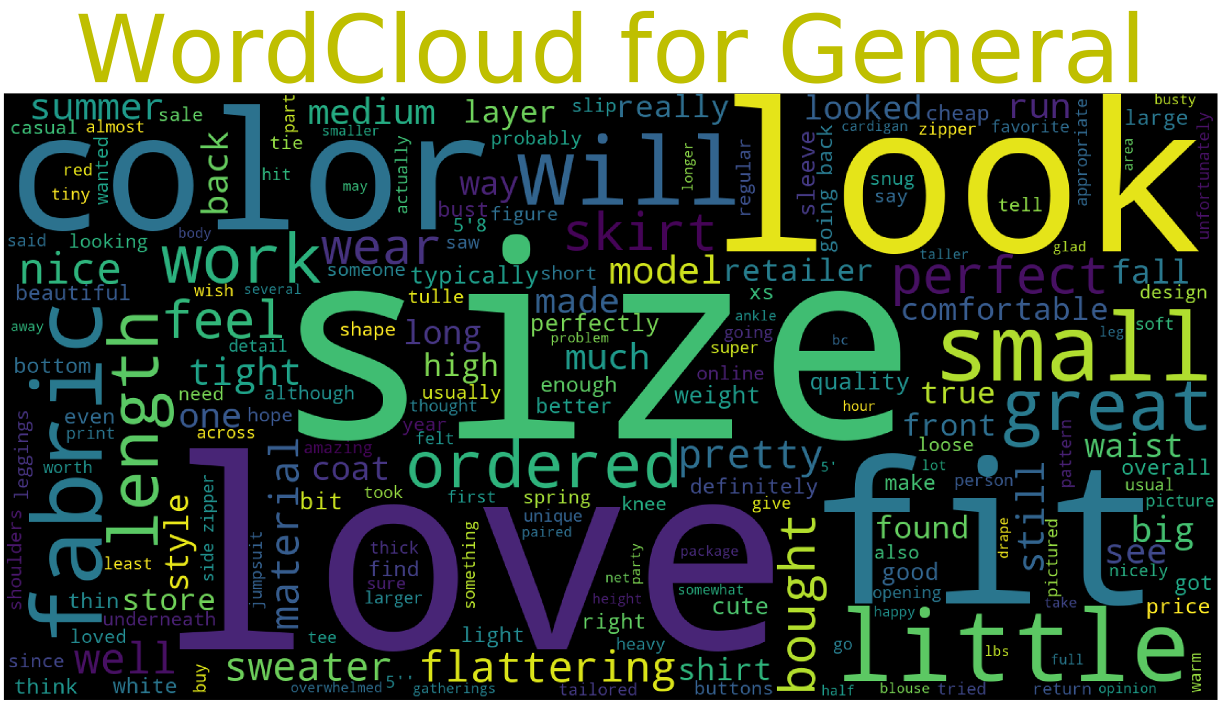

Figure 27: Word cloud generated based on script by [2]. The most frequent words used in review texts in *general-sized* apparels.

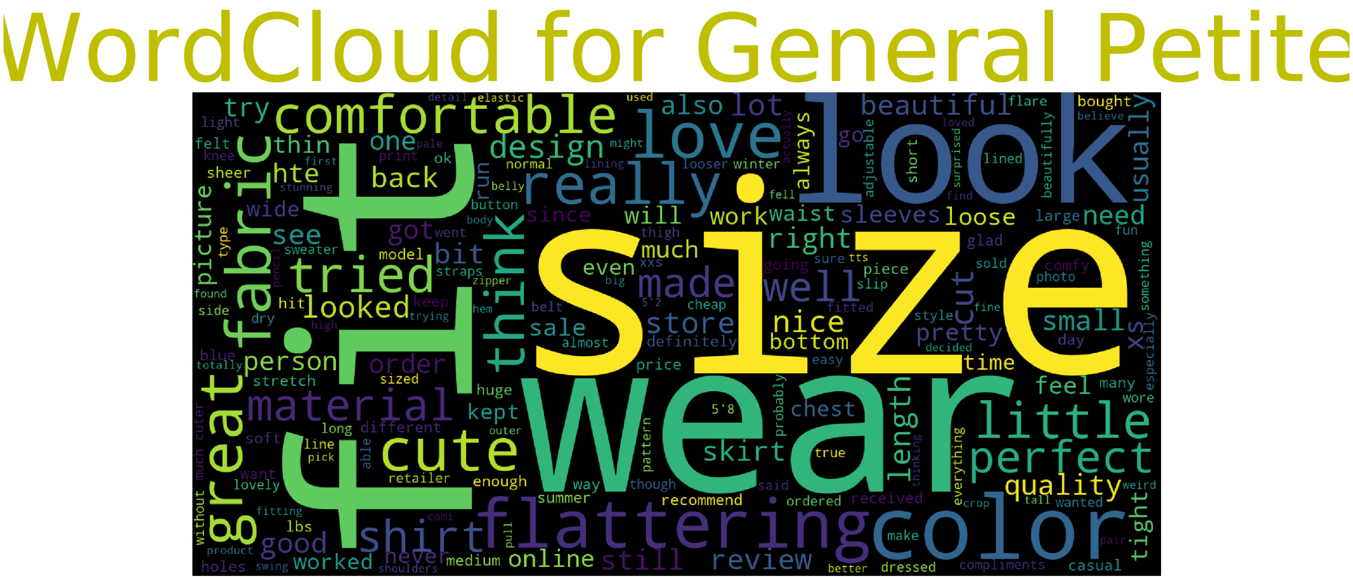

Figure 28: Word cloud generated based on script by [2]. The most frequent words used in review texts in *petite-sized* apparels.

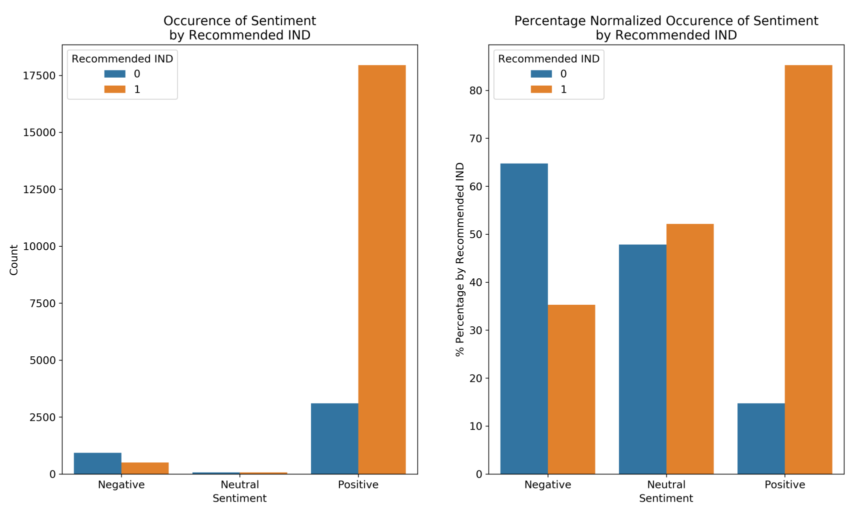

Figure 29: Plot generated based on script by [2]. The frequency distribution of sentiments per recommendation state in review texts.

List of Figures