Erasmus
School of
Economics

# TEXT ANALYTICS FOR MARKETING

## Assignment 2

EBD2 – Team 6

Kaan Güner Alkan – 575313

Narmin Qarazada – 428066

Victor de Andrade – 427707

Pham Anh Tuan – 532458

Erasmus University Rotterdam
Make it happen.

In the era of digitalization, e-commerce fashion companies are paying close attention to their consumers' feedback on their design and quality. Therefore, analyzing word embeddings and applying predictive analysis on reviews of the apparel products could generate more insights about which words or combinations of words might predict low rating or not-recommended products for further investigation and improvement.

The data set used in this report is called Women's Clothing E-Commerce from Kaggle. This is real commercial data that was anonymized. There are 23486 reviews in this data with 7 variables namely Clothing.ID, Review_ID (unique ID of each review), Age (of the reviewer), Review.Text, Rating (on a scale from 1 Worst to 5 Best), and Department.Name which contains 6 levels (Dresses, Tops, Intimate, Jacket, Bottoms, Trend), Recommended.IND (whether the reviewers want to recommend the product to others or not – 0 means no and 1 means yes).

For the preparation of part 1 (word embeddings) stemming is applied. The stop words are not removed, otherwise, the meaning of some word embeddings would not be clear. For part 2 (predictive analysis) blank reviews and reviews with unknown department names were removed from the data set. This left the Rating level 1,2,3,4, and 5 with the frequency of 821, 1549, 2823, 4908, 12527 respectively. Further investigation showed that almost all reviews with ratings of 4 and 5 had Recommended.IND = 1 and reviews with ratings of 1 and 2 had Recommended.IND = 0. This might lead to potential bias in prediction due to an imbalance of the outcome variable toward recommended products. Therefore, regards of predictive modeling, the report only used reviews with a rating from 1 to 4 whose frequencies of Recommended.IND = 0 and 1 were 4078 and 6023 respectively. There were then 10101 observations. After that, all the negative abbreviations in Review.Text like "isn't" were transformed into "is not" before stemming with and without negators. After that, 15 factors from PCA, Nr_of_word (the number of words in stemmed reviews), NRC dictionary's emotions derived from the original Review. Text, unigrams, and bigrams of the Review.Text were added in the data set to build predictive modeling with the outcome variable as Recommended.IND. The data set then contains 274 features.

**Stage 1: Word embeddings**

Each word or review is represented in a high-dimensional space. When using the GloVe model, similarities can be found between those words. After the cleaning process of the dataset, Dimensions 5, 7, and 10 are selected for a meaningful interpretation. Table 1 shows the first five words for these three dimensions. All those words are presented as values, which are vector values. For example, the word "heel" occurs 0.6794577852 in Dimension 5. All words with a positive value have the same 'direction' and all words with a negative value have the opposite 'direction' in latent space. Meaning the words with a negative value are least similar to the word "heel". The words with a positive value are most similar to the word "heel". For a meaningful interpretation, similarities were found between words represented by the selected three dimensions. This selection is made based on the Cosine similarity matrix.

Table 1: GloVe word embedding dimensions

| | | | | | |
|---|---|---|---|---|---|
| Dimension 5 | Heel<br>0.6794577852 | Blous<br>0.6505881090 | Recommend<br>0.5104262444 | Type<br>- 0.0058139484 | Flow<br>- 0.0099517969 |
| Dimension 7 | Favorit<br>0.737931759 | Heel<br>0.451438561 | Denim<br>0.450431663 | Neck<br>- 0.006015574 | White<br>- 0.013203648 |
| Dimension 10 | Size<br>0.668430249 | Flow<br>0.2238559957 | Nice<br>0.5580486692 | Review<br>- 0.0070847907 | Recommend<br>- 0.0143157575 |

Table 2 shows this matrix (only a part of the whole matrix is shown in Table 2). GloVe most similar words are the words that have the highest Cosine similarity with a specific word. This means that all words with a positive value are strongly related to each other. And therefore, all the words with a negative value are strongly related to each other. In Table 2 in every row, the first three words are positive and the last two words are negative. Dimension 5 is represented by the words "heel", "blous", "recommend", "type" and "flow". But to truly understand the meaning of the similarities between these words we have to check them in Table 2. As stated above, Table 2 shows the Cosine similarity matrix. When looking at the word "heel" in Dimension 5 we see that the words "boot", "knee", "flat", "return", and "sweater" have similarities between each other. Note that the words “boot”, “knee” and “flat” are strongly related, because they all have positive values. The words “return” and “sweater” are then related because they have negative values. These rules also apply to Dimensions 7 and 10. When reviewing all the words which dimension 5 describes and the similarities between those words in Table 2, we can state that reviews or words mainly regard types of shoes, clothing for the upper part of the body (such as shirt, top, blouse, etc.), body types (shape, curvi, etc.), positive experiences (highly recommended) and flowy type of clothing items. Dimension 7 is represented by the words “favorit”, “heel”, “denim”, “neck” and “white”. Instead of checking the cosine similarity for the word “heel” again, we can take a look at the word “favorit”. “Favorit” is related to “pair”, “love”, “feminin”, “snug” and “larger”. Dimension 10 is represented by the words “size”, “flow”, “nice”, “review” and “recommend”. Again, we can check the cosine similarity for the word “size”. The words “fit”, “top”, “perfect”, “suit” and “embroideri” are related to “size”. This means that Dimension 10 is represented by the reviews about the fitting, whether the clothing is flowy or comfy, how nice the fabric feels, how the customers review the store and size, etc. The rules for the positive and negative values apply for these similarities, as mentioned above. When reviewing the words in the dimension and cross-referencing them with Table 2 the true meaning of the similarities can be found. As mentioned for Dimension 5, even the reviews and/or words can be classified into a certain category. Next, the selected element which has the largest score in absolute value is the word “favorit” from Dimension 7 with a score of 0.737931759. This score does not have a semantic meaning, because the meaning is as mentioned above, captured in the direction in the latent space described by word embeddings. However, it implies that the computer can extract that much (0.74) information from Dimension 7. In other words, Dimension 7 of the word “favorit” participates 0.74 in representing all of the words from the vocabulary.

Table 2: Cosine similarity matrix

| | | | | | | |
|---|---|---|---|---|---|---|
| **Dimension 5** | **Heel** | Boot<br>0.4481994 | Knee<br>0.4232189 | Flat<br>0.4159516 | Return<br>- 0.1768523 | Sweater<br>- 0.1785727 |
| | **Blous** | Top<br>0.53018323 | Shirt<br>0.50884946 | Print<br>0.45874674 | Pull<br>- 0.06780536 | Wore<br>- 0.07056281 |
| | **Recommend** | True<br>0.47925558 | Cut<br>0.45351884 | Larger<br>0.44261041 | Heavi<br>- 0.09612379 | Dark<br>- 0.10675521 |
| | **Type** | Curvi<br>0.4867363 | Shape<br>0.4860651 | Drape<br>0.4220809 | Front<br>- 0.1417340 | Bra<br>- 0.1434286 |
| | **Flow** | Flowi<br>0.5625611 | Drape<br>0.5252662 | Comfi<br>0.3686963 | Buy<br>- 0.1323856 | Decide<br>- 0.1372309 |
| **Dimension 7** | **Favorit** | Pair<br>0.5385849 | Love<br>0.4821433 | Feminin<br>0.4284598 | Snug<br>- 0.1413593 | Larger<br>- 0.1425127 |
| | **Denim** | Jean<br>0.5817615 | Pair<br>0.4329883 | Pant<br>0.4082727 | 5<br>- 0.1425688 | Refer<br>- 0.1500413 |
| | **Neck** | Necklin<br>0.6139620 | Arm<br>0.5352507 | Cut<br>0.5232565 | Excit<br>- 0.1085090 | Compliment<br>- 0.1211435 |
| | **White** | Black<br>0.6933996 | Red<br>0.6251939 | Navi<br>0.6183982 | Curvi<br>- 0.1076522 | Figur<br>- 0.1162432 |
| | **Size** | Fit<br>0.73909775 | Top<br>0.72370606 | Perfect<br>0.70814434 | Suit<br>- 0.02086892 | Embroideri<br>- 0.02585039 |

| | | | | | | |
|---|---|---|---|---|---|---|
| **Dimension 10** | **Nice** | Love 0.8477136030 | Top 0.7947170817 | Feel 0.7656494068 | Dark - 0.026224244 | Excit - 0.031588377 |
| | **Review** | Store 0.717421478 | Size 0.698188851 | Fit 0.581929286 | Low -0.028688917 | Boot -0.042327203 |

Finally, Table 3 shows similarities as an illustration of word arithmetic. For this analysis, we try to find an answer to the question of when consumers are most disappointed during certain seasons. For this, we use the following word similarity equation: disappointment + summer-winter. This means that "disappointment" shifts from winter to summer. Hence, "disappointment" is stronger related to "summer" than to "winter". The following words with positive values are regarded to be disappointments for summer and the words with negative values are regarded to be disappointments for winter (they are therefore opposites from one and another). With this analysis, we can advise clothing departments which clothes to emphasize in those seasons to improve the quality of certain items. For example, the word "shirt" indicates that it often disappoints during the summer. This is true because a shirt was mostly bought during the summer. But also, the "fabric" is a disappointment for the customers during the summer. The same applies to the word "coat". It often disappoints during the winter, because it is mostly bought then.

Table 3: Word arithmetic

| **disappoint** | 0.6796803 |
|---|---|
| **wash** | 0.5252591 |
| **return** | 0.5104493 |
| **person** | 0.5037366 |
| **pretti** | 0.4981245 |
| **fabric** | 0.4967115 |
| **shirt** | 0.4927419 |
| **summer** | 0.4889989 |
| **retail** | 0.4682715 |
| **white** | 0.4681498 |
| **spring** | -0.1217553 |
| **add** | -0.1273240 |
| **lb** | -0.1444187 |
| **highli** | -0.1559248 |
| **coat** | -0.1697664 |
| **belt** | -0.1769464 |
| **boot** | -0.2161526 |
| **warm** | -0.3410718 |
| **cardigan** | -0.3513360 |
| **winter** | -0.4251302 |

**Predictive modelling**

In this report, 7 predictive models were constructed to forecast the outcome of Recommended.IND as the independent variable to find which features in the review text might influence the intention to recommend the product of the customers. All models contained Age and Nr_of_word as independent variables and other features as required in the assignment. Especially, an interaction between Age and Nr_of_word was added in Lm.interaction_all and Glm.all_clothing to examine the influence of age on the length of the reviews. ANOVA tests and AIC scores also shown that models with more features were better at capturing the variance of the data set.

After comparing the performance of prediction, the glm.all_clothing was selected as the final model since it had the lowest AIC score of 8109.699. This means Glm.all_clothing had the best goodness-of-fit and explained the greatest amount of variation using the fewest possible independent variables. It also had the highest training accuracy (73.71%) although it experienced minor overfitting. Compared to Glm.all_clothing, Lm.interaction_all and lm.all_clothing had higher testing accuracy (72.88 % and 72.58%), however, their AIC scores were much higher at around 8622. Lasso regularized version of Lm.interaction_all also provided lower accuracy (around 72%). Other models all had higher AIC scores and lower prediction accuracy compared to Glm.all_clothing. For a more detailed comparison, please check the table below..

Table 3: Prediction performance comparison

| No | Model | Description | AIC | Accuracy | |
|---|---|---|---|---|---|
| | | | | Training data set | Testing data set |
| **1** | Lm.interaction_all | Linear model contained all PCA factors, Age, Nr_of_word, all emotions, all unigrams, all bigrams, and interaction between Age and Nr_of_word as independent variables | 8622.78 | 73.57% | 72.88% |
| **2** | Lm.all_clothing | Linear model contained all PCA factors, all emotions, all unigrams, all bigrams, Age, and Nr_of_word | 8621.82 | 73.66% | 72.58% |
| **3** | **Glm.all_clothing** | **Generalized linear model contained all PCA factors, Age, Nr_of_word, all emotions, all bigrams and unigrams, and the interaction between Age and Nr_of_word.** | **8109.69** | **73.71%** | **72.21%** |
| **4** | Lm.nodict_clothing | Linear model contained all PCA factors, Age, Nr_of_word, all unigram, and all bigram | 8762.04 | 72.53% | 72.05% |
| **5** | Lm.onlyemotions_clothing | Linear model contained only all emotions, Age, and Nr_of_word | 9667.37 | 63.79% | 64.41% |
| **6** | Lm.words_bigrams_clothing | Linear model contained Age, Nr_of_word, unigrams and bigrams | 9202.32 | 69.79% | 68.93% |
| **7** | Lasso.mod_clothing | Lasso regularization of Lm.interaction_all | - | 72.14% | 72.08% |

In Glm.all_clothing, the word “wear”, “disgust”, and PCA factor8 which captured features corresponding to the “fabric” are the top three most important predictive features. For “wear, consider two reviews that are almost the same. Review 1 does not contain the word "wear". Review 2 is a copy of review 1, but also mentions the word "wear". The odd of recommended product for review 2 will be 99.7% point lower. For “disgust” consider two reviews that are almost similar, review 1 does not contain the word "disgust". Review 2 is a copy of review 1 but also mentions the word "disgust". The odd recommended product of review 2 is 36.5% point lower. For PCA factor 8, consider two reviews. Review 1 does not have features captured by PCA factor8. Review 2 has features captured by PCA factor 8, the odd of the recommended product of review 2 is 99.45% point lower. In this model, the interaction between Age and Nr_of_word was not in the top 25 most important features.

These features were also the top 3 most important predictive features of Lm.interaction_all and Lm.all_clothing. These 3 models possessed the highest accuracy and lowest AIC scores compared to other models. For “disgust” it makes sense that the review containing it might be about an extremely negative experience, which leads to the products not recommended by the customers. However, for “wear” and PCA factor 2, they might make more sense if they are associated with low rating reviews. PCA factor 8 was also within the top 3 most important features in Lm.nodict_clothing. In this model, compared to an almost similar review, the Recommended.IND score of the review with features captured by PCA factor 8 was also lower. The word “wear” was also one of the most predictive features in the Lasso.mod_clothing. Other models have different most important predictive features. Regards of Lm.nodict_clothing, “beauti” as in “beauty” or “beautiful” and “fabric” increase the prediction point, which also makes sense since they might be referred to as the praise of the products. In Lm.onlyemotions_clothing, reviews categorized as “joy” had higher Recommended.IND score compared to those which were not in this emotion; Nr_of_words also increased the Recommended.IND score. Lm.words_bigrams_clothing had the word “return”, “look”, and “bit”. It is reasonable that the reviews with “return” and “look” had a lower prediction of Recommended.IND score than the similar reviews without it since if the product was returned, it was not likely to be recommended, and “look” might be associated with unexpected design in low-rating reviews. Whereas, “bit” has a positive impact on Recommended.IND since word often came with minor flaws. In the Lasso model, “wear”, “nice”, and “review_size” all had a negative impact on the prediction of Recommended.IND score. “review_size” might make sense if it came with low rating reviews like the “review_size” was not as accurate as actual size for example, however, “nice” might only make sense if it followed negators in the low – rating reviews. For a more detailed interpretation of each model, please refer to Table 2 below.

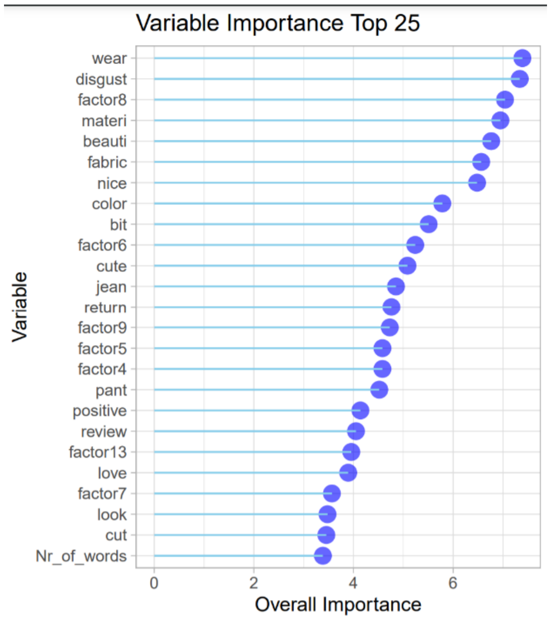


Figure 1: Variable importance based on T-value of the Glm.all_clothing model

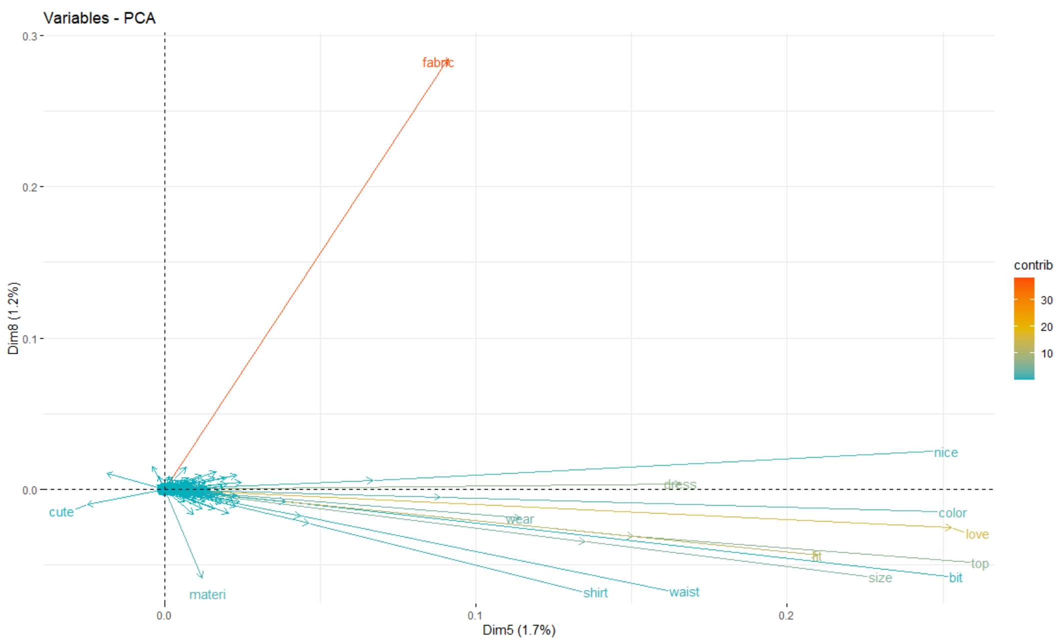


Figure 2: PCA analysis of dimension 8 and 5

Table 5: Interpretation and insight of different models

| No | Model | Description | Top 3 most important predictive features |
|---|---|---|---|
| **1** | Lm.interaction_all | Linear model contained all PCA factors, Age, Nr_of_word, all emotions, all unigrams, all bigrams, and interaction between Age and Nr_of_word as independent variables | - **Wear**: Consider two reviews that are almost the same. Review 1 does not contain the word "wear". Review 2 is a copy of review 1, but also mentions the word "wear". The Recommended.IND score for review 2 will be 1.062 lower.<br>- **Disgust**: Consider two reviews that are almost similar, review 1 does not contain the word "disgust". Review 2 is a copy of review 1, but also mentions the word "disgust". The prediction of Recommended.IND score review 2 is 0.018 lower.<br>- **Factor8**: Consider two reviews. Review 1 does not have features captured by PCA factor8. Review 2 has features captured by PCA factor 8, its prediction for Recommended.IND is 0.86 lower. |
| **2** | Lm.all_clothing | Linear model contained all PCA factors, all emotions, all unigrams, all bigrams, Age, and Nr_of_word | - **Wear:** Consider two reviews that are almost the same. Review 1 does not contain the word "wear". Review 2 is a copy of review 1, but also mentions the word "wear". The prediction for Recommended.IND score for review 2 will be 1.059 lower.<br>- **Disgust:** Consider two reviews that are almost similar, review 1 does not contain the word "disgust". Review 2 is a copy of review 1, but also mentions the word "disgust". The prediction for Recommended.IND score of review 2 is 0.0814 lower.<br>- **Factor8:** Consider two reviews. Review 1 does not have features captured by PCA factor8. Review 2 has features captured by PCA factor 8, its prediction of Recommended.IND score is 0.857 lower. |
| **3** | **Glm.all_clothing** | **Generalized linear model contained all PCA factors, Age, Nr_of_word, all emotions, all bigrams and unigrams, and the interaction between Age and Nr_of_word.** | - **Wear:** Consider two reviews that are almost the same. Review 1 does not contain the word "wear". Review 2 is a copy of review 1, but also mentions the word "wear". The odd of recommended product for review 2 will be 99.7% point lower.<br>- **Disgust**: Consider two reviews that are almost similar, review 1 does not contain the word "disgust". Review 2 is a copy of review 1, but also mentions the word "disgust". The odd of recommendation product of review 2 is 36.5% point lower.<br>- **Factor 8** Consider two reviews. Review 1 does not have features captured by PCA factor8. Review 2 has features captured by PCA factor 8, its odd of recommended product is 99.45% point lower. |
| **4** | Lm.nodict_clothing | Linear model contained all PCA factors, Age, Nr_of_word, all unigram, and all bigram | - **Beauti**: Consider two reviews that are almost the same. Review 1 does not contain the word "Beautiful". Review 2 is a copy of review 1, but also mentions the word "wear". The prediction of Recommended.IND score for review 2 will be 0.244 higher.<br>- **Factor8**: Consider two reviews that are almost similar, review 1 does not contain the word "disgust". Review 2 is a copy of review 1, but also mentions the word "Factor 8". The prediction of Recommended.IND score of review 2 is 1.023 lower.<br>- **Fabric**: Consider two reviews. Review 1 does not have features captured by PCA fabric. Review 2 has features captured by PCA fabric, its prediction of Recommended.IND score is 1.714 higher. |
| **5** | Lm.onlyemotions_clothing | Linear model contained only all emotions, Age, and Nr_of_word | - **Disgusting:** Consider two reviews that are almost the same. Review 1 was not categorized as "Disgusting". Review 2 was categorized as "Disgusting" based on sentiment analysis. The prediction of Recommended.IND score for review 2 will be 0.154 lower. |

| | | | |
|---|---|---|---|
| | | | - **Joy**: Consider two reviews that are almost similar, review 1 was not categorized as "Joy". Review 2 was categorized as “Joy". The prediction of Recommended.IND score of review 2 is 0.044 point higher.<br>**-Number_of_words:** Consider two almost similar reviews, if the number of words increases by 1, the prediction for Recommended.IND score of that review will increase by 0.0041. |
| **6** | Lm.words_ bigrams_cl othing | Linear model contained Age, Nr_of_word, unigrams and bigrams | - **Return:** Consider two reviews that are almost the same. Review 1 does not contain the word "Return”. Review 2 is a copy of review 1, but also mentions the word "Return". The prediction for Recommended.IND score of review 2 will be 0.173 lower.<br>- **Look**: Consider two reviews that are almost similar, review 1 does not contain the word “Look". Review 2 is a copy of review 1, but also mentions the word “Look". The prediction for Recommended.IND score of review 2 is 0.173 lower.<br>**-Bit:** Consider two reviews. Review 1 does not have features captured by PCA bit. Review 2 has features captured by PCA bit, its prediction for Recommended.IND score is 0.133 higher. |
| **7** | Lasso.mod _clothing | Lasso regularization of Lm.interaction_all | - **Wear:** Consider two reviews that are almost the same. Review 1 does not contain the word "Wear”. Review 2 is a copy of review 1, but also mentions the word "Wear". The prediction of Recommended.IND score for review 2 will be 0.440 lower.<br>- **Review_Size**: Consider two almost similar reviews. Review 1 does not contain the word “Review_Size”. Review 2 is a copy of review 1, but also mentions the word “Review_size”. The prediction of Recommended.IND score will be decreased by 0.258.<br>**-Nice:** Consider two almost similar reviews. Review 1 does not have the word “nice”. Review 2 is a copy of review 1, but also mentions the word “nice”. Then its prediction of Recommeded.IND score is 0.256 lower. |