eCommerce NLP: Ranking Forecast on Hair Care Products Based on Product Functionality and Aspect-based Sentiment Analysis

Group 13
Anqi Dong (A0218944W), Hui Xinyao (A0148006M), Li Xinyi (A0218908W)，
Zhang Yixuan (A0218975M), Zhao Dongyu (A0206513R)

Abstract

In recent years, with more ecommerce shopping options, consumers tend to shop online more frequently. As a result, consumers are willing to engage in online community to share their suggestions and feedbacks on the shopping experience. As the top ecommerce platform, Amazon has overwhelmed amount of unexploited review texts with high values for either the companies to enhance product quality or the ecommerce platform to improve the targeted recommendation system. This project integrates NLP technique in building a predictive model.

The aim of this project is to evaluate the key factors contributing to the Hair care product's sales rank on Amazon. Besides the product's metadata information and the review rating/votes, new features about product functionality fulfillment and aspect-based sentiment analysis (ABSA) on four general aspects (price, deliver, customer service and package) are created and used in the modelling. Various ABSA methods and product-level aggregation approaches are evaluated with respect to model performance. Eventually, the study confirmed the best approach is to use mixed sentiment methods with all average feature grouping strategy.

1. Introduction

1.1 Problem Statement

The study was inspired by the mega trend that consumers tend to shop more online than offline in recent years. Especially with Covid19, the trend accelerated dramatically. As a result, eCommerce is counting a higher percentage in overall global commerce transactions. Consequently, more online reviews were left by brand consumers.

By studying the reviews, brands could understand the true voice of their customers, receive first-hand product opinions, and eventually engage with them timely.

To unveil the insights of eCommerce reviews to the product ranking in key eCommerce platform, the study will analyze 2014 Amazon eCommerce data in Beauty Care – Hair Care category. And the study will process the reviews into key words and sentiments, then leverage modern predictive models to determine how they influence product ranking.

1.2 Product Functionality

With growing population of netizens as well as increasing popularity of social media, more user-generated-content (UGC) are created by users to share information and/or opinions with other users (Tang et al., 2014). It could be best illustrated by e-commerce where reviews are posted by consumers after transactions. Exploring reviews can have significant business values for three parties. For consumers, summarization on product feedbacks can offer them a quick impression on the products and subsequently facilitate their purchase decision. For ecommerce platform, reviews analysis result can be considered as basis of recommendation system so that proper products can be recommended to target consumers. Moreover, for brand companies, product feedback integration can help improving the product with regards to the aspect and further boost the sales (Fan et al., 2020).

In the current hair care products market, the most popular functionalities are focused on cleansing, dandruff, growth (including preventing hair loss), moisture, repair and color protection. Functionality keywords appearing in reviews, no matter if being "promised" in product titles or not, will be extracted respectively to measure whether a product has met the standard it promised or whether it has other extra functionalities that appeal to consumers' need.

1.3 Sentiment Analysis

Sentiment analysis is the area which studies people's viewpoints, sentiments, reviews, attitudes and emotion

Project Git-Hub: https://github.com/xinyao9661/5153-Ecommerce-NLP

forms to the entity and its attributes (Liu, 2012). It can be further divided into three subclasses – document-, sentence- and aspect-based sentiment analysis. In recent years, the sentiment analysis tasks have attracted more attention due to more and more open-source dataset available as well as the technical advance in processing texts based on language models such as BERT (Devlin et al.,2018) and GPT (Radford et al., 2019).

Document level sentiment analysis gives a primary option of a context. On top of sentiment analysis for the entire reviews, the project further explores reviews with regards to pre-defined aspects, including delivery, packaging, price and customer service. The purpose of aspect-based sentiment analysis (ABSA) is to extract sentiment information with regards to the target attribute. The flow of ABSA includes three steps, aspect information extraction, aspect sentiment classification and aspect information integration. In research from Fan et al. (2020), there are a wide spectrum of existing technologies to handle aspect feature extraction, including method based on frequency, rules, topic model, conditional random field and deep learning. Moreover, a number of technologies to tackle aspect sentiment classification, which are methods based on dictionary, support vector machine, deep learning, are available. In this project, we adopted two methodologies to carry out ABSA. They are methods based on frequency followed by sentiment classification with reference to dictionary and methods built upon deep learning and transformers. The tools and packages we used include *SentimentIntensityAnalyser* from NLTK Vader package and open-source Python Package, *aspect-based-sentiment-analysis*, built based on BERT and Tensorflow.

1.4 Overview of Methodology

The methodology of the work is arranged as following. Firstly, the relevant meta-product and reviews information were extracted for the selected products, shampoo and conditioners. Subsequently, the data was cleaned and preprocessed. Engineered features were created from two considerations, product functionality and ABSA. Taking into account of product meta information, intrinsic review information, and engineered functionality & ABSA features, a salesRank forecasting model was constructed and its result was further evaluated when the two different ABSA approaches were in use. In the end, the key factors signifying the product popularity were evaluated and discussed. In addition, limitation and future work for the project was illustrated in the last section.

2. Basic Facts of Data

2.1 Data Description

This dataset contains product reviews and metadata from Amazon, including 142.8 million reviews spanning May 1996 to July 2014. It includes reviews (ratings, text, helpfulness votes), product metadata (descriptions, category information, price, brand, and image features), and links (also viewed/also bought graphs). The summary of the dataset can be found in Table 1. In this work, analysis was based on review data and metadata (2014) of hair products (mainly shampoo and conditioner).

Table 1. Contents of Metadata and Reviews Dataset

meta	reviews
asin	asin
	helpful
imUrl	overall (rating)
salesRank	reviewText
title	reviewTime
price	reviewerID
brand	reviewerName
	summary

The two tables were merged on the common key asin. After excluding products with no review records, there are 13,132 products and 133,159 reviews in total.

2.2 Exploratory Data Analysis

By conducting exploratory data analysis, tables and charts are produced to understand the basic data distribution in meta and review data sets. Figure 1 and Figure 2 show the top 15 brands with most products asins and total reviews respectively. Many brands are among the top under both criteria. Figure 3 shows the distribution of overall ratings reflected by all the reviews, where most ratings are in the range between 4.5 and 5.

Figure 1. Top 15 Brands with Most Products

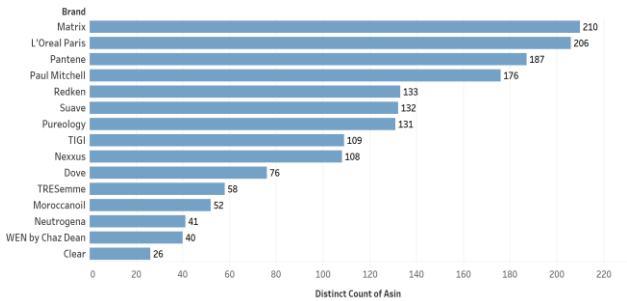


Figure 2. Top 15 Brands with Most Reviews

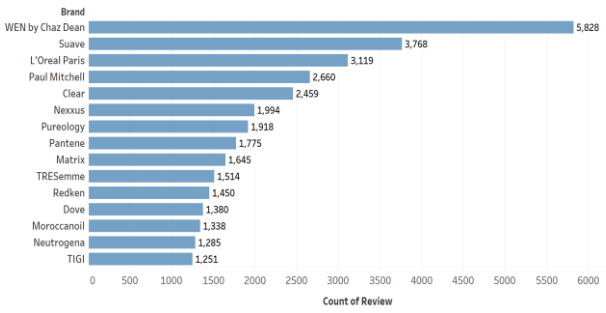
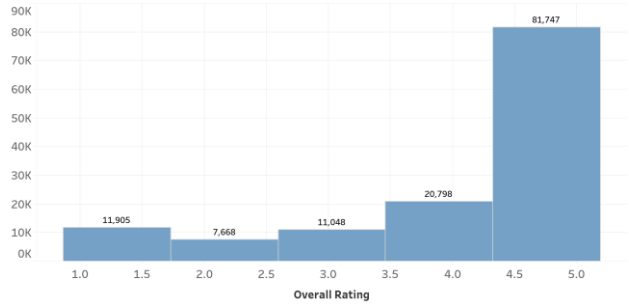


Figure 3. Distribution of Overall Rating



Distribution of the number of reviews per product can be found in Figure 4 below. About 31.6% of the products only have one review. 37.3% of the products have greater than or equal to 5 reviews.

Distribution of length of reviews is shown in Figure 5. A majority of reviews are short with 20-40 words whilst reviews with a size more than 4,000 characters exist as well. Note that customers who leave long reviews tend to be either very positive or negative.

Figure 4. Distribution of Number of Reviews Per Product

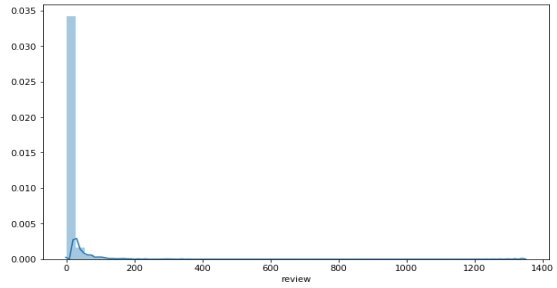
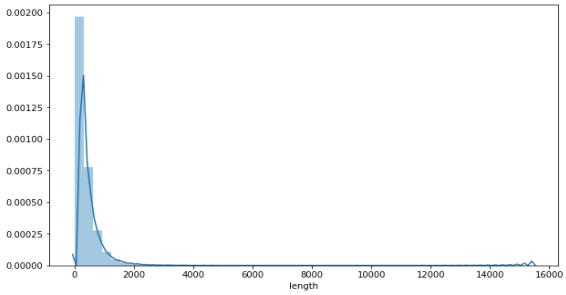


Figure 5. Distribution of Character Length of Reviews



3. Text Mining

In this section, various text mining techniques has been applied to generate the weighted sentiment score on different aspects of products and keywords indicator on product features. Followed by this, the features have been aggregated on a product level for modeling purpose.

3.1 Product Functional Keywords Extraction

Due to the nature of this hair care products being studied, both title (product names) and their reviews can reflect many different functional keywords. Matching the keywords in titles and reviews for each product can help with the understanding of whether a product has met the standard it promised or not, as well as whether a product has some extra functions other than what is expected.

Regarding the two aspects of functional keywords, a two-part extraction and matching procedure was taken to approach this idea. Please see below for examples on each of these parts.

- **Functional Keywords in Product Names**

Title kms headremedy dandruff conditioner 8.5oz

Review

'Really works for me. Dandruff evidently decreased. Will definitely purchase again.'

For example, for this particular KMS conditioner, "dandruff" is a promised functional keyword in the product title, which should be extracted in the first place.

From a review mentioning the promised function, the keyword should be extracted and this review will thus be considered to remark that the product has reached the basic standard.

One point to be noted is that there can be many different words and expressions standing for one particular kind of function. For example, "repair" and "damage" both indicate the repairing function for damaged hair.

Therefore, a manually crafted list is needed to extract functional keywords as comprehensive as possible.

- **Functional Keywords not in Product Names**

Title kms headremedy dandruff conditioner 8.5oz

Review

'Not only controls dandruff, but also strengthens hair. Nice product.'

Taking the same KMS conditioner as example, aside from "dandruff", the promised functional keyword, "strengthens" is also a functional keyword that should be extracted.

From a review mentioning some functions other than the promised function, this review will be considered to remark that the product has some extra functions.

Note that not all reviews remarking extra functions mentioned the promised function. In some cases, consumers found the product did not appear to have the basic function or they simply just did not mention in the review since it should be the baseline which did not need special attention, while it did have some other effects. For this kind of situations, if the extra functions are not taken into account, the product will be considered mostly bad reflected from the features, which is not precise because it might still have something good.

3.2 Overall Sentiment Score

Among all reviews, a majority of them are conveying an overall impression towards the product instead of being particular with certain aspects or functionality. Hence, document-level sentiment analysis is necessary and being carried out per review.

In this project, Sentiment Intensity Analyzer using VADER from NLTK package is adopted to obtain the sentiment score. This function acquires the sentiment valence by checking single word and bigrams and mapping with a lexicon and idioms dictionary respectively. Hence, relevant pre-processing including expansion of abbreviations is necessary. For example, "can't" should be replaced by "cannot".

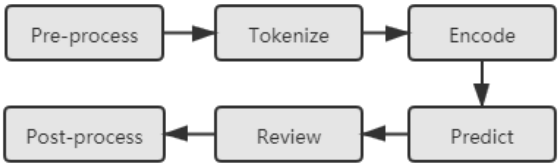
3.3 Sentiment by Aspect

Price, delivery, customer service and package have been selected as the aspects to be analyzed in the light of literature review and frequency of word occurrence in review data. To obtain a more solid result, two approaches, which are based on deep learning or customized method, on generating sentiment by aspects have been applied.

- **Methods Based on Deep Learning**

The first method is to apply a pretrained BERT Classifier which recognized the pattern and gave a sentiment score for each aspect. We used the open-source Python package *aspect-based-sentiment-analysis* implemented by Rolczynski (Scalac & Rolczyński, 2021). The aim of the package is to classify the sentiment of potentially long texts for several aspects, which can lead to improvement of reliability on review-based prediction. The pipeline in the used package is illustrated in Figure 6..

Figure 6. Package Pipeline



The package is built based on BERT transformer and Tensorflow. Firstly, the pipeline preprocesses the texts by splitting long sentences by using text splitter. As long sentence tends to return a neutral score, short sentences containing sufficient explicit and contextual aspect-level information are used as proper entity for measuring the aspect level sentiment score. The pipeline then tokenizes and further prepares for a set of encoding layers, which takes into account of the words embedding as well as their position in a sequence.

A pre-trained BERT aspect-based sentiment classifier will then be used to make the prediction. In more details, a Basic Reference Recognizer first measures the cosine similarity between a text and the aspect and predicts whether the text relates to an aspect by a simple logistic regression. Following on, a Base Pattern Recognizer uses attentions and their gradients to discover patterns which the model uses for prediction. Only crucial pattern is identified and unimportant pattern will be excluded out. As last, the model returns the neural, positive and negative score for the given aspect. Review and post-process step can be utilized for interpreting the result.
The package is very powerful in capturing the various relations between words by constructing the language model. However, upon deploying, certain limitations of the package are observed. To begin with, the package cannot handle texts which are longer than 2000 in character length. Very long texts are omitted in the analysis. More importantly, the prediction result is inaccurate if the review text does not mention about the aspect.

Figure 7. ABSA Illustration

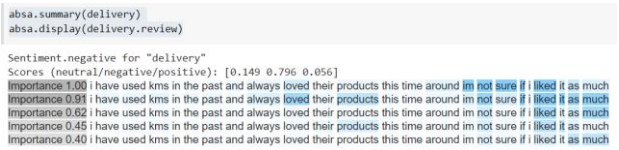


Figure 7 is an illustration of the pattern recognised by the package and the scores measured for the aspect of delivery. Even though the text is not concerned with delivery at all, the packge tends to give a very high negative score for the aspect of interest.

Hence, the limitation is addressed by extracting relevant reviews by first applying keywords on the four aspects-price, delivery, customer service and package. The reviews after filtering are then sent to the package for analysis. For reviews which are not concerning the given aspect, zero scores are assigned to them. The summary of keywords used for each aspect as well as the number of reviews extracted is presented in the table below.

Table 2. Summary of reviews which are sent to aspect-based-sentiment-analysis package for analysis

ASPECT	KEYWORDS	#REVIEWS
PRICE	Price | Pricing | Cost | Charge	17,377
DELIVERY	Delivery | logistics | shipment | transportation | transport	1,243
CUSTOMER SERVICE	Service | Return | Exchange | Customer+Support| Experience | Care | Assistance	2,981
PACKAGE	Package | packaging | wrapping | pack	4,463

- **Customized Method Based on Frequency**

As the ready package *aspect-based-sentiment-analysis* might leads to biased sentiments, another simple approach based on *nltk* has been deployed.

Firstly, a list of keywords has been set for each aspect. For example, ['price', 'pricing', 'cost', 'charge'] for aspect *Price*. Each review is then divided into sentences using *nltk.sent_tokenize*. If the sentence contains at least one of the keywords, the sentiment score generated using *nltk.sentiment.vade* will be recorded. The final sentiment of the review for certain aspect will be the average of the sentiments score of sentences containing as least 1 keyword. If none of the keywords appear in the review text, the sentiment of the review on this aspect will simply be 0.

This approach is easy to implement and understand. However, it can also lead to potential bias. For example, it is possible that one sentence will contain the keywords from 2 aspects. Although it may not be common, in such scenario, approach 2 can generate same sentiment score for 2 different aspects based on same sentence.

Algorithm of Customized Approach

Input: review, keywords
Repeat for all reviews
  Initialize *Contain Key = False*
  Initialize *sentiment_lists=[]*
  for *s in sent_tokenize(review)*
    if *s contains keywords:*
      Contain Key = True
      senti= sentiment score of s
      sentiment_lists.append(senti)
  if *Contain Key:*
    return *sentiment_lists.mean()*
  else:
    return *0*

3.4 Product-Level Aggregation

To build the model for rank prediction of each product, review-based keyword indicators and sentiments by aspects score were aggregated on product-level using average value. It is worth noting that 2 approaches of aggregation on aspects based on sentiment scores has been applied.

First approach is to simply use the mean of all reviews as the feature of the product. Second is to remove all the reviews whose sentiment score of this aspect is 0, then use the average of remaining reviews as the representative statistics for this product. The first approach has relatively small margin, which can reflect how frequent certain aspect has been mentioned for each product. The second approach provides a stronger indication of sentiments, which can help to better distinguish the products. Further discussion on these 2 aggregation approaches will be elaborated in modelling section.

4. Prediction Model

4.1 Feature Engineering

Original data is on review level. Features contain basic information of specific review for certain products as well as extracted sentimental features from review.

- **Basic Information**

  Overall score, number of regarding review as 'helpful', number of regarding review as 'not helpful', price, brand, length of review, etc.

- **Extracted Functional Aspects**

  Cleansing, dandruff, growth, moisture, repair, color. Each aspect is signed whether it appears in review/title or not. Furthermore, 'Satisfactory' feature concerns whether functional aspects appearing in title also appear in review, to measure if the product is worthy of its title.

- **Extracted Sentimental Scores of General Product Aspects**

  Delivery, price, service, package. Each aspect has 4 score: polarity, positive, neural and negative.

To find relationship between products ranking and features, we grouped data by product id. Methods of dealing with different features are shown in Table 3.

Table 3. Summary of Methods Dealing with Features

Feature	Method
Overall score Helpful overall Helpful positive Review length Price Functional aspects Total sentimental score	Average
General product aspects	Average Average without 0
Brand	Take top 19 big brands as dummy and merge the rest as 'others' (20 columns overall)
Helpful sentimental score	Helpful overall * polarity score
Number of reviews	Count

To mention that there are 2 methods to deal with general product aspects:

- **Method 1:** Average all rows
- **Method 2:** Only average rows which are not 0 since the matrix is sparse.

Method 1 gives more consistent score, while method 2 gives more extreme score. For example, a product has 10 reviews, among which only one review mentions 'delivery' and has a high negative score. Method 1 will dilute negative score because it regards the rest of 9 reviews have a neural altitude of 'delivery' by default. However, method 2 will take that negative score as the overall negative score, since it regards only one review cares much about 'delivery'. Performance of different methods are compared in later models.

To avoid multicollinearity, neural sentimental scores are dropped. To make prediction more accurate, we also drop products with less than 3 reviews because too small number of review data may cause large bias in result.

There may have causal inversion problem between sales rank and number of reviews, since high rank will attract more customers to purchase and review on products. Also, relatively high correlation indicates that data leakage may happen. So we dropped number of reviews.

Sales rank is inconsistent in original data since there are other kinds of products ranked together. So it will be more accurate to care about percentage of rank instead of absolute rank number. Ranks are cut into three categories: 0-10%, 10%-40%, and 40%-100%, since sales rank conforms to long tail phenomenon.

4.2 Modelling

Random Forest, XGBoost and Neural Network models are applied to predict sales rank with 2 average methods. Also, 2 different sentimental analysis methods introduced before and their mixed method are tried to get better prediction result. The performance of models is shown in Table 4 and 5.

Table 4. Model Performance (All Average by Products)

	Sentiment from Package	Sentiment from Customized Code	Mixed Features
Best Model	XGBClassifier	XGBClassifier	XGBClassifier
Best Score	0.64	0.64	0.65
NN Score	0.61	0.62	0.63

Table 5. Model Performance (Average On Positive Numbers Only)

	Sentiment from Package	Sentiment from Customized Code	Mixed Features
Best Model	XGBClassifier	XGBClassifier	XGBClassifier
Best Score	0.65	0.63	0.64
NN Score	0.61	0.62	0.61

From the result, we can see that XGBoost with mixed sentimental features and all average method performs the best, with total accuracy 0.65.

From feature importance analysis result, sentimental score of review matters a lot. The 5 most influencing features are pos_price_package_avg, neg_price_avg, r_func_repair_avg, polarity_price_avg, neg_service_ and package_avg.

5. Insights

After comparing with above model setups, the study confirmed to use following model to interpret the insights:

- Features: all average by product, with mixed features
- Model: XGBoostClassifier
- Accuracy: 65%

However, since the predictive classes was break into 10% (high class), 30% (medium class) and 60% (low class), then our prediction baseline is 60%. Therefore, the review sentiments and keywords could only provide marginal contribution. (around 5%)

Despite that the overall contribution was marginal, the sentiment and keywords still contribute positively to the model. By checking classification reports:

Table 6. Model Evaluation for Different Classes

class	precision	recall	f1-score
Low class	0.71	0.86	0.78
Middle class	0.47	0.34	0.39
High class	0.59	0.34	0.43

It shows that the model can accurately predict lower class products using sentiment and keywords. However, the model is struggling on classifying the middle and high-class products.

The top 10 important features are shown in Table 7.

Table 7. Top 10 Important Features in XGBoost Model

Feature	Explanation	Importance
pos_price_package_avg	Sentiment, Positive on price	3.87%
neg_price_avg	Sentiment, negative on price	3.18%
r_func_repair_avg	Keywords, contains repair function	3.12%
polarity_price_avg	Sentiment, polarity score on price	2.37%
neg_service_package_avg	Sentiment, negative on service	2.29%
pos_delivery_avg	Sentiment, positive on delivery	1.99%
r_func_cleansing_avg	Keywords, contains cleansing function	1.92%
r_func_dandruff_avg	Keywords, contains dandruff removal function	1.89%
pos_shipment_package_avg	Sentiment, positive on shipment	1.80%
polarity_service_avg	Sentiment, polarity score on service	1.74%

The key features obviously indicated that both sentiment and keywords contribute on ranking prediction. In addition, sentiments from both package and hand-code are picked as top features, it means that the discussion on the two approaches are valid.

By looking into the features, the top sentiment features are about pricing, customer service, and delivery. They are also the key aspects on eCommerce business performance. Therefore, the study can further confirm that for eCommerce business, a reasonable pricing, a high-quality customer service, and an effective delivery are the key to success. In product function wise, hair-repair, hair-cleansing, and dandruff removal are the top functions that influence the product ranking. They reflected that customers nowadays are mainly suffer from hair damage and oily/dandruff hair.

In summary, the consumer insights for eCommerce brands are that their consumers care about product pricing, customer service and delivery. The product insights are that they should focus on hair-repair, hair-cleansing, and dandruff removal products advertisement.

6. Conclusions and Limitations

In conclusion, the study explored different review data sentiment analysis approaches, extracted tailored eCommerce key words, and predicted product classes with review data. With the successful prediction, the study proved that review keywords and sentiments could provide marginal influence to the product ranking. Secondly, the study discussed key features of the hair care industry and their impact to the product classes.

Lastly, for future works, the study could focus on Advanced sentiment analysis, Integrating review analysis with other eCommerce analysis, and Individual level prediction explanations.

Acknowledgments

The authors are grateful to Prof. Zhao Rui and all the teaching assistants for helpful discussions. This work is supported in part of a coursework BT5153 final report.

References

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Fan, S., Yao, J., Sun, Y., & Zhan, Y. (2020). A Summary of Aspect-based Sentiment Analysis. Journal of Physics: Conference Series, 1624, 022051. https://doi.org/10.1088/1742-6596/1624/2/022051

Liu, B. (2012). Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies, 5(1), 1–167. https://doi.org/10.2200/s00416ed1v01y201204hlt016

Scalac, & Rolczyński, R. ł. (2021, March). Do You Trust in Aspect-Based Sentiment Analysis? Testing and Explaining Model Behaviors. https://rafalrolczynski.com/2021/03/07/aspect-based-sentiment-analysis/

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9.

Tang, T. Y., Fang, E. E., & Wang, F. (2014). Is Neutral Really Neutral? The Effects of Neutral User-Generated Content on Product Sales. Journal of Marketing, 78(4), 41–58. https://doi.org/10.1509/jm.13.0301