Automatic Aspect-Based Sentiment Analysis (AABSA) from Customer Reviews

Ella Jiaming Xu[1], Bo Tang[2],
Xiao Liu[1], and Feiyu Xiong[2]

[1] Stern School of Business, New York University
{jx1258, xliu}@stern.nyu.edu
[2] Alibaba Group
{tangbo.t, feiyu.xfy}@alibaba-inc.com

Abstract. Online review platforms provide enormous information for users to evaluate products and services. However, the sheer volume of reviews can create information overload that could increase user search costs and cognitive burden. To reduce information overload, in this paper, we propose an Automatic Aspect-Based Sentiment Analysis (AABSA) model to automatically identify key aspects from Chinese online reviews and conduct aspect-based sentiment analysis. We create a hierarchical structure of hypernyms and hyponyms, apply deep-learning-based representation learning and clustering to identify aspects that are the core content in the reviews, and then calculate the sentiment score of each aspect. To evaluate the performance of the identified aspects, we use an econometric model to estimate the impact of each aspect on product sales. We collaborate with one of Asia's largest online shopping platforms and employ the model in its product review tagging system to help consumers search for product aspects. Compared with benchmark models, our model is both more effective, because it creates a more comprehensive list of aspects that are indicative of customer needs, and more efficient because it is fully automated without any human labor cost.

Keywords: Aspect-Based Sentiment Analysis · Representation Learning · Deep Learning · Sentiment Analysis · Econometric Model

1 Introduction

Online reviews are critical for multiple stakeholders. Consumers can obtain rich information from reviews to evaluate products and services. Firms can leverage reviews to gain insights on customer needs and opportunities to improve their products. Despite the enormous informational value provided by reviews, the sheer volume of reviews has created a problem of information overload. Consumers cannot easily process information of thousands of reviews to understand the strengths and weaknesses of each product fully. Firms cannot easily glean insights from immense unstructured review content of their own products and competitors'. Although review rating, usually on a five-point Likert scale, is a

 In: N. Chhaya, K. Jaidka, J. Healey, L. H. Ungar, A. Sinha (eds.): Proceedings of the 3rd Workshop of Affective Content Analysis, New York, USA, 07-FEB-2020, published at http://ceur-ws.org

useful summary statistic, it is a single-dimensional value that fails to capture the multi-dimensional facets of products.

To overcome the information overload problem, a few online platforms, such as Yelp and TripAdvisor, have started to provide aspect-based sentiment scores on review pages, which help customers select reviews containing the selected aspects. Current literature also pay growing attention to identifying customer needs from online reviews [2, 28, 33, 30, 37, 6, 34, 7, 25]. However, both the platforms and the literature face several fundamental problems:

First, many previous solutions applied the supervised learning approach to define aspects manually and then match them to corresponding reviews [6, 2, 34, 33]. This approach has many drawbacks. First of all, e-commerce platforms, such as Amazon and Alibaba, often cover a wide range of product categories. The aspects that are relevant to one product category might be irrelevant for another category. For example, sound quality is important for the TV category but not for the sofa category. Therefore, it is time-consuming to identify aspects manually for each product category. Moreover, e-commerce evolves rapidly. New product categories constantly arrive, and customer tastes are continuously changing. It is hard to keep up with the trend and manually identify aspects for each newly developed category and new customer needs. For example, in 2019, a new product category, anti-smoking smart lighter, was introduced. And new aspects such as social connectedness and windproof need to be defined and added. Furthermore, even if defining a set of aspects is feasible, annotating large datasets is highly demanding on human labor costs and time. Last but not least, with significant human intervention, uncontrollable bias may arise. Second, although some previous papers also proposed automatic aspect detection, they selected the most frequently mentioned aspects as the core product aspects [40, 6, 2]. The drawback of this approach is that it ignores word similarities. For example, although hue and color each might not be the most frequently mentioned keyword, they can be constructed as an important aspect jointly. Some follow-up research applied word embedding techniques, such as word2vec and wordnet clustering, to capture word similarities [40, 36, 38]. But these works fail to capture complex semantic relationships of aspect keywords.

Third, previous literature assumes that all the aspects are at the same level [6, 2, 34]. However, there are limitations to flattening the aspect structure. Consider a laptop retailer aiming at improving the quality of laptops. Quality, undeniably, is an important aspect of a laptop, but it is an abstract aspect that consists of many sub-aspects like durability and speed. A review without the keyword "quality," but with more specific words such as "durability" and "speed" could also reflect a consumer's overall sentiment towards the quality of a laptop.

In summary, the following questions are left unsolved to conduct aspect-based sentiment analysis:

1. How can we identify aspects automatically with reduced cost and improved flexibility?

2. How can we better capture the semantic relationship among keywords to construct aspects?

3. Does the hierarchical structure among aspects exist, and can the hierarchical structure improve the comprehensiveness of identified aspects?

In this paper, we develop the Automatic Aspect-Based Sentiment Analysis (AABSA) model to extract hierarchical-structured product aspects from online consumer reviews. Specifically, we provide solutions to the three questions mentioned above:

1. We propose a fully automated aspect-based sentiment analysis model (AABSA). The model can create aspect-based sentiment scores from online reviews without any human intervention or domain knowledge. In the AABSA model, we applied k-means clustering to put sentence embeddings into groups and select the center words from clusters as aspects. A prominent advantage of the k-means clustering is that it is an unsupervised learning model so that we do not need to pre-determine the aspects manually. AABSA could automatically identify the aspects, aspect structure, and the number of aspects. No labels are needed for the learning process, leaving it on its own to find structure in its input. The model saves the time of defining labels and allows us to identify aspects automatically.

2. We introduce the Bidirectional Encoder Representations from Transformers (BERT) model to transfer short reviews into sentence embeddings and cluster them [9]. Unlike recent language representation methods, BERT jointly considers both left and right side context of words in all layers and helps us better capture the semantic relationship among aspects.

3. We develop a hierarchical aspect structure consisting of hypernym aspects, which are defined as the core content that can summarize the semantics, and hyponym aspects, which are defined as the sub-aspects of hypernym aspects. We first cluster sentence embeddings and identify center words of clusters as hypernym candidates. We then applied PageRank to build a weighted word map with synonyms of hypernym candidates and applied PageRank to identify hypernyms [26]. An essential advantage of the model is that the hypernyms and hyponyms are not necessarily the words that appear most frequently, but those that can capture the theme of the entire sentence. The hierarchical structure significantly increases the comprehensiveness and accuracy of identified aspects.

In summary, this paper makes several substantive and methodological contributions. We propose an innovative method to identify product aspects by introducing a hierarchical system. We demonstrate three comparative advantages of the proposed model against benchmark methods: 1) improved comprehensiveness, 2) better prediction accuracy on sales, and 3) full automation without time-consuming hand-coding. The method has been employed in Alibaba's Chinese product review tagging system to help consumers search for product aspects.

2 Literature Review

2.1 Aspect Identification

Online review platforms allow customers to express their attitudes towards products and services freely, and customers rely on online reviews to make decisions.

The sheer volume of online reviews makes it difficult for a human to process and extract all meaningful information. Hence, research on identifying product aspects from user-generated content and ranking their relative importance has been prolific in the past few decades [31, 40, 42, 12]. The most common methods rely on focus groups, experiential interviews, or ethnography as input. Trained professional analysts, then review the input, manually identify customer needs, remove redundancy, and structure the customer needs [34, 16, 1]. [40] identified important aspects according the frequency and the influence of consumers' opinions given to each aspect on their overall opinions by a shallow dependency parser. [42] then extended [40]'s paper by performing extensive evaluations on more products in more diverse domains and more real-world applications. [12] applied an automatic clustering approach to aspect identification. One common limitation of these approaches is that they assume the frequency that an aspect appears is positively correlated with its importance. However, high-level and abstract concepts, such as "quality," may not appear very frequently in the reviews. Still, the associated low-level, concrete concepts, such as durability and conformance, may appear very frequently. The approaches, as mentioned above, could fail to detect important high-level and abstract aspects. We instead propose a method that can rely on the hierarchical structure between hypernyms and hyponyms to detect important aspects. And our method is fully automatic, not relying on any human labor cost.

In the marketing field, researchers often rely on existing psychological and economic theory to pre-define a list of aspects and then extract the pre-defined aspects from user-generated reviews [41, 24, 19, 35, 8, 10, 20]. However, this approach is theory-driven instead of data-driven. Therefore, it is hard to generalize across contexts. For example, one paper that extracted the "health" aspect from weight-loss products might not be relevant for another product category, such as automobiles. In contrast, we propose a data-driven method that can extract the most relevant aspects tailored to the specific context. And our method is domain knowledge agnostic, not relying on human expertise.

2.2 Aspect Sentiment Analysis

Sentiment analysis is a type of subjectivity analysis that aims to identify opinions, emotions, and evaluations expressed in natural language [27]. The main goal is to predict the sentiment orientation by analyzing opinion words and expressions and detect trends. Sentiment analysis plays an important role in identifying customer's attitudes towards brands, and recent studies are paying more attention to developing more fine-grained aspect-based sentiment analysis on user-generated content. Previously, researchers studied extraction of evaluating expressions from customer opinions [4, 17, 27, 32, 43, 14]. [14] extracted features and summarized opinions from consumer reviews by part-of-speech tagging and built an opinion word list. [4] summarized the sentiment of reviews for a local service and focused on aspect-based summarization models.

With the development of machine learning techniques, researchers applied these advanced techniques to sentiment analysis. [23] introduced support vec-

tor machines (SVMs) and unigram models to sentiment analysis. [22] applied Naive Bayes to analyze aspect-based sentiment. [18] applied the Maximum Entropy (MaxEnt) classification to classify consumer messages into either positive or negative. [27] researched the performance of various machine learning techniques, including MaxEnt classification, and showed that MaxEnt classification was powerful with classifying reviews. Researchers then applied deep learning methods such as XLNet and LSTM to conduct sentiment analysis [39, 11, 15, 29]. In this paper, we tested both MaxEnt classification and Fasttext and found that MaxEnt outperforms Fasttext because the e-commerce platform has created a rich sentiment vocabulary pertaining to product reviews. The compared results of MaxEnt and Fasttext is listed in Table 2 of Appendices.

3 AABSA Model Framework

In this section, we describe the details of the structure of our aspect-sentiment analysis model, AABSA. We start with an overview of its framework, which consists of two main components: aspect identification and sentiment analysis. We then describe the baseline model to be compared with.

3.1 Aspect-based Sentiment Analysis Problem

The aspect-based sentiment analysis problem is to identify product aspects from a review document, and the aspects represent the most important customer needs in the document. Having identified the aspects, we then need to associate sentiment scores with every aspect. We make two assumptions. First, we assume that each sentence is possible to be associated with more than one aspect. Second, we assume that the hierarchical structure exists among aspects. We classify aspects into hypernyms and hyponyms. Hypernyms are the core content that can summarize the theme of the content and are often abstract and involve various sub-aspects. For example, "battery" is a core theme in the camera category identified by a previous research [2]. However, from the retailer's perspective, "battery" cannot provide them with detailed and comprehensive information on the direction to improve the battery aspect. As a result, we identify the sub-aspects as their hyponyms. For example, if the "battery" is identified as a hypernym, then "battery life" and "battery production place" are its possible hyponyms. Hyponyms could provide more exact direction for product improvement.

3.2 AABSA Model Framework

Our model consists of eight steps:

1. Pre-process reviews. We collected reviews from Alibaba, one of the biggest e-commerce platforms in Asia. In our research, we analyze reviews for two product categories: camera and toothbrush. There are two reasons why we choose these two categories. First, camera and toothbrush are common products and they are widely analyzed in marketing literature and we are able to compare our

results with previous works. Second, the camera represent the high-end product categories and the toothbrush represent the lower-end and more daily-used products. We can compare the impact of reviews on the sales of them and generate business insights. We divide the entire review document into short sentences and identify informative sentences, which were defined by the company's existing internal rules. For example, the sentence "Very good" is classified as uninformative, whereas the sentence "The battery can last more than 10 hours" is classified as informative.

2. Train word embeddings. The hierarchical aspect structure is based on the relationships between hypernyms and hyponyms, which are represented by word similarities. Concerning quantitative similarity representations, we convert words into vectors using the Word2vec algorithm and eliminate the lower-frequency words in synonym pairs [21].

3. Train sentence embeddings. In order to measure the similarity between reviews quantitatively, we convert the most frequent 50% short sentences into sentence embeddings with BERT [9].

4. Select hypernym candidates. We assume that a few core words, which are defined as hypernym candidates, could summary each sentence. Hence, we apply k-means clustering to generate semantic categories and select the most important words, whose accumulated cosine distances to their cluster centers are the shortest within the clusters, as hypernym candidates. We then filter invalid words among hypernyms candidates.

5. Further, we introduce the concept of hyponyms to assist in the subsequent sentiment analysis step. We select the words closest to the hypernym candidates in each cluster as hyponyms candidates. We then select the words closest to the hyponym candidates as their subordinates. Then we construct a weighted word network comprising hyponym candidates, hypernym candidates, and their subordinates. We then apply PageRank to select hyponym candidates according to their relative importance.

6. Merge hypernyms and hyponym candidates. We find that there are overlaps among hypernyms and hyponym candidates. To avoid redundancy, we rank all the hypernyms according to their importance and merge the hypernym and hyponym candidates to finalize hypernyms. If a hypernym belongs to several hypernym sets, then we merge the hypernym with its hyponym candidates to the hyponym set of the highest-ranked hypernym.

7. Match hypernyms to reviews. We select the words closest to hypernyms and hyponyms from the content and then apply a regular expression matching to match them to reviews.

8. We use the Maximum Entropy (MaxEnt) classification to classify reviews sentences associated with each aspect into positive, neutral, or negative. The sentiment score of reviews of each product is aggregated at the week level.

The framework and algorithm of AABSA model is shown in Algorithm 1 and Figure 1.

Algorithm 1 Aspect Identification

Input: Reviews: $\{R_i\}_{i=1}^{I}$
Output: Hypernym set H_1 and hyponym set H_2

Learn word vectors from reviews: $\{WV_j\}_{j=1}^{J}$ = **Word2Vec**($\{R_i\}_{i=1}^{I}$)
Divide all reviews $\{R_i\}_{i=1}^{I}$ into short reviews $\{SR_i\}_{i=1}^{SI}$
Calculate vector representation of short reviews $\{SRV_i\}_{i=1}^{SI}$ with BERT
Cluster $\{SRV_i\}_{i=1}^{SI}$ into k clusters with k-means clustering
Calculate the center of each cluster m: C_m
for m in clusters **do**
    for word w_j in cluster m **do**
        Calculate the frequency of w_j in cluster m: N_{mj}
        for instance n_{w_j} of word w_j **do**
            Calculate the cosine distance between w_j and C_m: $D_n(w_j, m)$
        Calculate importance of w_j in m: $F_{mj} = \sum_{n_{w_j}=1}^{N_{mj}} D_n(w_j, m)$
    Select word $\hat{w_j}$ with highest F_{mj} as the hypernym in cluster m
Form hypernym set H_1
for w_{1i} in H_1 **do**
    for w_{2j} in **mostSimilarN**(w_{1i}) **do**
        Add edge(w_{1i},w_{2j}) (=D(WV_{1i},WV_{2j})) to Net_N
        for w_{3k} in **mostSimilarN**(w_{2j}) **do**
            Add edge(w_{2j},w_{3k}) to Net_N
Rank words using **WeightedPageRank** (Net_N)
for $h_j \in H_1$ **do**
    Select TopN words with highest ranking as hyponym set H_{2j}
Sort hypernyms by descending importance
for $h_{1j} \in H_1$ **do**
    for $h_{1i} \in H_1$ **and** i>j **do**
        if $h_{1i} \in H_{2j}$ **then**
            Merge h_{1i} and H_{2i} with H_{1j}
Sentiment analysis using MaxEnt

Pre-process reviews In general, online reviews are complex sentences consisting of complicate sentiments and are composed of both informative and uninformative contents [34]. For example, in a review such as "I just got this camera today, and it looks fantastic but it's too heavy for me!", the first clause is uninformative since it is irrelevant to the camera's aspects, while the second clause describes the customer's positive attitude towards its appearance and negative attitude towards its weight. To better identify sentiments and informative contents, we separate original comments into single sentences and then automatically eliminate uninformative single sentences with regular expression matching with predefined regulations. Then we automatically eliminate the stop-words, numbers, brand names, and punctuation.

Train word embeddings To measure the similarities and also figure out synonyms quantitatively, we need to transfer words into vectors with word embed-

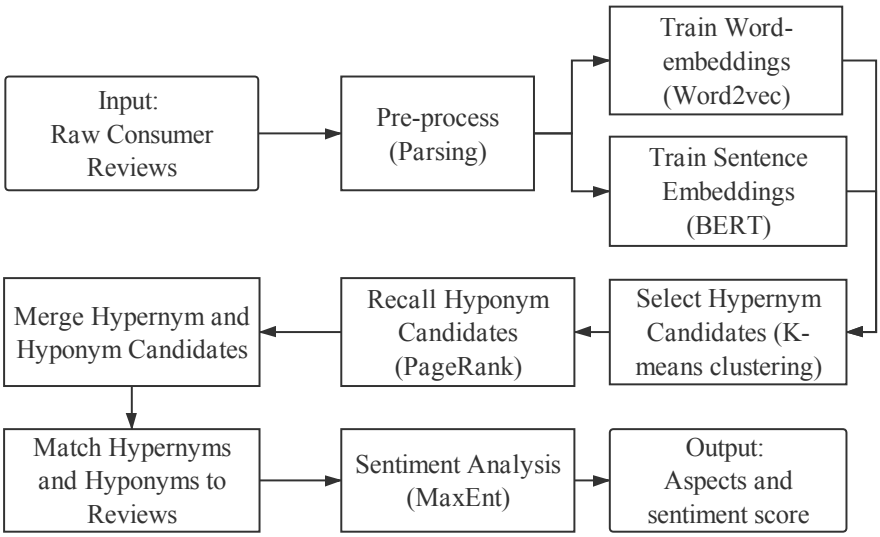

Fig. 1. AABSA Framework

ding. Word embedding is a representation of document vocabulary utilizing the context of a word, including semantic and syntactic similarity and word relationships. With word embedding, words used in similar contexts have similar representations, and the cosine similarity between word vectors could quantitatively represent similarities between words. We apply a skip-gram word2vec model to train word embeddings [21]. Skip-gram takes as its input a large corpus of text and produces a vector space, typically of several hundred dimensions, with each unique word in the corpus being assigned a corresponding vector in the space [21].

Train Sentence Embedding (with BERT) Sentence embeddings are useful for keyword expansion and are used to identify the relationship between words and sentences. In order to quantify the relationship between the sentences and discover the latent customer needs, we formulate sentence embeddings and extract keywords from the sentence clusters afterward. Consider the following examples:

"The toothbrush hair is super soft, and it really protects my son's teeth!"

"I am really disappointed that its toothbrush hair too soft, and it cannot clean my teeth."

These two sentences are different expressions of opposite attitudes towards the same product aspect, but they are similar in the semantic structure. In earlier works, researchers often created sentence embeddings by directly taking the average of word embeddings, which ignores the semantic and concatenate relationships between sentences [9]. For example, word2vec would produce the same word embedding for the word "soft" in both sentences. Language models

training word embeddings only use directionless or unidirectional context and match each word to a fixed representation regardless of the context within which the word appears. Discussing the same aspect in a similar semantic structure might have different meanings. In this paper, we apply BERT (Bidirectional Encoder Representations from Transformers) to obtain sentence embeddings.

BERT is a deep learning-based model in natural language processing, and the architecture is a multi-layer bidirectional Transformer encoder. It is designed to learn deep bidirectional representations from the unlabeled text by cooperatively considering both sides of context. BERT trains contextual representations on text corpus and produces word representations that are dynamically informed by the words around them. In contrast to previous efforts that read text sequentially either from left to right or right to left, BERT introduces more comprehensive and global word relationships to the word representation. BERT is bidirectional, generalizable, has high-performance, and universal. Since the pre-training procedure is comparatively hardware-demanding and time-consuming, we use BERT's own pre-built pre-training model, Chinese_L-12_H-768_A-12, which was trained by Google with Chinese Wikipedia data, as our pre-training model.

Select Hypernym Candidates After training sentence embeddings, we cluster them with the k-means clustering algorithm. We assume that sentence vectors within a cluster describe the same customer needs, and a limited number of core words could summarize the opinions of each cluster. To exploit variety and comprehensiveness, we select the non-repeated central words of each of the top 10 largest clusters as hypernym candidates. Both silhouette coefficients and BIC determine the optimal number of clusters.

The process of selecting hypernym candidates is as follows. Denote embedding of sentence i in cluster m as s_{mi}, word j as w_j, and if w_j appears in s_{mi} then indicator a_{mij} is 1, otherwise is 0. The number of sentence embeddings in cluster m is N_m and the number of w_j appearance in s_{mi} is n_{mij}. We first calculate the cosine distance between s_{mi} to its cluster center, d_{mi}, and the distance is proportional to its representativeness. We sum up the cosine similarities between w_j and the cluster center as its importance in cluster m:

$$F_{mj} = \sum_{i=1}^{N_m} \sum_{j=1}^{n_{cij}} d_{mi} a_{mij} \tag{1}$$

The cosine distance also represents the similarities between words in a sentence and its cluster center. Since words repeatedly appear in different sentences, we sum up the cosine distances between words and their cluster center as their final distances. The words with the largest similarity within each cluster are the most core words, and we select the top two words from each cluster as hypernym candidates. The whole process of selecting hypernym candidates is shown in Figure 2. Hypernyms candidates are then finalized after eliminating repeated candidates chosen from all clusters.

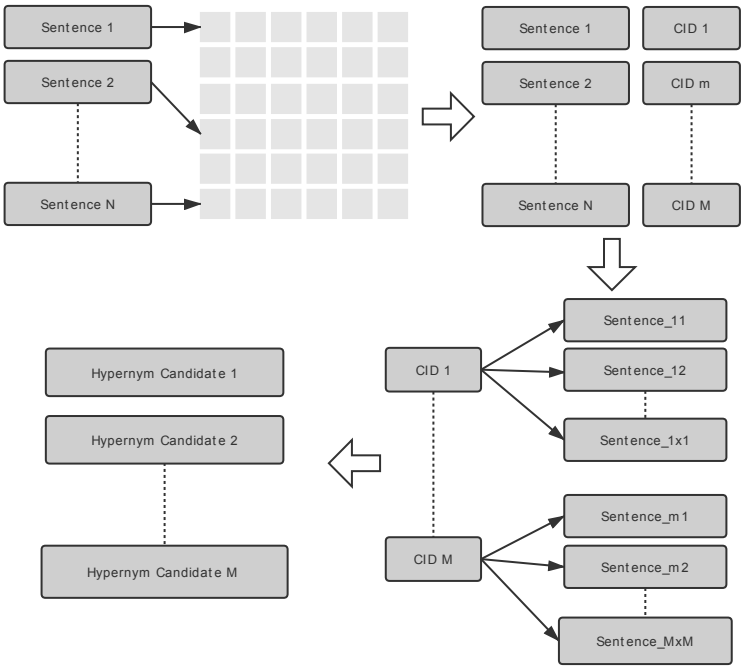

Fig. 2. Select Hypernyms

Recall Hyponym Candidates (with PageRank) As we mentioned earlier, hyponyms provide retailers with more detailed and granular information about a product. Another purpose of introducing hyponyms is that also they help matching hypernyms to more related reviews. We select the closest words to each hypernym candidate as second-order related words and again select the closest words to each second-order related words as third-order ones. Then we construct weighted directed wordnet where weights are determined by distances between pre-trained word embeddings. The process of building the wordnet is shown in Figure 3. Then we apply the PageRank algorithm to generate the final hyponyms according to their relative closeness and importance. PageRank is an iterative algorithm that determines the importance of a web page based on the importance of its parent page [26, 5, 13]. The core idea of PageRank is that the rank of an element is divided among its forward links evenly to contribute to the ranks of the pages they point to. After PageRank of each element is obtained, we select words with the highest PageRank between hypernyms as hyponym candidates. Compared with selecting the closest words to hypernyms and hyponyms, the main advantage of using PageRank is that it uses the entire graph rather than a small subset to estimate relative relationships between words. As a result, it enlarges the diversity, reliability, and richness of identified aspects.

The procedure of calculating PageRank is described as follows. Let F_i be the set of words that word i points to and B_i be the set of words that points to

i. Let $N_i = |F_i|$ be the number of links from i and let c be a factor used for normalization. PageRank of i is then

$$R(i) = c \sum_{v \in B_i} \frac{R(v)}{N_v} \tag{2}$$

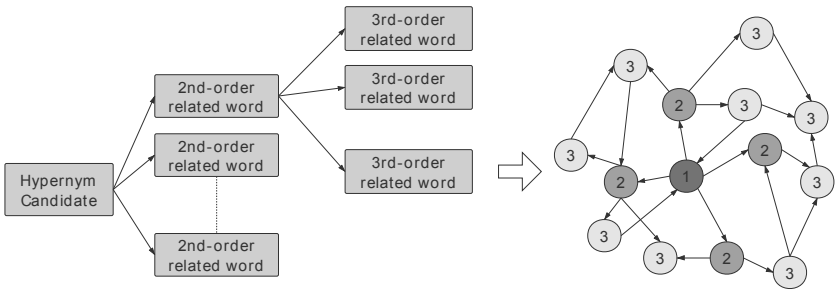

Fig. 3. Build wordnet

Merge Hypernyms and Hyponyms After finalizing the recall process, we noticed that there are overlaps between hypernyms and hyponyms. For example, hypernym candidate A is also a hyponym of hypernym candidate B. Overlaps would cause redundancy and confusion when mapping sentiment to aspects in the following steps. In order to further improve the precision of the constructed aspect lexicon and investigate the internal similarity between hypernym candidates, we merge hypernyms and hyponym candidates.

Match hypernyms to Reviews The next step is matching hypernyms to the reviews discussing corresponding aspects. With the pre-trained word embeddings, we select the closest words to hyponyms and match them with hypernyms and hyponyms to reviews with regular expression matching.

Sentiment Analysis In the next step of the AABSA model, we need to identify the sentiment evaluation of identified aspects. We applied the MaxEnt classification algorithm to the sentiment classification problem. MaxEnt models are feature-based models and could solve feature selection and model selection. MaxEnt classification is proved to be effective in a number of natural language processing applications [27, 3]. The goal is to assign a class c to a given document d to maximize $P(c|d)$, which is calculated as below:

$$P_{ME}(c|d) = \frac{1}{Z(d)} exp(\sum_i \lambda_{i,c} f_{i,c}(d, c)) \tag{3}$$

where $Z(d)$ is a normalization function. $F_{i,c}$ is a aspect function for aspect f_i and class c. $F_{i,c}(d, c') = 1$ if $n_i(d) > 0$ and $c' = c$. The $\lambda_{i,c}$ is a aspect-weighted

parameter and a large $\lambda_{i,c}$ means that f_i is considered a strong indicator for class c.

Now, each review can be represented as a vector consists of aspects and sentiment evaluations. We measure the overall sentiment evaluation of aspect i in week t as:

$$sentiment_{it} = \frac{\sum_{j=1}^{n_{it}} \frac{\sum_{k=1}^{m_{ijt}} sentiment_{kjt}}{m_{ijt}}}{n_{it}} \tag{4}$$

where n_{it} is the number of reviews that mention aspect i in week t and m_{ijt} is the time of i's appearance in review j in week t.

3.3 BASELINE MODEL

The baseline model is developed by [2] on deriving product aspects by mining consumer reviews. It mainly consists of four steps: pre-process content, eliminate synonyms, obtain core word candidates, and select hypernyms. After splitting reviews into sentences and remove stopwords, they calculate the TF-IDF value of each word and convert words into one-hot vectors with TF-IDF values of context words. Then they cluster word vectors with k-means clustering and choose the center words of clusters as hypernyms. The major differences are the training process of word vectors and the application of hierarchical structure of aspects. The architecture of the baseline model is indicated in Figure 4.

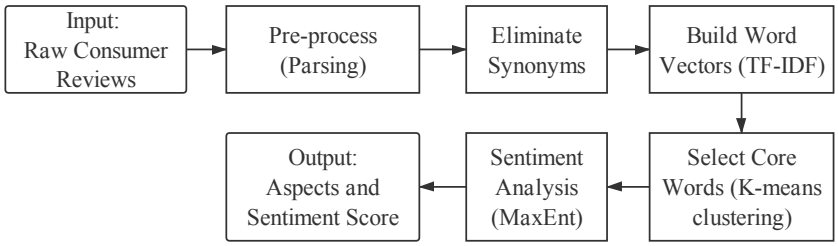

Fig. 4. Baseline Model Framework

4 Empirical Applications

In this section, we evaluate the AABSA model with review data drawn from product categories "Toothbrush" and "Camera" provided by Alibaba. In section 4.1, we first describe our data set. Then in section 4.2, we describe the identified aspects with AABSA.

4.1 Data

Alibaba Group is one of the largest e-commerce companies in Asia, which was first launched in 1999 in China. It is commonly referred to as the "Chinese

Amazon." As of June 2019, Alibaba has 755 million active users in more than 200 countries. It has three biggest digital shopping platforms, Alibaba, Taobao, and Tmall, which focus on B2B, C2B, and B2C business separately. Since 2010, Alibaba has launched sales on singles day in November and Spring Festival in January or February. In our work, we used panel data from 20 weeks between March and July to avoid fluctuation caused by the sales effect. For each item, we observe reviews, ratings, weekly sales, and essential attributes (e.g., price, weight, popularity), which are defined by retailers when the products were launched. The full data set consists of 295,628 reviews of 13,944 camera products and 18,550,956 reviews of 147,337 toothbrush products. In the pre-processing step, we use the reviews to build a vocabulary of nouns from which we select hypernyms and hyponyms. In the sentiment analysis step, we hired human taggers to classify aspect sentiments into three categories: positive, neutral, and negative.

4.2 Selecting Aspects

Tables 1 and Table 2 describe the top 10 hypernyms and 50 hyponyms from camera and toothbrush reviews. We find that aspects obtained from our model provide more detailed and comprehensive information on product aspects and customer needs. Each aspect captured by the AABSA model represents a detailed aspect, and it could provide clear instructions for firms to perform product improvement. For example, a positive "pixel" aspect indicates that the photo taken by the camera is clear. However, some words obtained by the baseline model, such as "cell phone" and "camera" are broad-defined aspects, and it is hard for firms to make specific improvements given this information.

To make an apple-to-apple comparison, among the aspects identified by the baseline model, we select the 10 most frequently-mentioned aspects. They are shown in Table 3.

5 Experimental Results

We describe two main sets of results: i) performance comparison of our AABSA and the benchmark model and ii) the marketing insights. First, we select the most popular and frequent 10 aspects from all hypernyms and hyponyms for AABSA and the baseline model. The aspects are shown in Table 4.

Accuracy To compare the performance of AABSA and the benchmark model, we create an econometric model to estimate the impact of each aspect on product sales. The intuition is that if the identified aspects are more useful for consumers and firms, they should be better predictors of product sales. We calculate the percentage of positive reviews of an aspect in the past 180 days of the week t as the support rate of the aspect and then use the support rate to predict product sales in week t in linear regression. In our model, there are 9 hypernyms and 1 hyponym of "price," "offline." We then report the performance of our model and the baseline model in terms of sales prediction accuracy and analyze the prediction power of each aspect. The regression result of cameras is

Table 1. Hypernyms and Hyponyms of Cameras

hypernyms	hyponyms
price	unworthy, value, half-price, replacement, incredible
pixel	effect, clear, figure, recorder, sense of camera
function	advancement, night-vision, fun, stabilization, fish-eye
packaging	beautiful packaging, delicate packaging, solid packaging, thickness, protection
outlook	chic, draft, specialty, delicacy, eye catching
efficacy	outstanding, slow motion, movie, sense of color, flashlight
photo	outstanding, clean, sense of color, color tune, high definition
battery	charging battery, battery life, duration, camera battery, forbidden
hue	brightness, gradation, sense of color, beauty, charm
color	outlook color, fashion, delicacy, red, pink

shown in table 5. The first column reports the estimates from AABSA, and the second column reports estimates from the baseline model.

We can make several inferences from the regression coefficients. First, in our model, coefficients for every aspect are significant, and the adjust r-squared is 5% higher than that of the baseline model. Second, we find that while positive reviews on most aspects have positive effects on sales, positives reviews on offline stores have negative effects on sales. One plausible explanation for this effect is that the offline and online stores are of competitive relationships, and customers would tend to switch to offline stores if they read related positive reviews on online retailing platforms.

The regression result of toothbrushes is shown in table 6. In the toothbrush category, the coefficients are all significant, and our model also out-performances the baseline model by 2%. However, the improvement is not as much as in the camera category. A plausible explanation is that toothbrushes are daily necessities, and they are much cheaper than cameras. As a result, when consumers purchase toothbrushes, they would spare less time to read textual reviews, and the sales prediction power of reviews is weakened.

Comprehensiveness We then compare the comprehensiveness of the aspects identified by our model to develop some intuition of what drives the performance discrepancy. [34] identified 6 primary customer needs and 22 secondary customer needs of oral care products with a machine-learning hybrid method and then classified the needs into the primary group and the secondary group. We compare the toothbrush aspects extracted from the AABSA model with aspects

Table 2. Hypernyms and Hyponyms of Toothbrushes

hypernyms	hyponyms
price	offline, average price, market price value package, retail store
brush head	children, easy to clean, gum pain, soft toughness
package	box, bag, small bag, foam bag, hole
smell	chocolate,juice, orange, mellow, fragrant
service	sincere
attitude	passionate
efficacy	outstanding, white, enhance, disease, stain removal
brush hair	thick, soft, weak, toughness, plentiful
gift	floss, pencil sharpener, color pen, origami case
color	brown, pink, beige, purple, green

Table 3. Top 10 Aspects Identified by the Baseline Model

Category	Aspects
Camera	clear, price, item, photo, satisfaction efficacy, camera, delivery, style, cell phone
Toothbrush	affordable, discount, offline, satisfaction, cheap, easy to use, delivery, purchase, strength, praise

extracted from [34]'s model. The comparison table is listed in Table 1 in the Appendices.

In [34]'s results, "feel clean and fresh" captures the customer's own oral feeling while and after using oral care products, and in AABSA's results, "easy to clean" hypernym aspect also describes the customers' feelings' of the oral smell after brushing teeth, and "toothbrush hair" and "toothbrush head" captures the comfort while using the toothbrush; "strong teeth and gums" describes the aspect of preventing gingivitis and protecting the gum, and we identified "gum pain"; "Product efficacy" describes the efficacy of oral care products, which focus

Table 4. Top 10 Aspects Identified by AABSA

Product Category	Aspects
Camera	price, price_offline, pixel, features, package, exterior, efficacy, photo, battery, color
Toothbrush	price, brush head, package, smell, service, attitude, efficacy, brush hair, gift, color

Table 5. Estimation Results of Cameras

AABSA model		Baseline model	
price	0.050 *	clear	0.110 ***
	(0.023)		(0.030)
price_offline	-0.062 *	price	0.198 ***
	(0.027)		(0.020)
pixel	0.127 ***	item	0.340 ***
	(0.016)		(0.019)
function	0.144 ***	photo	0.101 ***
	(0.020)		(0.021)
package	0.209 ***	satisfaction	0.002
	(0.016)		(0.057)
exterior	0.307 ***	efficacy	0.140 ***
	(0.021)		(0.019)
efficacy	0.215 ***	camera	0.289 ***
	(0.017)		(0.019)
photo	0.166 ***	delivery	-0.162 ***
	(0.019)		(0.029)
battery	0.202 ***	style	0.064 *
	(0.022)		()0.030
color	0.215 ***	cell phone	0.219 ***
	(0.023)		(0.244)
Adjust R-Squared: 0.294		Adjust R-Squared: 0.244	

on a more subjective aspect. It matches "efficacy" in our model, which reflects the effect of using the toothbrush; "Convenience" describes the convenience of using the oral product to reach the cleaning perspective, and AABSA identified "easy to clean" which also describes the toothbrush's ability to clean teeth; and "Shopping/product choice" describes the competitiveness between brands. From the above results, we can conclude that our model can create a more comprehensive list of aspects than [34].

6 Discussions

In this paper, we propose an innovative method to identify product aspects by introducing a hierarchical system. Compared with the previous aspect identification and sentiment analysis model, the AABSA model improves the comprehensiveness and the prediction accuracy on sales, and it is fully automatic in aspect-identification without time-consuming hand-coding. The method has been adopted by Alibaba's product review tagging system to help consumers search for product aspects.

References

1. Alam, I., Perry, C.: A customer-oriented new service development process. Journal of services Marketing **16**(6), 515–534 (2002)

Table 6. Estimation Results of Toothbrushes

AABSA model		Baseline model	
price	0.060 ***	affordable	0.038 ***
	(0.009)		(0.006)
brush head	0.379 ***	discount	0.266 ***
	(0.019)		(0.007)
package	0.544 ***	offline	-0.354 ***
	(0.010)		(0.010)
smell	0.726 ***	satisfaction	0.381 ***
	(0.011)		(0.006)
service	0.087 ***	cheap	0.316 ***
	(0.008)		(0.007)
attitude	0.773 ***	easy to use	0.246 ***
	(0.007)		(0.008)
efficacy	0.381 ***	delivery	0.073 ***
	(0.010)		(0.007)
brush hair	0.603 ***	purchase	0.376 ***
	(0.033)		(0.006)
gift	0.620 ***	strength	0.444 ***
	(0.010)		(0.009)
color	0.659 ***	praise	-0.212 ***
	(0.013)		(0.010)
Adjust R-Squared: 0.186		Adjust R-Squared: 0.164	

2. Archak, N., Ghose, A., Ipeirotis, P.G.: Show me the money!: deriving the pricing power of product features by mining consumer reviews. In: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 56–65. ACM (2007)
3. Berger, A.L., Pietra, V.J.D., Pietra, S.A.D.: A maximum entropy approach to natural language processing. Computational linguistics **22**(1), 39–71 (1996)
4. Blair-Goldensohn, S., Hannan, K., McDonald, R., Neylon, T., Reis, G., Reynar, J.: Building a sentiment summarizer for local service reviews (2008)
5. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. Computer networks and ISDN systems **30**(1-7), 107–117 (1998)
6. Chakraborty, I., Kim, M., Sudhir, K.: Attribute sentiment scoring with online text reviews: Accounting for language structure and attribute self-selection (2019)
7. Che, W., Zhao, Y., Guo, H., Su, Z., Liu, T.: Sentence compression for aspect-based sentiment analysis. IEEE/ACM Transactions on audio, speech, and language processing **23**(12), 2111–2124 (2015)
8. Chevalier, J.A., Mayzlin, D.: The effect of word of mouth on sales: Online book reviews. Journal of marketing research **43**(3), 345–354 (2006)
9. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
10. Dhar, V., Chang, E.A.: Does chatter matter? the impact of user-generated content on music sales. Journal of Interactive Marketing **23**(4), 300–307 (2009)
11. Gray, S., Radford, A., Kingma, D.P.: Gpu kernels for block-sparse weights. arXiv preprint arXiv:1711.09224 (2017)

12. Hadano, M., Shimada, K., Endo, T.: Aspect identification of sentiment sentences using a clustering algorithm. Procedia - Social and Behavioral Sciences **27**, 22 – 31 (2011). https://doi.org/https://doi.org/10.1016/j.sbspro.2011.10.579, http://www.sciencedirect.com/science/article/pii/S1877042811024062, computational Linguistics and Related Fields
13. Haveliwala, T.: Efficient computation of pagerank. Tech. rep., Stanford (1999)
14. Hu, M., Liu, B.: Mining opinion features in customer reviews. In: AAAI. vol. 4, pp. 755–760 (2004)
15. Johnson, R., Zhang, T.: Supervised and semi-supervised text categorization using lstm for region embeddings. arXiv preprint arXiv:1602.02373 (2016)
16. Kaulio, M.A.: Customer, consumer and user involvement in product development: A framework and a review of selected methods. Total quality management **9**(1), 141–149 (1998)
17. Kobayashi, N., Inui, K., Matsumoto, Y.: Extracting aspect-evaluation and aspect-of relations in opinion mining. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL). pp. 1065–1074 (2007)
18. Lee, H.Y., Renganathan, H.: Chinese sentiment analysis using maximum entropy. In: Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2011). pp. 89–93. Asian Federation of Natural Language Processing, Chiang Mai, Thailand (Nov 2011), https://www.aclweb.org/anthology/W11-3713
19. Lee, T.Y., Bradlow, E.T.: Automated marketing research using online customer reviews. Journal of Marketing Research **48**(5), 881–894 (2011)
20. Liu, X., Lee, D., Srinivasan, K.: Large scale cross category analysis of consumer review content on sales conversion leveraging deep learning. Available at SSRN 2848528 (2017)
21. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013)
22. Mubarok, M.S., Adiwijaya, Aldhi, M.D.: Aspect-based sentiment analysis to review products using naïve bayes. In: AIP Conference Proceedings. vol. 1867, p. 020060. AIP Publishing (2017)
23. Mullen, T., Collier, N.: Sentiment analysis using support vector machines with diverse information sources. In: Proceedings of the 2004 conference on empirical methods in natural language processing. pp. 412–418 (2004)
24. Netzer, O., Feldman, R., Goldenberg, J., Fresko, M.: Mine your own business: Market-structure surveillance through text mining. Marketing Science **31**(3), 521–543 (2012)
25. Nguyen, T.H., Shirai, K.: Phrasernn: Phrase recursive neural network for aspect-based sentiment analysis. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. pp. 2509–2514 (2015)
26. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: Bringing order to the web. Tech. rep., Stanford InfoLab (1999)
27. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up?: sentiment classification using machine learning techniques. In: Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. pp. 79–86. Association for Computational Linguistics (2002)
28. Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., Mohammad, A.S., Al-Ayyoub, M., Zhao, Y., Qin, B., De Clercq, O., et al.: Semeval-2016 task 5: Aspect based sentiment analysis. In: Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016). pp. 19–30 (2016)

29. Radford, A., Jozefowicz, R., Sutskever, I.: Learning to generate reviews and discovering sentiment. arXiv preprint arXiv:1704.01444 (2017)
30. Ruder, S., Ghaffari, P., Breslin, J.G.: A hierarchical model of reviews for aspect-based sentiment analysis. arXiv preprint arXiv:1609.02745 (2016)
31. Snyder, B., Barzilay, R.: Multiple aspect ranking using the good grief algorithm. In: Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference. pp. 300–307 (2007)
32. Tadano, R., Shimada, K., Endo, T.: Effective construction and expansion of a sentiment corpus using an existing corpus and evaluative criteria estimation. In: Proceedings of the 11th Conference of the Pacific Association for Computational Linguistics (PACLING2009). pp. 211–216. Citeseer (2009)
33. Thet, T.T., Na, J.C., Khoo, C.S.: Aspect-based sentiment analysis of movie reviews on discussion boards. Journal of information science **36**(6), 823–848 (2010)
34. Timoshenko, A., Hauser, J.R.: Identifying customer needs from user-generated content. Marketing Science **38**(1), 1–20 (2019)
35. Tirunillai, S., Tellis, G.J.: Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent dirichlet allocation. Journal of Marketing Research **51**(4), 463–479 (2014)
36. Tsai, Y.L., Wang, Y.C., Chung, C.W., Su, S.C., Tsai, R.T.H.: Aspect-category-based sentiment classification with aspect-opinion relation. In: 2016 Conference on Technologies and Applications of Artificial Intelligence (TAAI). pp. 162–169. IEEE (2016)
37. Wang, W., Pan, S.J., Dahlmeier, D., Xiao, X.: Recursive neural conditional random fields for aspect-based sentiment analysis. arXiv preprint arXiv:1603.06679 (2016)
38. Xue, B., Fu, C., Shaobin, Z.: A study on sentiment computing and classification of sina weibo with word2vec. In: 2014 IEEE International Congress on Big Data. pp. 358–363. IEEE (2014)
39. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., Le, Q.V.: Xlnet: Generalized autoregressive pretraining for language understanding. arXiv preprint arXiv:1906.08237 (2019)
40. Yu, J., Zha, Z.J., Wang, M., Chua, T.S.: Aspect ranking: Identifying important product aspects from online consumer reviews. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1. pp. 1496–1505. HLT '11, Association for Computational Linguistics, Stroudsburg, PA, USA (2011), http://dl.acm.org/citation.cfm?id=2002472.2002654
41. Zha, Z., Yu, J., Tang, J., Wang, M., Chua, T.: Product aspect ranking and its applications. IEEE Transactions on Knowledge and Data Engineering **26**(5), 1211–1224 (May 2014). https://doi.org/10.1109/TKDE.2013.136
42. Zha, Z.J., Yu, J., Tang, J., Wang, M., Chua, T.S.: Product aspect ranking and its applications. IEEE transactions on knowledge and data engineering **26**(5), 1211–1224 (2013)
43. Zhu, J., Wang, H., Tsou, B.K., Zhu, M.: Multi-aspect opinion polling from textual reviews. In: Proceedings of the 18th ACM conference on Information and knowledge management. pp. 1799–1802. ACM (2009)

7 Appendix

Table 1. Comprehensiveness Comparison

[34]'s oral care attributes	AABSA
Feel clean and fresh	
Clean feeling in my mouth	easy to clean
Fresh breath all day long	smell
Pleasant taste and texture	toothbrush hair, toothbrush head
Strong teeth	
Prevent gingivitis	gums, efficacy_disease
Product efficacy	
Able to protect my teeth	efficacy
Whiter teeth	efficacy_whitening
Effectively clean hard to reach areas	easy to clean
Knowledge and confidence	
Gentle oral care products	
Oral care products that last	
Tools are easy to maneuver and manipulate	
Knowledge of proper techniques	-
Long-term oral care health	
Motivation for good check-ups	
Able to differentiate products	
Convenience	
Efficient oral care routine	easy to clean
Oral care "away from the bathroom"	-
Shopping/product choice	
Faith in the products	-
Provides a good deal	price_value
Effective storage	package
Environmentally friendly products	-
Easy to shop for oral care items	-
Product aesthetics	-

Table 2. Robustness check: MaxEnt and Fastttext

MaxEnt	Precision	Recall	F1-score	support
0	0.910	0.900	0.905	23830
1	0.939	0.946	0.942	39055
Avg/total	0.928	0.928	0.928	39055
Fasttext	**Precision**	**Recall**	**F1-score**	**support**
0	0.867	0.869	0.868	23830
1	0.920	0.919	0.919	39055
Avg/total	0.900	0.900	0.900	62885